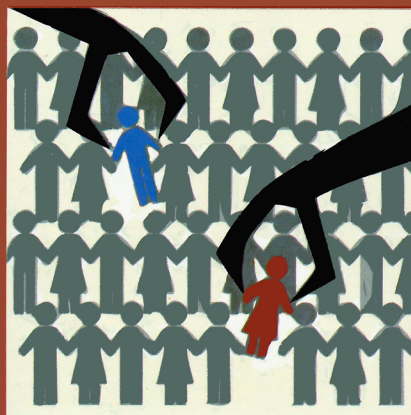


NCBA

Design of Experiments and Sampling Methods



K C Bhuyan

Design of Experiments
and
Sampling Methods



"This page is Intentionally Left Blank"

Design of Experiments and Sampling Methods

Dr K C Bhuyan

Department of Mathematics

Faculty of Science and Information Technology

American International University, Bangladesh

Also Author of Probability, Distribution Theory and Statistical Inference,
Advanced Biostatistics and Multivariate Analysis

New Central Book Agency (P) Ltd

LONDON

HYDERABAD ERNAKULAM BHUBANESWAR

NEW DELHI KOLKATA PUNE GUWAHATI

NCBA

REGD OFFICE

8/1 Chintamani Das Lane, Kolkata 700 009, India
email: ncbapvtltd@eth.net

OVERSEAS

NCBA (UK) Ltd, 149 Park Avenue North, Northampton, NN3 2HY, UK
email: ncbauk@yahoo.com

EXPORT

NCBA Exports Pvt. Ltd, 212 Shahpur Jat, New Delhi 110 049
email: ncbalexports@ncbapvtltd.com

BRANCHES

2/27 Ansari Road, 1st Floor
Daryaganj, New Delhi 110 002
email: ncbadel@ncbapvtltd.com

Shop Nos. 3 & 4, Vinayak Towers
681/B Budhwar Peth, Appa Balwant Chowk
Pune 411 002
email: ncbapune@ncbapvtltd.com

Shop No. 15, Bharati Towers, Ground Floor
Forest Park
Bhubaneswar 751 001
email: ncbabub@ncbapvtltd.com

House No. 3-1-315, 1st Floor
Nimboliadda, Kachiguda, Hyderabad 500 027
email: ncbahydb@ncbapvtltd.com

GSS Shopping Complex, 1st Floor
Opposite College Post Office, Convent Road
Ernakulam 682 035
0484-4021054/53

Radhamayee Complex (Second Floor)
Opp. SBI, Panbazar, Jaswanta Road
Guwahati 781 001
email: ncbaguwahat@ncbapvtltd.com

DESIGN OF EXPERIMENTS AND SAMPLING METHODS • Dr K C Bhuyan

© Copyright reserved by the Author

Publication, Distribution, and Promotion Rights reserved by the Publisher

All rights reserved. No part of the text in general, and the figures, diagrams, page layout, and cover design in particular, may be reproduced or transmitted in any form or by any means—electronic, mechanical, photocopying, recording, or by any information storage and retrieval system—without the prior written permission of the Publisher

First Published: January 2017

Revised Edition: 2021

PUBLISHER

New Central Book Agency (P) Ltd
8/1 Chintamani Das Lane, Kolkata 700 009

TYPESETTER

Anin, Kolkata

PRINTER

New Central Book Agency (P) Ltd
Web-Offset Division, Dhulagarh, Sankrail, Howrah

TECHNICAL EDITOR

Dipan Roy

PROJECT TEAM

Pradip Biswas and Prodip Baidya

ISBN: 978-93-90530-68-7

Dedicated to

my wife, Sova

daughter, Simon and son, Dipan

"This page is Intentionally Left Blank"

Contents

Preface

xiii

DESIGN OF EXPERIMENTS

Chapter 1 : Design of Experiments and Analysis of Variance	3–26
1.1 Introduction	3
1.2 Technique of Analysis of Variance	4
1.3 Linear Model for Analysis of Variance	5
1.4 Regression Analysis and Analysis of Variance	6
1.5 Assumptions in Analysis of Variance	7
1.6 Consequences of Violation of Assumptions	7
1.7 Terms Related to Experiment	8
1.8 Design of Experiment	8
1.9 Conditions for Efficient Experiment	10
1.10 Estimation and Test of Hypothesis	11
1.11 Multiple Comparison	20
1.12 Estimation of Missing Observation	24
Chapter 2 : Multi-Way Classification	27–71
2.1 One-Way Classification	27
2.2 Two-Way Classification	33
2.3 Two-Way Classification with Several (Equal) Observations Per Cell	40
2.4 Two-Way Classification with Several (Unequal) Observations Per Cell	44
2.5 Two-Way Classification with Several (Unequal) Observations Per Cell with Interaction	53
2.6 Three-Way Classification	57
2.7 Three-Way Classification with Several (Equal) Observations Per Cell	63
Chapter 3 : Basic Design	72–122
3.1 Introduction	72
3.2 Completely Randomized Design (CRD)	72
3.3 Randomized Block Design (RBD)	75
3.4 Randomized Block Design with More than One (Equal) Observations Per Cell	78
3.5 Efficiency of Randomized Block Design	81
3.6 Advantages, Disadvantages and Uses of Randomized Block Design	82
3.7 Randomized Block Design with Missing Observation(s)	83
3.8 Latin Square Design (LSD)	90
3.9 Analysis of Latin Square Design with Missing Observations	97
3.10 Efficiency of Latin Square Design	107
3.11 Advantages, Disadvantages and Uses of Latin Square Design	108

3.12	Analysis of Latin Square Design with Several (Equal) Observations Per Cell	109
3.13	Analysis of p Latin Square Designs	114
3.14	Orthogonal Latin Square Designs	119
3.15	Graeco Latin Square Design	121
Chapter 4 : Factorial Experiment		123–212
4.1	Introduction	123
4.2	3^n -Factorial Experiment	136
4.3	4^n -Factorial Experiment	145
4.4	p^n -Factorial Experiment	149
4.5	Generalized Interaction	151
4.6	Confounded Factorial Experiment	151
4.7	Fractional Replication of Factorial Experiment	165
4.8	Advantages and Disadvantages of Factorial Experiment	172
4.9	Advantages and Disadvantages of Confounding	173
4.10	Asymmetrical Factorial Experiment	173
4.11	Split-Plot Design	178
4.12	Estimation of Missing Value in Case of Split-Plot Design	187
4.13	Split-Split-Plot Design	188
4.14	Split-Block Design (Split-Plot Design with Sub-units in Strips)	204
Chapter 5 : Incomplete Block Design		213–235
5.1	Introduction	213
5.2	Balanced Incomplete Block (BIB) Design	214
5.3	Analysis of BIB Design	218
5.4	Partially Balanced Incomplete Block (PBIB) Design	229
5.5	Analysis of PBIB Design	231
Chapter 6 : Covariance Analysis		236–265
6.1	Definition	236
6.2	Covariance Analysis in Case of Completely Randomized Design (CRD) with One Concomitant Variable	237
6.3	Covariance Analysis in Completely Randomized Design with Two Concomitant Variables	243
6.4	Covariance Analysis in Randomized Block Design with One Concomitant Variable	247
6.5	Covariance Analysis in Randomized Block Design with Two Concomitant Variables	252
6.6	Technique of Covariance Analysis in Analysing Data of Randomized Block Design with Missing Observations	257
6.7	Covariance Analysis in Latin Square Design with One Concomitant Variable	262

Chapter 7 : Variance Component Analysis	266–280
7.1 Introduction	266
7.2 Assumptions in Variance Component Analysis	267
7.3 Method of Variance Component Analysis	267
7.4 Variance Component Analysis in Two-Way Classification	269
7.5 Steps in Calculating $E(MS)$ for Variance Component Analysis	271
7.6 Demerits of Analysis of Variance Technique in Variance Component Analysis	274
7.7 Variance Component Analysis for Three-Way Classification	276
Chapter 8 : Nested Classification	281–297
8.1 Introduction	281
8.2 Two-Stage Nested Classification	282
8.3 Three-Stage Nested Classification	285
8.4 Unbalanced Three-Stage Nested Classification	289
8.5 Nested and Cross Classification	295
Chapter 9 : Group of Experiments	298–308
9.1 Introduction	298
9.2 Analysis of Groups of Randomized Block Designs	299
9.3 Analysis when Error Variances are Heterogeneous	300
Chapter 10 : Construction of Design	309–326
10.1 Introduction	309
10.2 Some Mathematical Concept Related to Construction of Design	309
10.3 Construction of Incomplete Block Design Using Primitive Root	311
10.4 Construction of Design from Other Design	314
10.5 Construction of Incomplete Block Design from Orthogonal Latin Square Design	316
10.6 Optimum Design	318

SAMPLING METHODS

Chapter 11: Elementary Discussion on Sampling Methods	329–340
11.1 Concept of Sampling	329
11.2 Scope of Sampling	329
11.3 Important Points to be Considered During Sampling	330
11.4 Different Sampling Methods	330
11.5 Advantages and Limitations of Sampling	332
11.6 Census and Sample Survey	333
11.7 Merits and Demerits of Census and Sample Survey	333
11.8 Principal Steps in a Sample Survey	333
11.9 Sampling Error and Precision	336

11.10	Reliability	337
11.11	Determination of Sample Size	337
11.12	Non-Sampling Error	339
11.13	Bias	340
Chapter 12	Simple Random Sampling	341–356
12.1	Definition and Estimation of Parameter	341
12.2	Estimation of Proportion in Case of Simple Random Sampling	350
Chapter 13	Stratified Random Sampling	357–392
13.1	Definition	357
13.2	Allocation of Sample Size in Different Strata	366
13.3	Estimation of Proportion from Stratified Random Sample	373
13.4	Relative Precision of Simple Random and Stratified Random Sampling	380
13.5	Estimation of Gain in Precision due to Stratification	382
13.6	Method of Construction of Strata	384
13.7	Number of Strata	385
13.8	Effects of Error Due to Estimation of Stratum Size	386
13.9	Determination of Sample Size	387
13.10	Effect of Deviation from Optimum Allocation	389
13.11	Stratified Sampling with Varying Probabilities	391
13.12	Post-stratification	392
Chapter 14	Systematic Sampling	393–407
14.1	Introduction	393
14.2	Method of Selecting Systematic Sample	393
14.3	Advantages and Disadvantages of Systematic Sampling	395
14.4	Method of Estimation in Systematic Sampling	396
14.5	Systematic Sampling when Population Units are in Random Order	405
14.6	Systematic Sampling in Populations with Linear Trend	406
Chapter 15	Ratio and Regression Estimator	408–448
15.1	Ratio Estimator	408
15.2	Estimation of Variance of Ratio Estimator from Sample	412
15.3	Comparison of Ratio Estimator and Simple Estimator	415
15.4	Unbiased Ratio Estimator and its Variance	415
15.5	Unbiased Ratio Type Estimator	420
15.6	Conditions Under which Ratio Estimator is a Best Linear Unbiased Estimator	422
15.7	Ratio Estimator when \bar{X} is not Known	423
15.8	Difference Estimator	427
15.9	Ratio Estimator in Case of Stratified Sampling	429

15.10	Optimum Allocation in Case of Ratio Estimator	432
15.11	Ratio Estimator to Estimate the Parameter Related to Qualitative Variable	433
15.12	Regression Estimator	434
15.13	Bias of Regression Estimator	435
15.14	Variance of Regression Estimator	436
15.15	Comparison of Regression Estimator, Ratio Estimator and Simple Estimator	440
15.16	Regression Estimator in Case of Stratified Random Sampling	442
Chapter 16 : Cluster Sampling		449–470
16.1	Introduction	449
16.2	Method of Estimation in Cluster Sampling	450
16.3	Method of Estimation in Cluster Sampling when Clusters are of Unequal Sizes	452
16.4	Estimation of Sampling Variance in Case of Cluster Sampling	454
16.5	Relative Efficiency of Cluster Sampling	456
16.6	Cluster Sampling with Varying Probabilities	458
16.7	Cluster Sampling to Estimate Proportions	468
Chapter 17 : Two-Stage Sampling		471–501
17.1	Definition	471
17.2	Advantage and Use of Two-Stage Sampling	471
17.3	Estimation of Parameter in Two-Stage Sampling	472
17.4	Estimation of Parameters in Two-Stage Sampling with Unequal First Stage Units	479
17.5	Estimation of Proportion in Two-Stage Sampling with Equal First Stage Units	486
17.6	Allocation of Sample Sizes in Two-Stages	187
17.7	Two-Stage Sampling with Varying Probabilities	492
17.8	Method of Estimation in Two-Stage Sampling with Varying Probabilities	493
17.9	Two-Stage Sampling With Varying Probabilities at Each Stage	497
Chapter 18 : Three-Stage Sampling		502–514
18.1	Three-Stage Sampling with Equal First-Stage Units	502
18.2	Three-Stage Sampling with Unequal First-Stage and Second-Stage Units	506
18.3	Allocation of Sample Sizes in Three-Stages	512
Chapter 19 : Multiphase Sampling		515–527
19.1	Introduction	515
19.2	Double Sampling for Stratification	516
19.3	Double Sampling for Ratio Estimator	518
19.4	Double Sampling for Regression Estimator	522
19.5	Optimum Allocation in Double Sampling for Ratio Estimator	525
19.6	Optimum Allocation in Double Sampling for Regression Estimator	525
19.7	Double Sampling for Difference Estimator	526

Chapter 20 : Sampling with Varying Probabilities	528–557
20.1 Introduction	528
20.2 Method of Selection of Sampling Units with Varying Probabilities	528
20.3 Method of Estimation in PPS Sampling with Replacement	530
20.4 PPS Sampling without Replacement	535
20.5 Method of Estimation in PPS Sampling without Replacement	539
20.6 Other Methods of Sampling with Varying Probabilities	552
20.7 Sampling Procedures where Inclusion Probability is Proportional to Size (π PS Sampling)	555
Chapter 21 : Non-Sampling Error	558–580
21.1 Introduction	558
21.2 Effects of Non-Response	559
21.3 Technique for Adjustment of Non-Responses	560
21.4 Call Backs and its Effects	561
21.5 Politz-Simmons Technique	563
21.6 Response Errors	565
21.7 Determination of Optimum Number of Investigators	571
Appendix—1–7	581–592
Index	593–594

Preface

The sources of statistical data are the experiments—controlled or uncontrolled—in any field of inquiry. The sample survey is the source of data from uncontrolled experiment; and design of experiment is the source of data from controlled experiment. The experimental results are analysed and interpreted using statistical tools. The analytical procedure for the data of design of experiments is of one type and it is different from that of the data collected through sample survey. Both the analytical procedures are discussed in the book with examples.

Out of 21 chapters of the book the first 10 chapters are dedicated to the topics of design of experiments while last 11 chapters cover the aspects of sampling methods. However, the applied research workers in any field may use the book for analysing the experimental data. All theoretical discussions are amply supported by examples.

In writing this book I have used the books and papers of several well known authors and researchers. Specially, the works of Cochran, Cochran and Cox, Des Raj, Deming, Das and Giri, Hansen, Hurwitz and Madow, Kempthorne, Murthy, Sukhatme and Sukhatme, and Yates are to be duly acknowledged with gratitude.

My sincere thanks go to Mr. Amitabha Sen, Director, New Central Book Agency (P) Ltd and the entire production team of the publisher for successfully completing the work in time.

Any suggestion whatsoever from any quarter to improve the text will be highly appreciated.

K C Bhuyan

"This page is Intentionally Left Blank"

Design of Experiments

"This page is Intentionally Left Blank"

Chapter 1

Design of Experiments and Analysis of Variance

1.1 Introduction

The objective of experiments in agriculture, industry, life science, physical science, engineering sciences and in social sciences is to test the significance regarding the impacts of levels of factors involved in the experiment. The input of any factor is assumed to be absent or insignificant and the researcher needs to test the significance of the hypothesis based on the assumption. The test of hypothesis is performed using the data collected through experiment. The method of experimentation or the method of collection of responses for certain factor level is known as design of experiment.

The experiments conducted in different fields mentioned above are classified into three. These are (i) varietal trial experiment, (ii) factorial experiment, and (iii) experiment on bio-assay.

Varietal trial experiment : Let us consider that a rice research institute discovers 5 varieties of rice and the researcher needs to identify a variety which is best in terms of economy of cultivation. The investigator, in practice, needs to perform some experiments that is he needs to cultivate the rice varieties in homogeneous agricultural conditions so that data on production of rice varieties are collected. The collected data will be used to verify the assumption of the investigator related to one or more varieties of rice. Here the cultivation of rice varieties is known as varietal trial experiment.

Factorial experiment : It is mentioned that rice varieties are cultivated under homogeneous agro-climatic conditions. In practice, a particular variety of rice may have more yielding capacity under a particular level of irrigation in presence of a particular dose of fertilizer. This involves three factors in cultivation. These are rice variety, fertilizer and irrigation. The experiment may be conducted using different varieties of rice, different levels of a fertilizer and different doses of irrigation. Such experiment where different factor levels are used to get the required responses is known as factorial experiment.

Experiment on bio-assay : In medical science, experiments are performed to identify the dose(s) of a particular medicine against a disease. The response of a medicine may vary from individual to individual depending on the stage(s) of disease. The investigator conducts the experiment to identify the doses of a medicine for patients of different ages, of different body condition scores, etc. Such experiment is usually known as experiment on bio-assay.

Whatever be the experiment, it is conducted to collect information. The method of collection of data may vary from situation to situation. This means that the design of experiments are of different types and each type depends on agro-climatic condition or on experimental situation. With the variation in the factors or factor levels or in the experimental conditions the responses usually vary. The sources of variation in the responses are more. The investigator needs to estimate the variance in the data and also needs to estimate the variances according to pre-identified sources of variation. The technique of partition of total variation in the data set into

different component variations according to pre-identified sources of variation is known as analysis of variance.

According to R.A. Fisher, "The analysis of variance is a technique to partition the total variance of a particular data set into component variances according to pre-identified sources of variation." This technique has been developed first by R.A. Fisher. Later on F. Yates and others contributed a lot in the field of design and analysis of experiments.

1.2 Technique of Analysis of Variance

Let us consider that k varieties of rice are cultivated in n plots, where i -th variety ($i = 1, 2, \dots, k$) is repeatedly cultivated in n_i plots such that $n = \sum n_i$ and all the plots are homogeneous in respect of size, shape and agro-climatic conditions. Let y_{ij} be the production of i -th variety of rice in j -th plot ($j = 1, 2, \dots, n_i$). The values of y_{ij} can be shown as below :

Variety of Rice	Production of rice (y_{ij})	Total $y_{i\cdot}$	Mean \bar{y}_i
R_1	$y_{11}, y_{12}, \dots, y_{1j}, \dots, y_{1n_1}$	$y_{1\cdot}$	\bar{y}_1
R_2	$y_{21}, y_{22}, \dots, y_{2j}, \dots, y_{2n_2}$	$y_{2\cdot}$	\bar{y}_2
\vdots	$\dots\dots\dots$	\vdots	\vdots
R_i	$y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{in_i}$	$y_{i\cdot}$	\bar{y}_i
\vdots	$\dots\dots\dots$	\vdots	\vdots
R_k	$y_{k1}, y_{k2}, \dots, y_{kj}, \dots, y_{kn_k}$	$y_{k\cdot}$	\bar{y}_k
Total		$y_{..} = G$	$\bar{y}_{..}$

Since rice varieties are cultivated in plots of homogeneous size, shape and agro-climatic conditions, the productions of a particular rice variety are assumed to be similar and each production of i -th variety is supposed to be \bar{y}_i . However, due to some uncontrolled sources of variation (in agricultural experiment soil fertility variation, insect bite, excessive sun light, etc. are some of the causes of uncontrolled sources of variation), the values of y_{ij} ($j = 1, 2, \dots, n_i$) may vary and within the observations of a particular variety of rice there is a variation. Such variation of observations can be measured by an amount $\sum \sum (y_{ij} - \bar{y}_i)^2$. Again, if it is known that the rice varieties are similar, there is no need of conducting the experiment. The usual assumption with new varieties of rice is that these are not homogeneous. Under this assumption, the means $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k$ are not same. There may be dispersion in these means which can be measured taking the deviations of the means about grand mean $\bar{y}_{..}$. The amount of dispersion can be measured by a quantity proportional to $\sum (\bar{y}_i - \bar{y}_{..})^2$.

The total variation of all the y_{ij} observations can be measured by an amount proportional to $\sum \sum (y_{ij} - \bar{y}_{..})^2$ and this variation is mainly due to the sources of variation within i -th set of observations and due to the variation between k sets of observation. Now, the total variation (total sum of squares) of y_{ij} ($i = 1, 2, \dots, k, j = 1, 2, \dots, n_i$) observations can be partitioned as follows :

$$\sum \sum (y_{ij} - \bar{y}_{..})^2 = \sum \sum [(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}_{..})]^2 = \sum n_i (\bar{y}_i - \bar{y}_{..})^2 + \sum \sum (y_{ij} - \bar{y}_i)^2$$

= sum of squares within + sum of squares between

$$SS \text{ (total)} = SS \text{ (within)} + SS \text{ (between)}.$$

The above technique of partitioning the total sum of squares into components sum of squares according to pre-identified sources of variation is known as analysis of variance technique.

1.3 Linear Model for Analysis of Variance

It has already been mentioned in the previous section that y_{ij} ($i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$) observations are collected on j -th occasion for i -th variety of rice. These observations are, usually, sample observations. Let us assume that i -th set of observations are drawn randomly from a normal population with mean μ_i and variance σ^2 [$y_{ij} \sim N(\mu_i, \sigma^2)$]. It is also mentioned that the i -th set of observations are not same. These are deviated by a certain amount, say e_{ij} , from the population mean. Here e_{ij} ($i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$) is the random component (error) or the amount due to uncontrolled source in y_{ij} observations. Therefore, the y_{ij} observations can be expressed by a linear mathematical function :

$$\begin{aligned} y_{ij} &= \mu_i + e_{ij}; \quad i = 1, 2, \dots, k; j = 1, 2, \dots, n_i \\ &= \mu + (\mu_i - \mu) + e_{ij} = \mu + \alpha_i + e_{ij}, \end{aligned} \quad (1)$$

where μ is the general mean of all observations under the assumption that all populations are similar, α_i is the amount of deviation of i -th population mean about the grand mean (usually called effect of i -th variety), where $\alpha_i = \mu_i - \mu$. The representation of y_{ij} observations as in (1) is a linear model. This model is essential for analysis of variance.

In section 1.2, it is mentioned that, except uncontrolled source of variation, the main variation in the set of observations is due to the variation in the variety. If in such agricultural experiment, the rice varieties are cultivated under fertilizer trial, say, using q levels of a particular fertilizer, then the total variation in the data set will be mainly due to two sources of variation, namely variety of rice and level of fertilizer. In such a case, the y_{ij} observations can be expressed by a linear model.

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl} \dots \quad (2)$$

$i = 1, 2, \dots, k; j = 1, 2, \dots, q; l = 1, 2, \dots, n_i$ (or n_{ij}), where the new term β_j is the effect of j -th level of fertilizer.

Model : It is a mathematical equation formed with random variable, mathematical variable and parameter(s). Thus, in (1) e_{ij} is introduced as a source of random component and its distribution may be assumed as the distribution of a random variable, the term μ_i and hence, α_i is arisen from the population parameter and, therefore, μ and α_i are parameters. The equation as shown in (1) is a model.

Linear Model : The equation which is comprised of random variable, mathematical variable and parameter and which is linear in observations and parameter(s) is known as linear model. Thus, equations (1) and (2) are linear models.

The general linear model is expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e,$$

where $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are parameters; x_1, x_2, \dots, x_k are mathematical variables; e is a random variable. Here due to functional form y is also a random variable, the observations of which can be collected through experimental design. The values of random variable e are not observed. However, these may be estimated through analysis. It is assumed that, for the purpose of analysis, e is distributed with mean zero.

So far we have discussed linear model. The model may be non-linear such as quadratic, exponential, etc. But for analysis of variance purpose we shall confine our discussion within linear model only.

In the models (1) and (2), $\alpha_i, \beta_j, (\alpha\beta)_{ij}, \mu$ are assumed as parameters, where α_i is termed as the effect of i -th variety and β_j is the effect of j -th level of fertilizer and $(\alpha\beta)_{ij}$ is the interaction

of i -th variety with j -th level of fertilizer. If the selected varieties for any experiment are not the population (only available) varieties, α_i is not a parameter. Let us consider that k varieties are randomly selected from a group of varieties for an experiment. In such a case the effect of varieties will be random variable. Thus, depending on the characteristic of the effects the model is of three types. These are (a) fixed effect model, (b) random effect model, and (c) mixed effect model.

Fixed effect model : The linear model in which the effects are fixed is called fixed effect model. In the following model,

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl}, \quad (3)$$

$i = 1, 2, \dots, p$; $j = 1, 2, \dots, q$; $l = 1, 2, \dots, r$ if the effects α_i , β_j and $(\alpha\beta)_{ij}$ are fixed effects of i -th level of A , j -th level of B and interaction of i -th level of A with j -th level of B , respectively, then the model(3) is a fixed effect model.

Random effect model : The linear model in which the effects except the additive constant (general mean) are random variables is called a random effect model.

Thus, in model (3), if α_i, β_j and $(\alpha\beta)_{ij}$ are random variables, the model is called random effect model. The analysis of data assuming random effect model is known as variance component analysis.

Mixed effect model : If some of the effects are fixed except additive constant and some are random variables in a model, it is called mixed effect model.

Let us consider that the effects α_i are fixed, the effects β_j are random variable and so $(\alpha\beta)_{ij}$ are random variable in model (3). Then the model is known as mixed effect models.

Unless otherwise assumed, the analysis of variance technique can be used to analyze the fixed effect, mixed effect and random effect models.

1.4 Regression Analysis and Analysis of Variance

The general linear model is assumed as

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_k x_{kj} + e_j, \quad (4)$$

$j = 1, 2, \dots, n$, where y_j is the j -th value of a dependent variable which depends on the values of k explanatory variables x_1, x_2, \dots, x_k ; $\beta_0, \beta_1, \dots, \beta_k$ are parameters; e_j is the j -th value of random component corresponding to y_j ; β_i 's ($i = 1, 2, \dots, k$) are parameters indicating the rate of change of y for unit change in the value of x_i , when x_i 's are orthogonal. The general linear model (4) is known as a regression model, if x_i 's are continuous variable taking values in the limit $-\infty$ to ∞ .

In matrix notation, the model is written as

$$Y = XB + U, \quad (5)$$

$$\text{where } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}_{n \times (k+1)}, \quad B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1) \times 1}, \quad U = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n \times 1}$$

The problem in regression analysis is to estimate the parameter vector B and is to test the significance of this parameter vector. Finally, the analysis leads to predict the value of y for a certain set of values of x_i 's.

In some instances, the values of x_i 's are either zero or 1 indicating the presence ($x_i = 1$) or absence ($x_i = 0$) of a variable. For example, let us consider that all values of x_1 in model (4) are 1 and the values of other x 's are zero. Then the model transforms to

$$y_j = \beta_0 + \beta_1 + e_j. \quad (6)$$

This model (6) is similar to the model (1) of section (1.3). Again, let us consider that $x_{1j} = 1$, $x_{2j} = 1$, $x_{3j} = 1$ for all j and $x_{4j} = 0, \dots, x_{kj} = 0$. Then, we can write :

$$y_j = \beta_0 + \beta_1 + \beta_2 + \beta_3 + e_j$$

$$\text{or, } y_{ij} = \beta_0 + \beta_i + e_j; \quad i = 1, 2, 3; \quad j = 1, 2, \dots, n. \quad (7)$$

This model (7) is a linear model exactly same as model (1). This latter model is known as analysis of variance model or experimental design model, where the explanatory variables are indicator variables indicating the presence or absence of the levels of a factor. Here β_i is the impact of i -th level of a factor (variable). In that sense, both regression parameters and parameters involved in analysis of variance model are similar. However, in respect of analysis, the regression analysis and analysis of variance are not exactly similar. In regression analysis emphasis is put on the estimate of the parameters and on the prediction of the value of dependent variable. But in analysis of variance emphasis is given not on the estimation of the parameter rather on the significance of the impacts and on the differences in impacts of different levels of a factor. In that sense, analysis of variance is a special type of regression analysis where there is difference in the structure of the models of two analysis.

1.5 Assumptions in Analysis of Variance

It has already been mentioned that analysis of variance involves the test of significance of the effects of factor or factor levels. The test is performed using the concept of sampling distribution, where the basis of the sampling distribution is that the samples are drawn independently from normal population. Thus, the main assumption for analysis of variance is that the random component is normally and independently distributed with mean zero and common variance σ^2 [Thus, the random component in model (1) is $e_{ij} \sim \text{NID}(0, \sigma^2)$]. This assumption implies that the observations are also normally and independently distributed with common variance σ^2 .

The assumption mentioned above is needed to apply the usual analysis of variance F -test. However, the normality and common variance of observations are not ensured in all experimental situation. Of course, the violation of the assumption does not invalidate the analysis. The problem of non-normality and/or heteroscedasticity may be avoided by using data transformation technique. The important transformations are (a) \sin^{-1} transformation, (b) arc sin transformation, (c) square root transformation. Besides the application of data transformation technique, nonparametric methods are also used as a mode of analysis. For details of data transformation technique Das and Giri (1986) and for nonparametric analysis of data Conover (1999) may be consulted.

1.6 Consequences of Violation of Assumptions

The estimation of parameters in the model is done using method of least squares. For this the errors need to be distributed independently and with common variance. However, the correlated or heterogeneous errors do not create any problem in obtaining unbiased estimators. Normality of data also is not needed to obtain unbiased estimators of parameters. But heterogeneous and correlated errors provide inefficient estimators which effect the test of significance. Non-normality of data creates the problem in using analysis of variance F -test.

1.7 Terms Related to Experiment

In the introductory section, it is mentioned that the experiment is conducted to investigate the characteristic of certain phenomenon (variety), where data on phenomenon are collected from some experimental units after the experiment. The important terms related to experiment are (a) treatment, (b) experimental plot, (c) block, (d) yields and (e) experimental error.

(a) Treatment : The phenomenon which is under investigation for possible comparison of the effects of its different levels or which are under investigation to test the significance of its effects is known as treatment. For example, if different rice varieties are used in any experiment to select the best variety, then variety of rice treated as treatment. If different doses of urea are used in any agricultural experiment to identify the best dose, then nitrogen fertilizer is a treatment. In a production process in any industry, if the process is needed to be identified as the best one, the process is considered as treatment.

The treatment has different levels. In rice varietal trial experiment, the different varieties of rice are the levels of treatment; in fertilizer trial experiment, the doses of urea are the levels of treatments; different production processes of the same industry are the levels of treatment.

(b) Experimental Plot : The experimental materials on which a particular level of a treatment is applied once, constitute the experimental plot. In agricultural experiment, a plot of land is considered as an experimental plot. In dairy science of feeding trial, a cow or a group of cows of same age or same height or same weight kept in a shed may be considered as a plot. In selecting proteins for a baby, the baby may be considered as an experimental plot.

(c) Block : A group of homogeneous plots constitute a block. For example, in a dairy feeding-trial a group of cows of same age or of same origin or of same lactation period may constitute a block. In agricultural experiment, a group of agricultural plots of same soil fertility level will constitute a block. Since all the plots in a block are homogeneous, the experimental results (yields) of different treatments are expected to be similar if the treatments are of homogeneous impacts.

(d) Yields : The outcome of an experiment is known as yield. If the I.Q. of boys of different sociocultural group of people is under investigation, the I.Q. of each boy is the yield. In varietal trial experiment with agricultural crop, the per acre production of the crop or per plot production of the plot is known as yield. In medical science, if doses of medicine are used to reduce systolic blood pressure of patients, then blood pressure of a patient after using medicine is the yield. The observations collected from all experimental plots (units) are the yields to be analysed.

(e) Experimental Error : The yield of an ongoing experiment is not affected only by treatment or factors used in the experiment. There are some sources of variation which are beyond the control of the researcher. This uncontrolled source of variation affects the outcome of the experiment. The error crept in the experimental output due to uncontrolled source of variation is known as experimental error. In agricultural experiment, the production of a crop may be affected by insects or by drought or flood. These sources of insect bites, drought or flood cannot be controlled by design of experiment and these sources affect the outcome of the treatment. The error arisen due to these uncontrolled sources of variation is known as experimental error. However, certain sources of variation are controlled by the design of experiment.

1.8 Design of Experiment

The mode of collection of data in any controlled experiment is known as design of experiment. The objective of the controlled experiment is to collect information as an outcome of treatment,

where collected information are analysed to infer about the significance of the effects of treatments. The inference will be efficient if collected data and analysis of data are accurate.

Therefore, the accuracy in data collection is an important aspect of controlled experiment. The method of accurate data collection for analysis according to pre-determined objective is known as design of experiment.

The accuracy in data collection increases if the treatments are allocated to the plots by a random process and each treatment is repeated in several plots so that at least one yield per treatment is ensured. The treatments are also allocated to the plots of a block where plots of a block are homogeneous in character(s). The grouping of homogeneous plots, random allocation of treatment to the plots or to the plots of block and repetition of allocation of a treatment are the main aspects of design of experiment. Thus, it is understood that the method of data collection depends on allocation of treatment to the experimental plot. However, the following points are to be considered in conducting an experiment :

- (i) the treatments or factors to be used in the experiment,
- (ii) the effects of treatments or factors are to be estimated,
- (iii) the experimental plots to be used and the number of plots (units) to be used in the experiment,
- (iv) the method of allocation of treatment to the plots,
- (v) the mode of analysis.

The basic principles of experimental design when experiment is conducted considering all the above points are (a) randomization, (b) replication, and (c) local control.

Randomization : The treatments in experimental plots are allocated in such a way that no treatment is favoured or disfavoured during allocation. This is possible if treatments are allocated to the plots by any random process. The random allocation of treatments to the plots is known as randomization.

The analysis of the data becomes efficient if the error variance in the data is estimated efficiently. Randomization helps in estimation of error variance efficiently. Randomization also helps in avoiding the correlated observations. For example, let us consider that the effects of two varieties of rice are to be investigated. If the two rice varieties are allocated to the two neighbouring plots, the two yields will be probably correlated. However, only randomization is not sufficient to avoid the correlated observations and hence, correlated errors. If the treatments are allocated in many plots, there is less chance that the two treatments will always be allocated to the neighbouring plots. Therefore, along with randomization the replication will also be done as a mode of experiment.

Replication : By replication we mean the allocation of a treatment in several plots. The number of replication of a treatment is usually denoted by r if the treatment is allocated in r plots. The value of r is determined depending on the resources available for the experiment.

It has already been mentioned that the replication of the treatment helps in estimating the error variance efficiently. The efficiency of the experiment increases with the decrease in the error variance and replication helps in reducing the error variance. The value of r can be determined depending on the value of error variance.

Let us consider that the \bar{y}_i and \bar{y}_j are the means of i -th and j -th treatment respectively; s^2 is the error variance of the experiment in which these two treatments are randomly allocated to the plots. The objective of the researcher is to test the significance that there is an amount of difference d in the above treatments. Assume that each treatment is replicated in r plots.

Then the test statistic is

$$t = \frac{|\bar{y}_i - \bar{y}_j|}{\sqrt{\frac{2s^2}{r}}}$$

If d has a given value for which the difference in the treatments is significant at $\alpha\%$ level of significance, then the value of t is, say t_0 , where

$$t_0 = \frac{|d|}{\sqrt{\frac{2s^2}{r}}} \quad \text{or,} \quad r = \frac{2s^2 t_0^2}{d^2}$$

In large experiment or with known value of s^2 , t_0 transforms to the value of z (normal variate, where $z = 1.96$ at 5% level of significance). Here the difference d is considered significant at $\alpha\%$ level of significance.

The above method of estimation of r is used to estimate the value of r for a future experiment depending on the information of some existing results. For example, let us consider that in a varietal trial experiment 20 treatments are allocated to 100 plots, where each treatment is replicated 5 times. The error variance of this experiment is $s^2 = 30.26$. The means of treatment-1 and treatments 2 are 81.26 and 92.54, respectively. If the difference in the mean of these treatments is considered significant for a value of $d = 5$, then the value of r for any future experiment will be

$$r = \frac{2 \times 30.26 \times (1.96)^2}{25} \approx 9.$$

Here $t_0 = 1.96$ is considered assuming a large experiment of 20 treatments each replicated 5 times. However, this rule is not followed in estimating the replication per treatment in all experimental situations. To take the value of t_0 , the error d.f. should be considered, since smaller d.f. (< 10) does not provide stable F -test.

Local control : The act of making group of homogeneous plots for allocating a group treatments to those plots is known as local control. In dairy science experiment, selection of a group of cows or animals of same age or same size or same type is known as local control. In agricultural experiment, selection of a group of plots of the same soil fertility is known as local control. In irrigated experiment, the plots under different levels of irrigation must be of same height so that water is distributed uniformly in all plots. The local control is also called the control of external source of variation, since it controls, to some extent, the sources of errors in the experiment. Due to local control the error variance is reduced and hence, efficiency of the experiment is increased.

1.9 Conditions for Efficient Experiment

To conduct efficient experiment the following points are to be considered :

(i) **The experiment should be free of error :** It is possible if all the plots used in an experiment are homogeneous. The effect of treatment is entangled with the variation in the plots and hence heterogeneity in the plots will not reflect the real treatment effect.

(ii) **The error variance must be estimated :** The test of significance of treatment effect and any inference related to effect depend on the estimate of error variance. If \bar{y}_i and \bar{y}_j are the means of i -th and j -th treatment depending on n observations each, then $V(\bar{y}_i - \bar{y}_j) = \frac{2\sigma^2}{n}$, where σ^2 is the error variance. To study the efficiency of the difference in treatment mean, $V(\bar{y}_i - \bar{y}_j)$ must be estimated, where estimate depends on the estimate of σ^2 . Both randomization and replication help in obtaining estimate of σ^2 efficiently.

(iii) **Treatment effect should be estimated with precision** : The precision of treatment effect is estimated by the reciprocal of the variance of treatment effect. Hence, minimum error variance provides maximum precision of the treatment effect. The precision of the estimate increased with the increases in replication of treatment. The precision is also increased if precise design is used for the experiment.

(iv) **Scope of experiment should be well mentioned** : The experimental result is considered as the result of a sample experiment and it is predicted for the population. For example, let us consider that a varietal trial is performed in the field of agriculture using 5 doses of nitrogen as urea. These doses are used in an experiment in a particular area with an objective of predicting the suitability of those doses for all the areas. If the doses are suitable only for a particular area, the feasibility of experiment will not be beyond question. Therefore, the experiment is to be conducted in such a way that its result is acceptable to every situation.

(v) **The experiment should be simple** : The experimental technique should be simple and easy. The complexity in the experiment may create problem in the analysis and hence, create problem in estimating the error variance. However, the complex experiment is not prohibited completely. The simplest experimental techniques (designs) are completely randomized design, randomized block design and latin square design.

1.10 Estimation and Test of Hypothesis

The model considered for analysis of variance is $Y = XB + U$, where the elements of X are either 0 or 1 according to the absence or presence of a treatment in a plot. Let us consider that $r(X) < (k + 1) = p$. The problem is to estimate the parameter vector B and is to test the significance of this parameter vector. For this, the assumptions are :

$$(i) U \sim \text{NID}(0, \sigma^2 I), \quad (ii) E(UU') = \sigma^2 I, E(U) = 0.$$

The problem is to estimate $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ and σ^2 . The estimation of parameters is done using method of least squares, where

$$\hat{U}'\hat{U} = (Y - X\hat{B})'(Y - X\hat{B}) \quad \text{and} \quad \frac{\delta \hat{U}'\hat{U}}{\delta \hat{B}} = 0 \text{ gives the normal equation}$$

$$X'X\hat{B} = X'Y \quad \text{or,} \quad \hat{B} = (X'X)^{-1}X'Y.$$

Since $r(X) < k + 1$, $(X'X)^{-1}$ does not exist and the estimate of B is not available. However, if the $r(X'X)$ and $r(X'X|X'Y)$ are same, the normal equations are consistent and solution of the normal equation, though not unique, is available. Here the estimator of B will be a linear function of Y . The problem is to investigate a linear function of Y as an estimator of B . For this, let us consider a C matrix of order $(k + 1) \times n$, where the elements of C matrix are fixed and independent of B . We need to verify whether $E(CY) = B$ or not. If CY is the unbiased estimate of B , then $E(CY) = E[C(XB + U)] = CXB$. This implies that $CX = I$. But it is not possible, since the rank of C will be at best p whereas the rank of I is greater than $k + 1$. Therefore, there will be no linear function of Y as an estimator of B . At this stage, the problem is to investigate an estimator of any linear function of B_i 's. Let us now define some functions of the elements of B vector.

Parameter : The elements in the B vector are parameters since these are unknown constants for the population observations :

Parametric function : A linear combination of the parameters is known as parametric function. Thus, $\beta_1 - \beta_2, \beta_1 + \beta_2 - 2\beta_3, \beta_1 + \beta_2 + \beta_3 - 3\beta_4$ are parametric functions. In general, $\lambda'\beta$ is a parametric function, where $\lambda' = (\lambda_1, \lambda_2, \dots, \lambda_{k+1})$.

Estimable function : A parametric function is linearly estimable if it has an unbiased estimate and if the estimator is linear combination of observations.

Contrast : $\sum_{i=1}^{k+1} \alpha_i \beta_i$ is called a contrast if $\sum \alpha_i = 0$. Thus, $\beta_1 - \beta_2$ is a contrast, where $\alpha_1 = 1, \alpha_2 = -1$ and $\sum_{i=1}^2 \alpha_i = 0$. Similarly, $\beta_1 + \beta_2 - 2\beta_3, \beta_1 + \beta_2 + \beta_3 - 3\beta_4, \frac{1}{2}(\beta_1 - \beta_2)$ are all examples of contrast.

Let $C_1 = \sum l_i \beta_i$ and $C_2 = \sum d_i \beta_i$ are two contrasts. Then C_1 and C_2 are called *orthogonal contrasts* if $\sum l_i d_i = 0$. For example, $\beta_1 + \beta_2 - 2\beta_3$ and $\beta_1 + \beta_2 + \beta_3 - 3\beta_4$ are two orthogonal contrasts, since $l_1 = 1, l_2 = 1, l_3 = -2, l_4 = 0; d_1 = 1, d_2 = 1, d_3 = 1$ and $d_4 = -3$ and $\sum l_i d_i = 1 \times 1 + 1 \times 1 - 2 \times 1 + 0(-3) = 0$.

Theorem : If $\beta_1, \beta_2, \dots, \beta_k$ are parameters, then there will be $(k-1)$ orthogonal contrasts of these parameters.

Proof : Let $C_1 = \sum l_{1i} \beta_i, C_2 = \sum l_{2i} \beta_i, \dots, C_m = \sum l_{mi} \beta_i$ be m orthogonal contrasts of the parameters. Consider another contrast $C = \sum l_i \beta_i$, such that $\sum l_i = 0$ but l_i 's are unknown. The contrast C will be orthogonal to other contrasts C_1, C_2, \dots, C_m , if at least one of the following equations has non-zero solution :

$$\sum l_i = 0, \sum l_{1i} l_i = 0, \sum l_{2i} l_i = 0, \dots, \sum l_{mi} l_i = 0.$$

There are k unknowns in these equations. The non-zero solution of one of these unknowns is available if the number of equations does not exceed $(k-1)$. Therefore, the value of m will be at best $(k-2)$. This implies that the number of orthogonal contrasts cannot exceed $k-1$.

Theorem : The parametric function $\lambda' \beta$ defined on the model $Y = X\beta + U$ under assumption $E(U) = 0$ and $E(UU') = \sigma^2 I$ is estimable, if and only if there is only one solution in the equation $X'Xr = \lambda$, where λ is a vector of known constants.

Proof : We need to prove that there is only one solution of r in the equation $X'Xr = \lambda$ and in that case $E(b'Y) = X'\beta$, where b is a vector. If $\lambda' \beta$ estimable, $E(b'Y) = b'X\beta = \lambda' \beta$. This implies that $b'X = \lambda'$ or, $X'b = \lambda$. This is possible if $r(X') = r(X'/\lambda)$. Hence, $r(X'X) = r(X'X/\lambda)$ and r has a solution in the equation $X'Xr = \lambda$. Now, in case of $X'Xr = \lambda$ has a solution for r , it can be written as $X'(Xr) = \lambda$ or, $X'b = \lambda$, where $b = Xr$.

Theorem : The best linear unbiased estimator (BLUE) of the estimable function $\lambda' \beta$ in case of the model $Y = X\beta + U$ with $E(U) = 0$ and $E(UU') = \sigma^2 I$ is $r'X'Y$, where r is a solution of the equation $X'Xr = \lambda$.

Proof : Consider the best linear unbiased estimator of $\lambda' \beta$ is $b'Y$, where $b = r'X' + a'$ and a has any value. Thus, b is a general vector. The value of a will be such that (i) $E(b'Y) = \lambda' \beta$ and (ii) $V(b'Y)$ will be minimum among the variances of other functions of Y under (i).

If $b'Y$ is unbiased, $E(b'Y) = b'X\beta = (r'X'X + a'X)\beta = \lambda' \beta$. This is true if $a'X = 0$, since $X'Xr = \lambda$ has a solution.

$$\begin{aligned} \text{Again, } V(b'Y) &= E[b'Y - \lambda' \beta]^2 = E[b'Y - \lambda' \beta][b'Y - \lambda' \beta]' \\ &= E[b'X\beta + b'U - \lambda' \beta][b'X\beta + b'U - \lambda' \beta]' \\ &= E[r'X'X\beta + a'X\beta + b'U - \lambda' \beta][r'X'X\beta + a'X\beta + b'U - \lambda' \beta]' \end{aligned}$$

$$\begin{aligned}
 &= E[\lambda'\beta + b'U - \lambda'\beta][\lambda'\beta + b'U - \lambda'\beta]' = E[b'UU'b] \\
 &= \sigma^2bb' = \sigma^2[r'X' + a'] [Xr + a] \\
 &= \sigma^2r'X'Xr + \sigma^2a'a.
 \end{aligned}$$

Here $V(b'Y)$ will be minimum if $a'a$ is minimum. This value $a'a$ will be lowest if $a = 0$. Then $a'X = 0$. Therefore, $b' = r'X'$ and the best linear unbiased estimator of $\lambda'\beta$ is $r'X'Y$.

The above theorem is well known as Gauss-Markov Theorem of linear estimator.

Corollary : If $\lambda'\beta$ is an estimable function and if b is available in case of $\lambda' = b'X'$, then the unbiased estimator of $\lambda'\beta$ is $b'X'Y$.

Theorem : If $X'Xr = \lambda$ has a solution and $\lambda'\beta$ is an estimable function, then for any value of r , $\lambda'\beta$ is the same estimate.

Proof : The model considered is $Y = X\beta + U$ and $E(U) = 0$, $E(UU') = \sigma^2I$. Consider that r_1 and r_2 are two solutions of $X'Xr = \lambda$. We need to prove that $r_1'X'Y = r_2'X'Y = \lambda'\hat{\beta}$.

Let us consider that $\hat{\beta}$ is the solution of $X'X\hat{\beta} = X'Y$. Then we have $r_1'X'X\hat{\beta} = r_1'X'Y$ and $r_2'X'X\hat{\beta} = r_2'X'Y$. But $r_1'X'X = r_2'X'X = \lambda'$. Therefore, $\hat{\lambda}\beta = \lambda'\hat{\beta} = r_1'X'Y$ and $\hat{\lambda}\beta = r_2'X'Y = \lambda'\hat{\beta}$. Hence, the estimate of estimable function $\lambda'\beta$ is $\lambda'\hat{\beta}$.

Theorem : For the model $Y = X\beta + U$ with $E(U) = 0$, $E(UU') = \sigma^2I$, if the rank of X matrix is p , then there are exactly p independent linear estimable functions.

Proof : We need to show that there are exactly p linear independent vector λ such that λ is the solution of $X'Xr = \lambda$. Here p is the rank of X . If $X'Xr = \lambda$ has q vectors $\lambda_1, \lambda_2, \dots, \lambda_q$ ($q > p$), then these q vectors are available such that $X'Xr_1 = \lambda_1, X'Xr_2 = \lambda_2, \dots, X'Xr_q = \lambda_q$.

It can be written as

$$X'X(r_1, r_2, \dots, r_q) = (\lambda_1, \lambda_2, \dots, \lambda_q)$$

or, $X'XR = \Lambda$, where $R = (r_1, r_2, \dots, r_q)$ and $\Lambda = (\lambda_1, \lambda_2, \dots, \lambda_q)$

Here R is of order $(k+1) \times q$. However, the rank of $X'X$ is p . Therefore, the rank of Λ will be at best p . Thus, there will be p linear independent estimable functions.

Now, let us consider that X_i is the i -th row of X . For any value of i , $X_i'\beta$ will be estimable. We need to show that $X'Xr = X_i$ has a solution which is r . This solution is available if $r(X'X) = r((X'X|X_i))$ and in that case $X_i'\beta$ is estimable. Again, X_i 's form p independent vectors and hence, there will be p independent linear estimable functions.

Theorem : If $\lambda_1'\beta, \lambda_2'\beta, \dots, \lambda_q'\beta$ are estimable functions, the linear combination of those is also estimable.

Proof : Let us consider that $\lambda = \sum a_i\lambda_i$, where a_1, a_2, \dots, a_q are known values. We need to prove that $\lambda'\beta = \beta \sum a_i\lambda_i$ will be estimable, if $\lambda = \sum a_i\lambda_i$.

Consider that r_i is the solution of $X'Xr_i = \lambda_i$ and $r = \sum a_i r_i$. Then r is the solution of $X'Xr = \lambda$. Therefore, $\lambda'\beta$ is estimable, where $\lambda'\beta$ is the linear combination of $\lambda_1'\beta, \lambda_2'\beta, \dots, \lambda_q'\beta$.

Theorem : If $\lambda'\beta$ is the estimable function, λ' is the linear combination of row of X .

Proof : Let us consider that $\psi = a'Y$ is the unbiased estimator of $\lambda'\beta$. Then $E(\psi) = E(a'Y) = a'X\beta = \lambda'\beta$. Then $\lambda' = a'X$ and it is the linear combination of the rows of X -matrix.

Definition : If $r'_i X'Y$ is the best linear unbiased estimator of estimable function $\lambda'_i \beta$, then $\sum a_i r'_i X'Y$ is the best linear unbiased estimator of $\sum a_i \lambda'_i \beta$.

Definition : The best linear unbiased estimator of each estimable function is the linear combination $X'Y = \lambda$, where $X'Y$ is the right side of the normal equation $X'X\beta = X'Y$.

Definition : The linear function of the observations will be in the error space, if and only if its expected value is zero for any value of β .

If $b'Y$ is in the error space, then $E(b'Y) = b'X\beta = 0$. This implies that $b'X = 0$ or, $X'b = 0$. Thus, b is orthogonal to the columns of X .

Theorem : If the best linear unbiased estimator is expressed through observation vector, then the coefficient of this best linear unbiased estimator and the coefficient of the linear function of observations will be orthogonal.

Proof : If $b'Y$ is in the error space, then b is orthogonal to the columns of X . Again, if $\lambda' \beta$ is the estimable function, λ' is the linear combination of rows of X . Therefore, the coefficient of error space is orthogonal to the coefficient of the best linear unbiased estimator.

Theorem : The covariance of any linear function of error space and the best linear unbiased estimator is zero.

Proof : Let $b'Y$ is in the error space and $\lambda' \hat{\beta}$ is the best linear unbiased estimator. Then.

$$\begin{aligned} \text{Cov}(b'Y, X' \hat{\beta}) &= \text{Cov}[b'Y, \lambda'(X'X)^{-1}X'Y] \\ &= b'\lambda'(X'X)^{-1}X'\sigma^2, \quad \because V(Y) = \sigma^2 I \\ &= b'X(X'X)^{-1}\sigma^2 = 0, \quad \because b'X = 0. \end{aligned}$$

Theorem : If $\lambda'_1 \beta$ and $\lambda'_2 \beta$ are two estimable functions, then the variances of their best linear unbiased estimators are $\sigma^2 r'_1 X'X r_1$ and $\sigma^2 r'_2 X'X r_2$, respectively.

Proof : Let us consider the model $Y = X\beta + U$. Assumptions for U vector are $E(U) = 0$, $E(UU') = \sigma^2 I$. Consider that r_1 and r_2 are the solutions of the equations $X'X r_1 = \lambda_1$ and $X'X r_2 = \lambda_2$, respectively. Then

$$\begin{aligned} \text{Cov}[\lambda'_1 \hat{\beta}, \lambda'_2 \hat{\beta}] &= E[\lambda'_1 \hat{\beta} - \lambda_1 \beta][\lambda'_2 \hat{\beta} - \lambda_2 \beta]' \\ &= E[r'_1 X'Y - \lambda'_1 \beta][r'_2 X'Y - \lambda'_2 \beta]'. \quad \because \lambda'_1 \hat{\beta} = r'_1 X'Y \text{ and } \lambda'_2 \hat{\beta} = r'_2 X'Y \\ &= E[r'_1 X'U][U'X r_2] = \sigma^2 r'_1 X'X r_2. \end{aligned}$$

To find the variance of the estimators, let us consider $\lambda_1 = \lambda_2$. Then $r_1 = r_2$ and then

$$\text{Cov}[\lambda'_1 \hat{\beta}, \lambda'_2 \hat{\beta}] = \sigma^2 r'_1 X'X r_1 = \sigma^2 r'_2 X'X r_2.$$

Example 1.1 : For the following models :

$Y_1 = \beta_1 + \beta_2 + u_1$, $Y_2 = \beta_1 + \beta_3 + u_2$, $Y_3 = \beta_1 + \beta_2 + u_3$, $\lambda_1 \beta_1 + \lambda_2 \beta_2 + \lambda_3 \beta_3$ is an estimable function, if $\lambda_1 = \lambda_2 + \lambda_3$.

Solution : Let us consider a linear function $a_1 Y_1 + a_2 Y_2 + a_3 Y_3$.

Its expected value is assumed to be $\lambda_1 \beta_1 + \lambda_2 \beta_2 + \lambda_3 \beta_3$. Then

$$\begin{aligned} E(a_1 Y_1 + a_2 Y_2 + a_3 Y_3) &= a_1(\beta_1 + \beta_2) + a_2(\beta_1 + \beta_3) + a_3(\beta_1 + \beta_2), \quad \because E(U_i) = 0 \\ &= \beta_1(a_1 + a_2 + a_3) + \beta_2(a_1 + a_3) + \beta_3(a_2). \end{aligned}$$

Since $E(a_1 Y_1 + a_2 Y_2 + a_3 Y_3) = \lambda_1 \beta_1 + \lambda_2 \beta_2 + \lambda_3 \beta_3$,

we have $a_1 + a_2 + a_3 = \lambda_1$, $a_1 + a_3 = \lambda_2$ and $a_2 = \lambda_3 \Rightarrow \lambda_1 = \lambda_2 + \lambda_3$.

Thus, if $\lambda_1 = \lambda_2 + \lambda_3$, $E\left(\sum_{i=1}^3 a_i Y_i\right) = \sum_{i=1}^3 \lambda_i \beta_i$.

Hence, $\sum_{i=1}^3 \lambda_i \beta_i$ is an estimable function, where the estimator is $\sum_{i=1}^3 a_i Y_i$.

Reparameterization of the model : It has already been mentioned that the experimental design model $Y = XB + U$ is not of full rank model. The rank of X matrix is $p < k + 1$. The normal equations to estimate B vector are :

$$X'X\hat{B} = X'Y \quad \text{or,} \quad \hat{B} = (X'X)^{-1}X'Y.$$

But $(X'X)^{-1}$ does not exist and there is no unique solution \hat{B} . Therefore, to estimate B the parameter vector must be reparameterized so that X matrix is of full rank.

Definition : By reparameterization of the model $Y = XB + U$ we mean the transformation of B vector into $\alpha = VB$ vector so that the elements of $\alpha = VB$ are estimable functions.

The matrix $X'X$ is positive semi-definite with rank p . Therefore, we can have a non-singular matrix W^* ($k \times k$) so that

$$W^*(X'X)W^* = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix},$$

where A is a $p \times p$ matrix, the rank of which is p . The W^* matrix can be partitioned into W and W_1 such that $W^* = (W, W_1)$, where W is of order $k \times p$. We can write :

$$\begin{pmatrix} W' \\ W_1' \end{pmatrix} X'X(W, W_1) = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}.$$

Thus, we have $W'X'XW = A$, $W_1'X'XW_1 = 0$.

Therefore, $r(W'X') = p$ and $W_1'X' = 0$.

Now, we can write : $Y = XW^*(W^*)^{-1}B + U$.

Let $(W^*)^{-1} = \begin{pmatrix} V \\ V_1 \end{pmatrix}$.

$$\begin{aligned} \text{Then } Y &= X(W, W_1) \begin{pmatrix} V \\ V_1 \end{pmatrix} B + U = XW(VB) + (XW_1)(V_1B) + U \\ &= XW(VB) + U, \quad \because XW_1 = 0 \\ &= Z\alpha + U, \quad \text{where } XW = Z, \alpha = VB. \end{aligned}$$

Here Z is a matrix of order $n \times p$ and $r(Z) = p$. Thus, the X matrix is transformed to a Z matrix of full rank and B vector is reparameterized to a vector of parameters α .

The normal equations of the new model are $Z'Z\hat{\alpha} = Z'Y$ and $\hat{\alpha} = (Z'Z)^{-1}Z'Y$.

Since Z is of full rank, $\hat{\alpha}$ is available.

If $Z'Z$ is a diagonal matrix, the reparameterization of $Y = XB + U$ is known as orthogonal reparameterization. If $\lambda'B$ is estimable, the same linear estimate of $\lambda'B$ is available from the reparameterized model.

Example 1.2 : Estimate the parametric function after reparameterization of the model :

$$y_{i1} = \mu + \beta_i + e_{i1}, \quad i = 1, 2 \quad \text{where } E(e_{i1}) = 0.$$

Solution : Since $E(e_{i1}) = 0$,

$$E(y_{11}) = \mu + \beta_1 \quad \text{and} \quad E(y_{21}) = \mu + \beta_2.$$

Therefore, $\mu + \beta_1$ and $\mu + \beta_2$ are estimable functions, we can write,

$$\mu + \beta_1 = (1, 1, 0) \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \lambda'_1 \beta \quad \text{and} \quad \mu + \beta_2 = (1, 0, 1) \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \lambda'_2 \beta.$$

Here, λ_1 and λ_2 are independent and hence, $\lambda'_1 \beta$ and $\lambda'_2 \beta$ are also independent. Let us write :

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \end{pmatrix} = V\beta.$$

Now, to write $Y = Z\alpha + U$, let us write :

$$W^* = (W, W_1) = (V^*)^{-1}, \quad \text{where } V^* = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \quad \text{since } V = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

We need to find V_1 so that V^* is non-singular. If $V_1 = (0, 1, 1)$, then V^* is non-singular.

$$\text{Therefore, } V^* = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad (V^*)^{-1} = W^* = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

$$\text{Hence, } W = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad Z = XW = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$$\text{and} \quad Z'Z\hat{\alpha} = Z'Y \text{ gives } \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

$$\therefore \hat{\alpha}_1 = \frac{Y_1}{3}, \quad \hat{\alpha}_2 = \frac{Y_2}{3}.$$

Thus, α_1 and α_2 are estimable. We have $\alpha_1 - \alpha_2 = \beta_1 - \beta_2 = \frac{1}{3}(Y_1 - Y_2)$.

Theorem : If $Y = Z\alpha + U$ is the orthogonal reparameterized form of the model : $Y = XB + U$, then the elements in α are uncorrelated.

Proof : For the model : $Y = Z\alpha + U$, the estimates of α are $\hat{\alpha} = (Z'Z)^{-1}Z'Y = D_1'Z'Y$,
 $\therefore Y = Z\alpha + U$ is the orthogonal reparameterized form of $Y = X\beta + U$. Here, D_1 is diagonal.
 We have

$$\begin{aligned} \text{Cov}(\hat{\alpha}) &= E(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)' = E[(Z'Z)^{-1}Z'Y - \alpha][(Z'Z)^{-1}Z'Y - \alpha]' \\ &= E[(Z'Z)^{-1}Z'(Z\alpha + U) - \alpha][(Z'Z)^{-1}Z'(Z\alpha + U) - \alpha]' \\ &= E[(Z'Z)^{-1}Z'U][U'Z(Z'Z)^{-1}] \\ &= \sigma^2(Z'Z)^{-1} = \sigma^2 D_1^{-1}. \end{aligned}$$

$\therefore \text{Cov}(\hat{\alpha}_i, \hat{\alpha}_j) = 0$, for $i \neq j$.

Example 1.3 : Estimate the parameters μ and α_i in the model :

$$y_{ij} = \mu + \alpha_i + e_{ij}; \quad i = 1, 2; \quad j = 1, 2.$$

Solution : We can write $Y = X\beta + U$,

$$\text{where } Y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad U = \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \end{bmatrix}.$$

The normal equations are : $X'X\hat{\beta} = X'Y$

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{pmatrix}$$

$$\begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{pmatrix}, \quad \text{where } y_{..} = \sum_{i=1}^2 \sum_{j=1}^2 y_{ij}, \quad y_{1.} = \sum_j y_{1j}, \quad y_{2.} = \sum_j y_{2j}.$$

This gives $y_{..} = 4\hat{\mu} + 2\hat{\alpha}_1 + 2\hat{\alpha}_2$; $y_{1.} = 2\hat{\mu} + 2\hat{\alpha}_1$; $y_{2.} = 2\hat{\mu} + 2\hat{\alpha}_2$.

We can write : $y_{..} = 4\hat{\mu} + 2\sum \hat{\alpha}_i$; $y_{i.} = 2\hat{\mu} + 2\hat{\alpha}_i$, $i = 1, 2$.

These normal equations are not independent since $r(X) = 2$, where X is a matrix of order 4×3 .

To get the solution of the above equations, the model $Y = XB + U$ can be reparameterized. Alternatively, since the rank of the X matrix is less by one than the number of columns, one restriction can be put to get the unique solution of the normal equations. The restriction is $\sum \hat{\alpha}_i = 0$. Under this restriction the estimates are

$$\hat{\mu} = \bar{y}_{..} \quad \text{and} \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$$

Definition : The experimental design model is : $Y = XB + U$.

Assumption : (i) $E(U) = 0$, (ii) $E(UU') = \sigma^2 I$. Also U is normally distributed.

Under this assumption Y is also normally distributed. Hence, linear function of Y follows joint multivariate normal distribution, where the parameters of this distribution are the mean, the variance and the covariance of the linear functions of Y .

Let $b'_i Y$ ($i = 1, 2, \dots, n - k$) are the $(n - k)$ linear functions of Y and these are normally distributed with mean zero and variance σ^2 , where $b'_i Y$ are considered in the error space. Then

$$\frac{SS(\text{error})}{\sigma^2} = \sum_{i=1}^{n-k} \frac{(b'_i Y)^2}{\sigma^2}$$

follows χ^2 -distribution with $(n - k)$ d.f.

Theorem : The sum of squares of error is independently distributed of the best linear unbiased estimator of any estimable function.

Proof : We know that $\lambda' \hat{\beta}$ is the best linear unbiased estimator of the estimable function $\lambda' \beta$. Again, $b'_i Y$ ($i = 1, 2, \dots, (n - k)$) is in error space. These two quantities $\lambda' \hat{\beta}$ and $b'_i Y$ follow multivariate distribution. We also know that $\text{Cov}(\lambda' \hat{\beta}, b'_i Y) = 0$. Thus, $\lambda' \hat{\beta}$ and $b'_i Y$ are independently distributed. We have

$$SS(\text{error}) = \sum_{i=1}^{n-k} \frac{(b'_i Y)^2}{\sigma^2}.$$

Therefore, $\lambda' \hat{\beta}$ and $SS(\text{error})$ is independently distributed.

Definition : The distribution of best linear unbiased estimators of any m linear estimable functions $\Lambda\beta$ is multivariate normal distribution with mean $\Lambda\beta$ and variance-covariance matrix $\Lambda(X'X)^{-1}\Lambda'\sigma^2$, where Λ is a matrix of order $m \times p$ with rank m . This distribution is independent of the distribution of error sum of squares.

Theorem : The distribution of $(\Lambda\hat{\beta} - \Lambda\beta)'[\Lambda(X'X)^{-1}\Lambda'\sigma^2]^{-1}(\Lambda\hat{\beta} - \Lambda\beta)$ is chi-square with m d.f.

Proof: We know that $\Lambda(X'X)^{-1}\Lambda'$ matrix is non-singular symmetric matrix of order $m \times m$. Its eigenvalues are positive. Let the eigenvalues be l_1, l_2, \dots, l_m . Then an orthogonal matrix A is available such that

$$\Lambda(X'X)^{-1}\Lambda' = A \text{diag} (l_1, l_2, \dots, l_m)A'$$

Also, we have

$$[\Lambda(X'X)^{-1}\Lambda']^{\frac{1}{2}} = A \text{diag} \left(l_1^{\frac{1}{2}}, l_2^{\frac{1}{2}}, \dots, l_m^{\frac{1}{2}} \right) A'$$

and
$$[\Lambda(X'X)^{-1}\Lambda']^{-\frac{1}{2}} = A \text{diag} \left(\frac{1}{l_1^{\frac{1}{2}}}, \frac{1}{l_2^{\frac{1}{2}}}, \dots, \frac{1}{l_m^{\frac{1}{2}}} \right) A'$$

The unbiased estimates of $\Lambda\beta$ is $\Lambda\hat{\beta}$, where $\Lambda\hat{\beta}$ is the linear combination of m normal variables. The joint distribution of this linear combination is also normal with mean zero. Let us write :

$$Z = [\Lambda(X'X)^{-1}\Lambda']^{-\frac{1}{2}}(\Lambda\hat{\beta} - \Lambda\beta).$$

Then $E(Z) = 0$, $\therefore E(\Lambda\hat{\beta}) = \Lambda\beta$.

$$\begin{aligned} V(Z) &= [\Lambda(X'X)^{-1}\Lambda']^{-\frac{1}{2}} V(\Lambda\hat{\beta})[\Lambda(X'X)^{-1}\Lambda']^{-\frac{1}{2}} \\ &= [\Lambda(X'X)^{-1}\Lambda']^{-\frac{1}{2}} [\Lambda(X'X)^{-1}\Lambda'\sigma^2][\Lambda(X'X)^{-1}\Lambda']^{-\frac{1}{2}} \\ &= \sigma^2 I. \end{aligned}$$

Therefore, $Z \sim \text{NID}(0, \sigma^2)$ and $\frac{1}{\sigma^2}Z'Z = (\Lambda\hat{\beta} - \Lambda\beta)'[\Lambda(X'X)^{-1}\Lambda'\sigma^2]^{-1}(\Lambda\hat{\beta} - \Lambda\beta)$ is distributed as chi-square with m d.f.

When $m = 1$, $\Lambda\beta$ transforms to one parametric function $\lambda'\beta$. In that case,

$$Z = (\lambda'\hat{\beta} - \lambda'\beta) \{ \lambda'(X'X)^{-1}\lambda \}^{-\frac{1}{2}} \sim N(0, \sigma^2)$$

and $(\lambda'\hat{\beta} - \lambda'\beta)^2 / [\lambda'(X'X)^{-1}\lambda\sigma^2]$ is distributed as chi-square with 1 degree of freedom.

Theorem : For the model $Y = X\beta + U$, when $U \sim \text{NID}(0, \sigma^2 I)$,

$$F = \frac{(\Lambda\hat{\beta} - \Lambda\beta)'[\Lambda(X'X)^{-1}\Lambda']^{-1}(\Lambda\hat{\beta} - \Lambda\beta)/m}{SS(\text{error})/(n-k)}$$

has variance ratio distribution with m and $(n-k)$ d.f.

Proof: We have already proved that the numerator of F is distributed as χ^2 with m d.f. Also we showed that $SS(\text{error})$ is distributed as χ^2 with $(n-k)$ d.f. Both numerator and denominator are independent. Therefore, F has variance ratio distribution with m and $(n-k)$ d.f.

When $m = 1$, $(\lambda'\hat{\beta} - \lambda'\beta) / [SS(\text{error})\{\lambda'(X'X)^{-1}\lambda\}/(n-k)]^{\frac{1}{2}}$ is distributed as Student's t with $(n-k)$ d.f.

Let us consider that $\Lambda\beta$ has a specific value, say d . Then, we can define :

$$W = (\Lambda\hat{\beta} - d)' \{ \Lambda(X'X)^{-1}\Lambda'\sigma^2 \}^{-1} (\Lambda\hat{\beta} - d).$$

Here d is a $m \times 1$ vector of fixed quantity. Let us consider $v = [\wedge(X'X)^{-1}\wedge']^{-\frac{1}{2}}(\wedge\hat{\beta} - d)$, where v indicates the m linear functions of $\wedge\hat{\beta}$. The joint distribution of v is m -variate normal and

$$\begin{aligned} E(v) &= \mu \text{ (say), where } \mu = [\wedge(X'X)^{-1}\wedge']^{-\frac{1}{2}}(\wedge\beta - d). \\ V(v) &= [\wedge(X'X)^{-1}\wedge']^{-\frac{1}{2}} \vee (\wedge\hat{\beta} - d)[\wedge(X'X)^{-1}\wedge']^{-\frac{1}{2}} \\ &= [\wedge(X'X)^{-1}\wedge']^{-\frac{1}{2}}[\wedge(X'X)^{-1}\wedge'\sigma^2][\wedge(X'X)^{-1}\wedge']^{-\frac{1}{2}} \\ &= \sigma^2 I. \end{aligned}$$

Let us consider that the elements in v are v_1, v_2, \dots, v_m , where each one is independently distributed with $\mu_1, \mu_2, \dots, \mu_m$, respectively and with common variance σ^2 . Therefore,

$$\frac{v'v}{\sigma^2} = \frac{1}{\sigma^2}(v_1^2 + v_2^2 + \dots + v_m^2)$$

is distributed as non-central Chi-square with m d.f. The non-centrality parameter is

$$\delta^2 = \frac{\mu'\mu}{\sigma^2} = \frac{1}{\sigma^2} (\mu_1^2 + \mu_2^2 + \dots + \mu_m^2).$$

Therefore, W is distributed as non-central chi-square with m d.f. But, if $\wedge\beta - d = 0$, $\delta^2 = 0$, then W is distributed as central chi-square.

Definition : Let $Y = X\beta + U$, where $E(U) = 0$, $E(UU') = \sigma^2 I$ and U is normally distributed. Then the distribution of $SS(\hat{\beta})/\sigma^2$ is distributed as non-central chi-square with p d.f. and with noncentrality parameter $\beta'(X'X)\beta/p\sigma^2$, where p is the rank of X matrix. Again, $SS(\hat{\beta})$ and $SS(\text{error})$ are independently distributed.

Cochran's Theorem : Let x_1, x_2, \dots, x_n be n standard normal variates. Then, $\sum x_i^2 = X'IX$, where $X' = (x_1, x_2, \dots, x_n)$ can be decomposed into k quadratic forms $Q_i = X'A_iX$, $i = 1, 2, \dots, k$. Thus,

$$X'IX = \sum_{i=1}^k Q_i = \sum_{i=1}^k X'A_iX, \text{ where } A_i \text{ has rank } n_i.$$

In such a case (i) each of Q_i is distributed as chi-square with n_i d.f., and (ii) all Q_i 's are independently distributed, if and only if $n_1 + n_2 + \dots + n_k = n$. Here n is the rank of I_n matrix.

Proof : The first part of the theorem implies second part, since $Q = \sum Q_i$ and each Q_i is distributed as chi-square with n_i d.f. Hence, Q is distributed as chi-square with $n_1 + n_2 + \dots + n_k$ d.f. Again, $\Sigma Q_i = X'IX$ is distributed as χ^2 with n d.f. Hence, $n = \Sigma^k n_i$.

Second part implies first part. We have $Q_i = X'A_iX$. Let us consider an orthogonal transformation $X = BY$, which transforms A_i to diagonal form. Also, consider that

$$X'B'A_iBX + X'B'(I - A_i)BX = X'IX.$$

The right side and the first component of the left side of the equation is diagonal. Hence, the second component of the left side is also diagonal. Since A_i has rank n_i , the $n - n_i$ principal diagonal elements of $B'A_iB$ are zero and the corresponding diagonal elements of $B'(I - A_i)B$ are 1. From the given condition, the rank of $B'(I - A_i)B$ is $n - n_i$. Hence, the $n - n_i$ principal diagonal elements of $B'(I - A_i)B$ are 1 and the corresponding diagonal elements of $B'A_iB$ are zero. The remaining diagonal elements of $B'(I - A_i)B$ are zero and the corresponding diagonal elements of $B'A_iB$ are 1. Therefore, A_i is an idempotent matrix which has n_i eigenvalues 1 and the remainings are 0. Hence, Q_i is distributed as χ^2 with n_i d.f.

In analysis of variance, we partitioned the total sum of squares into component sum of squares according to pre-identified sources of variations. The d.f. of total sum of squares is equal to the d.f. of the different components' sum of squares. By Cochran's theorem all the component sum of squares are independently distributed as chi-square. Except error sum of squares all the sum of squares follow non-central chi-square distribution. Hence, to test the significance of the parametric function mean sum of squares due to the estimate of parametric function is compared with the mean sum of squares due to error.

1.11 Multiple Comparison

The general linear model for analysis of variance is

$$y_{ij} = \beta_0 + \beta_i x_{ij} + e_{ij}; \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, n_i,$$

where $x_{ij} = 0$ or 1. The assumption is that $e_{ij} \sim \text{NID}|0, \sigma^2$. The objective of the analysis is to test the significance of the hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k, \text{ against}$$

H_A : at least one of the equalities does not hold good.

The test statistic for this hypothesis or any hypothesis related to the estimable function is F . If the hypothesis is rejected by F -test, we need to test the significance of the hypothesis

$$H_0 : \beta_i = \beta_l, \text{ against}$$

$$H_A : \beta_i \neq \beta_l, \quad i \neq l = 1, 2, \dots, k$$

This pairwise comparison of the impacts of factors is known as multiple comparison.

The methods of multiple comparison are (i) Student's t test, (ii) Student-Newman-Keuls test, (iii) Duncan's Multiple Range test, (iv) Dunnett's test.

Student's t -test : Let $\hat{\beta}_i$ be the estimates of β_i and $V(\hat{\beta}_i) = \frac{s^2}{n_i}$, where s^2 is the error mean square. The null hypothesis $H_0 : \beta_i = \beta_l$ or $H_0 : \beta_i - \beta_l = 0$. Under this hypothesis the estimate of $\beta_i - \beta_l$ is $\hat{\beta}_i - \hat{\beta}_l$. The estimated variance of this estimate is

$$v(\hat{\beta}_i - \hat{\beta}_l) = s^2 \left(\frac{1}{n_i} + \frac{1}{n_l} \right).$$

Therefore, under H_0 the test statistic is

$$t = \frac{\hat{\beta}_i - \hat{\beta}_l}{\sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_l} \right)}}.$$

This t follows student's t distribution with error d.f. Hence any pair of impacts β_i and β_l ($i \neq l = 1, 2, \dots, k$) can be compared using ' t '-test.

Let us consider that the hypothesis is to be tested at $\alpha\%$ level of significance, where the t -value at $\alpha\%$ level of significance with error d.f. is, say, t_0 . Also, consider that all n_i 's are same ($n_1 = n_2 = \dots = n_k = n$). Then, using t_0 , we get

$$t_0 \sqrt{\frac{2s^2}{n}} = |\hat{\beta}_i - \hat{\beta}_l|.$$

Here $t_0 \sqrt{\frac{2s^2}{n}}$ is known as critical difference (C.D.) or Least Significant Difference (L.S.D.).

Thus,

$$\text{C.D.} = t_0 \sqrt{\frac{2s^2}{n}}.$$

Now, if any of the $|\hat{\beta}_i - \hat{\beta}_l| \geq \text{C.D.}$, the null hypothesis is rejected. If n_i 's are not equal, the value of n is to be replaced by the harmonic mean of n_1, n_2, \dots, n_k .

The basic assumption for student's t test is that the observations of i -th factor must be independent of the observations of l -th factor. Therefore, if β_i and β_l are to be compared in any experiment, the i -th and l -th factor (treatment) are to be allocated in the plots so that the yields of these two treatments are independent. Otherwise the t -test will be affected. The comparison of the highest and the lowest yielding treatments can also be compared using C.D. Pearson and Hartley (1942, 1943) showed that the first kind of error in such a test becomes more than 5%.

The problem that arises in comparing the treatment means which are not independent is avoided by comparing the range. Such range tests are Student-Newman-Keuls test and Duncan's multiple range test.

Student-Newman-Keuls test : Let $\bar{x}_1, < \bar{x}_2 < \dots < \bar{x}_k$ be k means related to k treatments. The estimated variance of i -th mean is s^2/n_i , where s^2 is the error mean square.

The Studentized range based on k means is defined by

$$q_{k,f} = \frac{(\bar{x}_k - \bar{x}_1)\sqrt{n}}{s^2},$$

where f is the d.f. related to s^2 . The value of $q_{k,f}$ for different values of α are available in different books [Federer (1955), Winner (1971)]. If \bar{x}_i is based on n_i observations, then n is to be replaced by the harmonic mean of n_1, n_2, \dots, n_k . The value of $q_{k,f}$ is given for different values of k .

The Studentized critical values is given by

$$W_i = q_{\alpha,i,f} \sqrt{\frac{s^2}{n}}, \quad i = 2, 3, \dots, k,$$

where $q_{\alpha,i,f}$ is the value of Studentized range at $\alpha\%$ level of significance for range of i means with f d.f.

Test procedure : Let the range of k means be $\bar{x}_k - \bar{x}_1$. This range is to be compared with W_k . If $\bar{x}_k - \bar{x}_1 \geq W_k$, significant difference in the means is noted. At this stage we need to calculate W_{k-1} and it is to be compared with $\bar{x}_k - \bar{x}_2$ and $\bar{x}_{k-1} - \bar{x}_1$. If $\bar{x}_k - \bar{x}_2 \geq W_{k-1}$ and $\bar{x}_{k-1} - \bar{x}_1 \geq W_{k-1}$, the ranges of $(k-1)$ means are significantly different. The process is continued until an observed range of k means is found smaller than W_k . Finally, the means which are not found significantly different are shown in one group.

Duncan's multiple range test : This test is similar to that of Student-Newman-Keuls test except that the tabulated value $q_{\alpha,i,f}$ is replaced by its modified value as suggested by Duncan (1945). Here the ranges of means for different values of k are compared with

$$D_i = d_{\alpha,i,f} \sqrt{\frac{s^2}{n}}.$$

The test procedure remains same as it is followed in Student-Newman-Keuls test.

Example 1.4 : The systolic blood pressure (in mm of Hg) of some patients admitted in 4 different hospitals are as follows

Hospital	Systolic blood pressure of patients (y_{ij})	Total y_i	Means \bar{y}_i
1	80, 85, 86, 90, 95, 98, 82, 80, 87, 95, 100	978	88.91
2	85, 88, 90, 90, 92, 96, 101, 103, 115, 100, 100	1060	96.36
3	75, 80, 82, 86, 85, 85, 85, 85, 85, 80, 90	918	83.45
4	90, 96, 80, 98, 90, 95, 96, 95, 100, 98, 110	1048	95.27

Identify the hospitals where patients are experienced of lower blood pressure.

Solution : Here $k = 4$, $n_1 = n_2 = n_3 = n_4 = 11$. According to technique of analysis of variance, we have

$$y.. = G = 4004, C.T. = \frac{G^2}{nk} = \frac{(4004)^2}{11 \times 4} = 364364.$$

$$SS(\text{total}) = \sum \sum (y_{ij} - \bar{y}..)^2 = \sum \sum y_{ij}^2 - C.T. = 368337 - 364364 = 3973.$$

$$SS(\text{between}) = n \sum (\bar{y}_i - \bar{y}..)^2 = \sum \frac{y_i^2}{n} - C.T. = 365555.64 - 364364 = 1191.64.$$

$$SS(\text{error}) = SS(\text{total}) - SS(\text{between}) = 3973 - 1191.64 = 2781.36.$$

ANOVA Table

Sources of variation:	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	P -value
Hospital (between)	3	1191.64	397.21	5.71	2.84	0.00
Error (within)	40	2781.36	69.534			
Total	43					

The null hypothesis to be tested is

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$, against H_A : At least one of the equality does not hold good.

Here β_i is the impact of i -th hospital ($i = 1, 2, 3, 4$). Since F is greater than $F_{0.05}$, H_0 is rejected. The recorded blood pressure of patients in different hospitals are significantly different.

Here P -value = $\int_F^\infty f(F) dF$. If P -value ≤ 0.05 , H_0 is rejected.

Now the hospitals can be grouped according to the average blood pressure of patients. The mean blood pressure of patients per hospitals in ascending order are as follows :

$$\bar{y}_3 = 83.45, \bar{y}_1 = 88.91, \bar{y}_4 = 95.27, \bar{y}_2 = 96.36.$$

We have

i	$d_{\alpha, i, f}$	$D_i = d_{\alpha, i, f} \sqrt{\frac{s^2}{n}}$
2	2.86	7.19
3	3.01	7.57
4	3.10	7.79

Now, $\bar{y}_2 - \bar{y}_3 = 96.36 - 83.45 = 12.91 > D_4$

$\bar{y}_2 - \bar{y}_1 = 96.36 - 88.91 = 7.45 < D_3$

$$\begin{aligned} \bar{y}_4 - \bar{y}_3 &= 95.27 - 83.45 = 11.82 > D_3 \\ \bar{y}_4 - \bar{y}_1 &= 95.27 - 88.91 = 6.36 < D_2 \\ \bar{y}_1 - \bar{y}_3 &= 88.91 - 83.45 = 5.46 < D_2. \end{aligned}$$

The hospitals which are not different in respect of mean blood pressure are shown below by putting underline

$$H_3, \underline{H_1}, \underline{H_4}, H_2$$

For multiple comparison we use Student's *t* test, Student-Newman-Keuls test and Duncan's multiple range test. The last one is the modified one and it is improved test over Student-Newman-Keuls test. However, Montgomery (1984) has mentioned that *t* test is still better than other tests.

Dunnnett's test : The objective of the control experiment is to investigate the treatment effects to identify the best one. To identify the best one, sometimes a treatment with known result is also used in the experiment. Such a treatment is known as control treatment. The objective is to identify the superiority of new treatments over the old one. This comparison of treatment means with the mean of old one (or one with known result) is also under multiple comparison. It is usually known as comparison of treatments with a control treatment.

The comparison is done by Dunnnett's (1984) test, where tabulated value for multiple comparison is used from Dunnnett's table. The test statistic is :

$$D = d_{\alpha,k-1,f} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_c} \right)}$$

where $d_{\alpha,k-1,f}$ is the tabulated value at $\alpha\%$ level of significance from Dunnnett's table for $(k - 1)$ means with error d.f. (f); n_c is the number of replications for control treatment and n_i is the number of replications of i -th treatment ($i = 1, 2, \dots, k$) except control treatment. If $n_i = n_c = n$, then the test statistic is :

$$D = d_{\alpha,k-1,f} \sqrt{\frac{2s^2}{n}}$$

Here s^2 is the mean square error.

Example 1.5 : In a dairy farm 4 new varieties of dry food are introduced for the milking cows. The objective of introduction of these food is to get increased amount of milk. These varieties of food are given to cows of same age and of same lactation period. During experiment of feeding trial, a group of cows are kept with the experimental cows. The milk productions of one day during experiment are recorded for all the cows. The milk productions are shown below :

Food	Milk production (in kg) of cows	Total y_i	Mean \bar{y}_i
No	18.5, 18.6, 22.7, 25.2, 26.0, 20.4	131.4	21.90
1	17.2, 19.7, 23.4, 22.6, 24.0, 25.2	132.1	22.02
2	20.2, 21.4, 24.6, 26.7, 28.2, 26.0	147.1	24.52
3	24.2, 28.6, 27.3, 30.2, 30.0, 30.1	170.4	28.40
4	25.0, 18.6, 24.7, 28.9, 29.0, 24.2	150.4	25.07

Do you think that the new varieties of food are effective in increasing the milk production?

Solution : We have $k = 5$, $n_1 = n_2 = n_3 = n_4 = n_5 = n = 6$.

$$G = \sum \sum y_{ij} = 731.4, \quad \text{C.T.} = \frac{G^2}{nk} = \frac{(731.4)^2}{6 \times 5} = 17831.532.$$

$$SS(\text{food}) = \sum \frac{y_i^2}{n} - \text{C.T.} = \frac{108011.1}{6} - 17831.532 = 170.318.$$

$$SS(\text{total}) = \sum \sum y_{ij}^2 - \text{C.T.} = 18249.12 - 17831.532 = 417.588.$$

$$SS(\text{error}) = SS(\text{total}) - SS(\text{food}) = 417.588 - 170.318 = 247.27.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	P -value
Food	4	170.318	42.5795	4.30	2.76	< 0.01
Error	25	247.27	9.8908			
Total	29					

The null hypothesis is :

H_0 : The varieties of food are similar, against H_A : The varieties of food are different.

Since $F = 4.70 > F_{0.05}$, H_0 is rejected. The varieties of food are different. [P -value < 0.01, H_0 is rejected]

The new varieties of food will be considered effective if the mean production of milk due to introduction of any new variety of food is more than that of no dry food. Considering no dry food as control treatment we need to compare it with new varieties of food. This can be done by Dunnett's test, where the test statistic is :

$$H_0 : \alpha_0 = \alpha_i, \text{ against } H_A : \alpha_0 \neq \alpha_i, \quad i = 1, 2, 3, 4.$$

The test statistic is :

$$\begin{aligned} D &= d_{0.05, k-1, f} \sqrt{\frac{2s^2}{n}}, \quad k = 5, \quad f = 25 \\ &= 2.654 \sqrt{\frac{2 \times 9.8908}{6}} = 4.82. \end{aligned}$$

The differences between the means of control treatment (no dry food) and other treatments are :

$$\begin{aligned} |\bar{F}_0 - \bar{F}_1| &= |21.90 - 22.02| = 0.12, \quad |\bar{F}_0 - \bar{F}_2| = |21.90 - 24.52| = 2.62, \\ |\bar{F}_0 - \bar{F}_3| &= |21.90 - 28.40| = 6.50, \quad |\bar{F}_0 - \bar{F}_4| = |21.90 - 25.07| = 3.17. \end{aligned}$$

It is observed that the new variety F_3 is better than no dry food.

1.12 Estimation of Missing Observation

The yield of some of the experimental plots, specially in agricultural experiment, are lost due to some uncontrolled causes. In laboratory experiment also, some of the results of experimental plots may be damaged or lost during compilation. This loss of experimental result during experiment is termed as missing observation.

Due to missing observation the orthogonality of data of different treatments is lost and it creates problem in the application of usual analysis of variance technique as a mode of analysis of data. However, the analysis is not affected if there is only one pre-identified source of variation except the error in the data set (if completely randomized design is used). The

missing observation(s) creates problem in data analysis if the experiment is conducted through randomized block design or other complex designs.

The usual analysis of variance technique is not suitable in analysing the data having missing observation. The analysis is done after estimating the value of missing observation. Allan and Wishart (1930) have first developed the formula to estimate one missing observation in case of data of randomized block design. Yates (1933) has showed that the error sum of squares becomes minimum if the missing value is estimated by Allan and Wishart's formula. The iterative method has been developed by Yates to estimate several missing observations. The analysis is less affected if Yates method is used to estimate the missing value. However, if there are several missing observations in a small scale experiment, the error d.f. is reduced and the expected value of sum of squares due to treatment is increased affecting the F -test, specially in case of variance component analysis.

Let us consider that the observations of an experiment be y_1, y_2, \dots, y_n . Also consider that the observations of k plots of this experiment is lost and let these observation be x_1, x_2, \dots, x_k . Assume that the experimental results are linear function of parameters $\theta_1, \theta_2, \dots, \theta_m$. Then

$$y_i = a_{i1}\theta_1 + a_{i2}\theta_2 + \dots + a_{im}\theta_m + e_i \tag{8}$$

where $i = 1, 2, \dots, n$. Assume that $E(e_i) = 0$ and $V(e_i) = \sigma^2$.

In matrix notation, the equation (8) is written as

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \cdots a_{1m} \\ a_{21} & a_{22} \cdots a_{2m} \\ \dots & \dots \dots \dots \\ a_{n1} & a_{n2} \cdots a_{nm} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$Y = A\theta + U \tag{9}$$

The estimated sum of squares due to error for the model (9) is

$$(Y - A\hat{\theta})'(Y - A\hat{\theta}) = Y'Y - 2\hat{\theta}'A'Y + \hat{\theta}'A'A\hat{\theta}.$$

The normal equations to estimate the parameter vector θ using method of least squares is :

$$A'A\hat{\theta} = A'Y. \tag{10}$$

$\therefore \hat{\theta} = (A'A)^{-1}A'Y$, if A is a matrix of full rank.

The sum of squares due to estimate is $\hat{\theta}'A'Y$ and the sum of squares due to error is $Y'Y - \hat{\theta}'A'Y$.

Consider that the objective of the analysis of data is to test the null hypothesis :

$$H_0 : \theta_{l+1} = \theta_{l+2} = \dots = \theta_m = 0.$$

The model under H_0 is written as

$$Y = \begin{bmatrix} a_{11} & a_{12} \cdots a_{1l} \\ a_{21} & a_{22} \cdots a_{2l} \\ \dots & \dots \dots \dots \\ a_{n1} & a_{n2} \cdots a_{nl} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_l \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = A_1\varphi + U.$$

The sum of squares due to error from this model is

$$Y'Y - \hat{\varphi}'A'Y. \tag{11}$$

The sum of squares due to error under H_0 is

$$\hat{\theta}'A'Y - \hat{\varphi}'A'Y.$$

Now, let us replace x_1, x_2, \dots, x_k for missing observations. Then

$$E(Y) = A\theta, \quad E(X) = E \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \cdots b_{1m} \\ b_{21} & b_{22} \cdots b_{2m} \\ \dots & \dots \\ b_{k1} & b_{k2} \cdots b_{km} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{bmatrix} = B\theta.$$

The sum of squares due to error for both y and x observations is given by

$$(Y - A\bar{\theta})'(Y - A\bar{\theta}) + (X - B\bar{\theta})'(X - B\bar{\theta}).$$

The normal equation to estimate $\bar{\theta}$ is :

$$A'Y + B'X = (A'A + B'B)\bar{\theta} \quad \text{or,} \quad Y'A + X'B = \bar{\theta}'(A'A + B'B). \quad (12)$$

The sum of squares due to error is :

$$S_e^2 = Y'Y + X'X - \bar{\theta}'(A'Y + B'X). \quad (13)$$

The problem is to estimate X vector so that S_e^2 is minimum.

The method of least square gives $2X - \frac{d\bar{\theta}'}{dX}(A'Y + B'X) - B\bar{\theta} = 0$.

Also, we have $B = \frac{d\bar{\theta}'}{dX}(A'A + B'B)$.

Putting the value of B in normal equations, we get $X = B\bar{\theta}$.

If we put the value of X in (12), we get normal equation similar to (10). Again, replacing X in (13), we get error sum of squares similar to (11).

To analyse the data in presence of missing values, x_1, x_2, \dots, x_k are to be replaced by some assumed value so that the sum of squares due to error is minimum. The values of x_1, x_2, \dots, x_k are to be estimated from this sum of squares due to error. Let us consider the sum of squares due to error to estimate the X vector is

$$(Y - A_1\hat{\varphi})'(Y - A_1\hat{\varphi}) + (X - B_1\hat{\varphi})'(X - B_1\hat{\varphi}). \quad (14)$$

To get the value of $\hat{\varphi}$, the normal equations are :

$$A_1'Y + B_1'X = (A_1'A_1 + B_1'B_1)\hat{\varphi}, \quad (15)$$

where $B_1 = \frac{d\hat{\varphi}}{dX}(A_1'A_1 + B_1'B_1)$.

The sum of squares due to error $S_e'^2$ is

$$Y'Y + X'X - \hat{\varphi}'(A_1'Y + B_1'X). \quad (16)$$

Differentiating (16) w.r.t. X , we get

$$2X - \frac{d\hat{\varphi}'}{dX} [(A_1'Y + B_1'X) - B_1\hat{\varphi}] = 0.$$

$\therefore X = B_1\hat{\varphi}$.

Replacing the value of X in (15), we shall get the normal equations similar to previous one. If the value of X is placed in (16), the sum of squares similar to (11) will be obtained.

The theoretical aspect to the estimation of missing value is described above. The estimation of missing observation of a particular design will be discussed in explaining the analysis of data of that design.

Chapter 2

Multi-Way Classification

2.1 One-Way Classification

Let there be n observations y_{ij} ($i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$). These observations are classified into k classes according to the source of data. Consider that the observations of each class are the yield of each treatment, where k treatments are under investigation. The i -th treatment has n_i yields $y_{i1}, y_{i2}, y_{i3}, \dots, y_{in_i}$. Except the uncontrolled source of variation these n_i observations ($i = 1, 2, \dots, k$) are assumed to be homogeneous. The means of k treatment has \bar{y}_i ($i = 1, 2, \dots, k$) are expected to be heterogeneous since these are the means of k different treatments. Therefore, the total variation of all y_{ij} observations is mainly due to the variation of treatments. The total sum of squares of these observations can be partitioned into sum of squares due to treatment except the sum of squares due to error (uncontrolled source of variation). Hence, analysis of variance of such data set which arise from an experiment conducted to investigate the behaviour of a set of treatments is known as one-way classification. Here it is assumed that k treatments are allocated to n homogeneous plots, where i -th treatment

is replicated into n_i plots such that $n = \sum_{i=1}^k n_i$.

Model for One-Way Classification : The linear model for y_{ij} observations ($i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$) is

$$y_{ij} = \mu_i + e_{ij} = \mu + \mu_i - \mu + e_{ij} = \mu + \alpha_i + e_{ij},$$

where y_{ij} = yield of i -th treatment in j -th plot, μ is the general mean, $\alpha_i = \mu_i - \mu$ is the effect of i -th treatment and e_{ij} is the random error associated with j -th yield of i -th treatment.

Assumption : $E(e_{ij}) = 0$, $E(e_{ij}, e_{i'j'}) = \sigma^2$, if $i = i'$, $j = j'$
= 0, otherwise

Moreover, e_{ij} is normally distributed.

In matrix notation, the model is $Y = X\beta + U$, where

$$Y = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{k1} \\ y_{k2} \\ \vdots \\ y_{kn_k} \end{bmatrix}_{n \times 1}, \quad X = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times (k+1)}, \quad \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}_{(k+1) \times 1}, \quad U = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \\ \vdots \\ e_{k1} \\ e_{k2} \\ \vdots \\ e_{kn_k} \end{bmatrix}_{n \times 1}$$

Here, X is called design matrix. The sum of last k columns of X matrix is equal to the first column indicating the linear dependence of first column on other columns. The last k columns are independent. Hence, the rank of the design matrix is $(k+1-1) = k$.

The normal equations to estimate the parameters in the model are

$$X'X\hat{\beta} = X'Y.$$

It gives

$$n\hat{\mu} + \sum n_i \hat{\alpha}_i = y.. \quad (17)$$

$$n_i \hat{\mu} + n_i \hat{\alpha}_i = y_i. \quad (18)$$

Since the rank of the design matrix is k , k of the normal equations are independent. To get the unique solution of these normal equations we need to put one restriction which is $\sum n_i \hat{\alpha}_i = 0$. We have

$$\hat{\alpha}_k = -\frac{1}{n_k} \sum_{i'=1}^{k-1} n_{i'} \hat{\alpha}_{i'}.$$

From equation (18) we get

$$\hat{\mu} + \hat{\alpha}_i = \frac{y_i}{n_i}, \quad i = 1, 2, \dots, k.$$

The k -th equation is

$$\hat{\mu} + \hat{\alpha}_k = \frac{y_k}{n_k}.$$

Now, $\hat{\alpha}_i - \hat{\alpha}_k = \frac{y_i}{n_i} - \frac{y_k}{n_k}$.

$$\hat{\alpha}_i + \frac{1}{n_k} \sum_{i'=1}^{k-1} n_{i'} \hat{\alpha}_{i'} = \frac{y_i}{n_i} - \frac{y_k}{n_k}. \quad (19)$$

Also, we have $n\hat{\mu} = y..$ (20)

The equations in (19) and (20) can be written as

$$\begin{bmatrix} n & 0 & 0 & \cdots & 0 \\ 0 & 1 + \frac{n_1}{n_k} & \frac{n_2}{n_k} & \cdots & \frac{n_{k-1}}{n_k} \\ 0 & \frac{n_1}{n_k} & 1 + \frac{n_1}{n_k} & \cdots & \frac{n_{k-1}}{n_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \frac{n_1}{n_k} & \frac{n_2}{n_k} & \cdots & 1 + \frac{n_{k-1}}{n_k} \end{bmatrix} \begin{bmatrix} \mu \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{k-1} \end{bmatrix} = \begin{bmatrix} y_{..} \\ \frac{y_{1.}}{n_1} - \frac{y_{k.}}{n_k} \\ \frac{y_{2.}}{n_2} - \frac{y_{k.}}{n_k} \\ \cdots \\ \frac{y_{(k-1).}}{n_{k-1}} - \frac{y_{k.}}{n_k} \end{bmatrix}.$$

The above equations are called reduced normal equations and k of these are independent. Hence, unique solution of these equations are available. The solutions are :

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$$

These two groups of estimates are independent since the coefficient matrix of the reduced normal equations can be partitioned into two. Also, it is observed that

$$\begin{aligned} \text{Cov}(\hat{\mu}, \hat{\alpha}_i) &= \text{Cov}(\bar{y}_{..}, \bar{y}_{i.} - \bar{y}_{..}) = \text{Cov}(\bar{y}_{..}, \bar{y}_{i.}) - V(\bar{y}_{..}) \\ &= \frac{n_i \sigma^2}{nn_i} - \frac{\sigma^2}{n} = 0. \end{aligned}$$

Hence, $\hat{\mu}$ and $\hat{\alpha}_i$ are independent.

The sum of squares due to estimates is :

$$\begin{aligned} \hat{\beta}' X' Y &= (\hat{\mu} \hat{\alpha}_1 \hat{\alpha}_2 \cdots \hat{\alpha}_k) \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ \vdots \\ y_{k.} \end{bmatrix} = \hat{\mu} y_{..} + \sum \hat{\alpha}_i y_{i.} \\ &= n \bar{y}_{..}^2 + \sum n_i \bar{y}_{i.} (\bar{y}_{i.} - \bar{y}_{..}) = n \bar{y}_{..}^3 + \sum n_i \bar{y}_{i.} (\bar{y}_{i.} - \bar{y}_{..})^2 \\ &= S_1 + S_2. \end{aligned}$$

Total sum of square is $Y'Y = \sum \sum y_{ij}^2$.

Hence, we have error sum of squares as

$$\begin{aligned} S_3 = SS(\text{error}) &= \sum \sum y_{ij}^2 - n \bar{y}_{..}^2 - \sum n_i (\bar{y}_{i.} - \bar{y}_{..})^2 \\ &= \sum \sum y_{ij} - n \bar{y}_{..}^2 - \sum n_i (\bar{y}_{i.} - \bar{y}_{..})^2 \\ &= \sum \sum (y_{ij} - \bar{y}_{i.})^2. \end{aligned}$$

The objective of the analysis is to test the null hypothesis.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu \text{ (say)}$$

$$\Rightarrow H_0 : \mu_i - \mu = \alpha_i = 0, \quad i = 1, 2, \dots, k \text{ against } H_A : \alpha_i \neq 0.$$

$$\text{We have } E \sum n_i (\bar{y}_{i.} - \bar{y}_{..})^2 = E \sum n_i (\alpha_i - \bar{e}_{i.} - \bar{e}_{..})^2$$

$$\begin{aligned} &= E \sum n_i \alpha_i^2 + E \sum n_i (\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.} \bar{e}_{..}) \\ &\quad + 2E \sum n_i \alpha_i (\bar{e}_{i.} - \bar{e}_{..}). \end{aligned}$$

Under the assumption mentioned earlier,

$$\begin{aligned} E \sum n_i \alpha_i (\bar{e}_i - \bar{e}_{..}) &= 0 \\ E \sum n_i (\bar{y}_i - \bar{y}_{..})^2 &= \sum n_i \alpha_i^2 + \sum_{i=1}^k n_i \frac{\sigma^2}{n_i} + \sum_{i=1}^k n_i \frac{\sigma^2}{n} - \frac{2\sigma^2}{n} \sum n_i \\ &= \sum n_i \alpha_i^2 + (k-1)\sigma^2. \end{aligned}$$

Under $H_0 : \alpha_i = 0$,

$$\frac{E \sum n_i (\bar{y}_i - \bar{y}_{..})^2}{\sigma^2} = k - 1$$

Hence, $\sum n_i (\bar{y}_i - \bar{y}_{..})^2 \sim \chi^2 \sigma^2$ with $(k-1)$ d.f.

But if H_0 is not true the distribution of $\sum n_i (\bar{y}_i - \bar{y}_{..})^2$ is noncentral χ^2 with noncentrality parameter $\lambda = \frac{1}{2n} \sum n_i \alpha_i^2$. Also, we have

$$\begin{aligned} E \sum \sum (y_{ij} - \bar{y}_i)^2 &= \sum \sum (e_{ij} - \bar{e}_i)^2 \\ &= E \sum \sum e_{ij}^2 + E \sum \sum \bar{e}_i^2 - 2E \sum \sum e_{ij} \bar{e}_i \\ &= n\sigma^2 + \sum \sum \frac{\sigma^2}{n_i} - 2 \sum \frac{n_i \sigma^2}{n_i} \\ &= n\sigma^2 + k\sigma^2 - 2k\sigma^2 = (n-k)\sigma^2. \end{aligned}$$

$$\therefore \frac{E \sum \sum (y_{ij} - \bar{y}_i)^2}{\sigma^2} = n - k.$$

Thus, $\sum \sum (y_{ij} - \bar{y}_i)^2 \sim \chi^2 \sigma^2$ with $(n-k)$ d.f.

We have $\hat{\sigma}^2 = \frac{1}{n-k} \sum \sum (y_{ij} - \bar{y}_i)^2 = \text{M.S. (error)}$.

We have $E \text{ MS (treatment)} = \frac{E \sum n_i (y_i - \bar{y}_{..})^2}{k-1} = \sigma^2 + \frac{1}{k-1} \sum n_i \alpha_i^2$

and $E \text{ MS (error)} = \frac{E \sum n_i (y_i - \bar{y}_{..})^2}{n-k} = \sigma^2$.

Thus, $E \text{ MS (treatment)} \geq E \text{ MS (error)}$.

The equality sign will hold good if $\alpha_i = 0$. Therefore, to test the null hypothesis $H_0 : \alpha_i = 0$, the test statistic is

$$F = \frac{MS \text{ (treatment)}}{MS \text{ (error)}}.$$

Thus, F has variance ratio distribution with $(k-1)$ and $(n-k)$ d.f. If H_0 is not true, the distribution of the test statistic is noncentral F with noncentrality parameter $\lambda = \frac{1}{\sigma^2} \sum n_i \alpha_i^2$. The null hypothesis will be rejected if $F \geq F_{\alpha; k-1, n-k}$ or if P -value $\int_F^\infty f(F) dF \leq 0.05$, H_0 is rejected.

ANOVA Table

Sources	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$	P-value
Treatment	$k - 1$	S_2	$s_2 = \frac{S_2}{k - 1}$	$\frac{s_2}{s_3}$	$\sigma^2 + \frac{1}{k - 1} \sum n_i \alpha_i^2$	$\int_F^\infty f(F) dF$
Error	$n - k$	S_3	$s_3 = \frac{S_3}{n - k}$		σ^2	
Total	$n - 1$					

If $H_0 : \alpha_i = 0$ is rejected, we need multiple comparison or test of significance of any contrasts, say $\alpha_i - \alpha_{i'}$, $i \neq i'$. The null hypothesis to be tested is :

$$H_0 : \alpha_i = \alpha_{i'}, \quad i \neq i' = 1, 2, \dots, k \text{ against } H_A : \alpha_i \neq \alpha_{i'}$$

$$\Rightarrow H_0 : \alpha_i - \alpha_{i'} = 0.$$

The estimate of this contrast $\alpha_i - \alpha_{i'}$ is $\hat{\alpha}_i - \hat{\alpha}_{i'}$. The variance of the estimate is :

$$V(\hat{\alpha}_i - \hat{\alpha}_{i'}) = V(\bar{y}_i - \bar{y}_{i'}) = V(\bar{y}_i) + V(\bar{y}_{i'}) - 2\text{Cov}(\bar{y}_i, \bar{y}_{i'})$$

$$= \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right),$$

where σ^2 is estimated by MS (error) = s_3 .

Therefore, the test statistic is :
$$t = \frac{\bar{y}_i - \bar{y}_{i'}}{\sqrt{s_3 \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}}$$

This t follows Student's t distribution with $(n - k)$ d.f.

The calculated value of $t \geq t_{\frac{\alpha}{2}, n-k}$ leads us to reject the null hypothesis.

If $n_i = m$ (say), then to compare all the pairs of treatments, the CD is given by

$$CD = t_{\frac{\alpha}{2}, n-k} \sqrt{\frac{2s_3}{m}}.$$

It is also needed to test the significance of the contrast $\sum c_i \alpha_i$, where $\sum c_i = 0$. The null hypothesis is $H_0 : \sum c_i \alpha_i = 0$, against $H_A : \sum c_i \alpha_i \neq 0$.

The estimate of $\sum c_i \alpha_i$ is $\sum c_i \hat{\alpha}_i = \sum c_i \bar{y}_i$.

The variance of this estimate is $V\left(\sum c_i \bar{y}_i\right) = \sum c_i^2 V(y_i) = \sigma^2 \sum \frac{c_i^2}{n_i}$.

The test statistic is :
$$t = \frac{\sum c_i \bar{y}_i}{\sqrt{s_3 \sum \frac{c_i^2}{n_i}}}$$

This t follows Student's t distribution with $(n - k)$ d.f. The conclusion is to be drawn in a similar way as it is done in the previous case.

Example 2.1 : Four varieties of maize are cultivated in different plots of similar soil fertility. The plot size is 1 cm \times 1 cm. The number of cobs produced in different plots for different varieties are shown below :

Varieties of Maize	Number of cobs per plot (y_{ij})	Total y_i	Mean y_i
M_1	96, 102, 97, 105, 90	490	98.00
M_2	102, 112, 108, 115, 109, 100, 106	752	107.43
M_3	90, 95, 92, 95	372	93.43
M_4	98, 99, 105, 107, 108, 101	618	103.00

Analyse the data and group the maize varieties, if possible.

The variety M_3 is cultivated previously. Do you think that M_1 , M_2 and M_4 are better than M_3 ? Test the significance of $\alpha_1 + \alpha_2 - 2\alpha_4$.

Solution : We have $n_1 = 5$, $n_2 = 7$, $n_3 = 4$, $n_4 = 6$, $n = \sum n_i = 22$, $G = \sum \sum y_{ij} = 2232$,
 C.T. = $\frac{G^2}{n} = \frac{(2232)^2}{22} = 226446.545$.

$$SS \text{ (total)} = \sum \sum y_{ij}^2 - \text{C.T.} = 227466 - 226446.545 = 1019.455$$

$$SS \text{ (maize)} = \sum_{i=1}^4 \frac{y_i^2}{n_i} - \text{C.T.} = \frac{(490)^2}{5} + \frac{(752)^2}{7} + \frac{(372)^2}{4} + \frac{(618)^2}{6} - 226446.545$$

$$SS \text{ (error)} = SS \text{ (total)} - SS \text{ (maize)} = 1019.455 - 609.741 = 409.714.$$

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{\text{d.f.}}$	F	$F_{0.05}$	$F_{0.01}$	P -value
Maize	3	609.741	203.247	8.93	3.16	5.09	< 0.01
Error	18	409.714	22.762				
Total	21						

H_0 : The maize varieties are similar, H_A : The maize varieties differ significantly.

Since $F = 8.93$ is greater than both $F_{0.05}$ and $F_{0.01}$, the maize varieties differ highly significantly.

All the maize varieties are not similar. However, some varieties may be similar and they can be grouped accordingly by Duncan's multiple range test. The test is

$$D_i = d_{\alpha, i, f} \sqrt{\frac{s^2}{n_H}},$$

where n_H is the harmonic mean of n_i 's, and s^2 is the error mean square.

We have $n_H = 5$.

$$D_2 = 2.97 \sqrt{\frac{22.762}{5}}, \quad D_3 = 3.12 \sqrt{\frac{22.762}{5}}, \quad D_4 = 3.21 \sqrt{\frac{22.762}{5}}$$

$$= 6.34. \quad = 6.66. \quad = 6.85.$$

The means in ascending order are :

$$\bar{M}_3 = 93.00, \quad \bar{M}_1 = 98.00, \quad \bar{M}_4 = 103.00, \quad \bar{M}_2 = 107.43.$$

$$\begin{aligned} \overline{M}_2 - \overline{M}_3 &= 107.43 - 93.00 = 14.43 > D_4, & \therefore \text{the maize varieties are different.} \\ \overline{M}_4 - \overline{M}_3 &= 103.00 - 93.00 = 10.00 > D_3, & \therefore M_3 \text{ and } M_4 \text{ are different.} \\ \overline{M}_2 - \overline{M}_1 &= 107.43 - 98.00 = 9.43 > D_3, & \therefore M_1 \text{ and } M_2 \text{ are different.} \\ \overline{M}_1 - \overline{M}_3 &= 98.00 - 93.00 = 5.00 < D_2, & \therefore M_1 \text{ and } M_3 \text{ are similar.} \\ \overline{M}_4 - \overline{M}_1 &= 103.00 - 98.00 = 5.00 < D_2, & \therefore M_1 \text{ and } M_4 \text{ are similar.} \\ \overline{M}_2 - \overline{M}_4 &= 107.43 - 103.00 = 4.43 < D_2, & \therefore M_2 \text{ and } M_4 \text{ are similar.} \end{aligned}$$

Similar varieties are grouped giving underline below the varieties,

$$\underline{M_3, M_1, M_4, M_2}$$

Since M_3 is cultivated previously also, it is considered as control treatment. Other varieties are compared with M_3 by Dunnett's test, where the test statistic is :

$$D = d_{\alpha, k-1, f} \sqrt{\frac{2s^2}{n_H}} = 2.59 \sqrt{\frac{2 \times 22.762}{5}} = 7.82.$$

$$\overline{M}_1 - \overline{M}_3 = 98.00 - 93.00 = 5.00 < D, \quad \therefore M_1 \text{ and } M_3 \text{ are similar.}$$

$$\overline{M}_2 - \overline{M}_3 = 107.43 - 93.00 = 14.43 > D, \quad \therefore M_2 \text{ is better than } M_3.$$

$$\overline{M}_4 - \overline{M}_3 = 103.43 - 93.00 = 10.00 > D, \quad \therefore M_4 \text{ is better than } M_3.$$

We need to test the significance of $H_0 : \alpha_1 + \alpha_2 - 2\alpha_3 = 0, \sum c_i \alpha_i = 0, C_1 = 1, C_2 = 1, C_3 = -2$ against $H_A : \alpha_1 + \alpha_2 - 2\alpha_3 \neq 0$.

The estimate of the contrast is $\overline{y}_1 + \overline{y}_2 - 2\overline{y}_3 = -0.57$.

$$\text{The test statistic is : } t = \frac{|\sum c_i \hat{\alpha}_i|}{\sqrt{s^2 \sum \frac{c_i^2}{n_i}}} = \frac{0.57}{\sqrt{22.762 (\frac{1}{5} + \frac{1}{7} - \frac{4}{6})}} = 0.12.$$

Since $t_{0.05, 18} = 2.101 > t, H_0$ is accepted. The contrast is insignificant.

2.2 Two-Way Classification

Let there be pq observations y_{ij} ($i = 1, 2, \dots, p; j = 1, 2, \dots, q$) which can be classified into two broad classes according to factors A and B , where A has p levels and B has q levels. The classified observations can be shown as follows :

$B \backslash A$	B_1	B_2	\dots	B_j	\dots	B_q	Total $y_{i\cdot}$	Mean \overline{y}_i
A_1	y_{11}	y_{12}	\dots	y_{1j}	\dots	y_{1q}	$y_{1\cdot}$	\overline{y}_1
A_2	y_{21}	y_{22}	\dots	y_{2j}	\dots	y_{2q}	$y_{2\cdot}$	\overline{y}_2
\vdots	\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
A_i	y_{i1}	y_{i2}	\dots	y_{ij}	\dots	y_{iq}	$y_{i\cdot}$	\overline{y}_i
\vdots	\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
A_p	y_{p1}	y_{p2}	\dots	y_{pj}	\dots	y_{pq}	$y_{p\cdot}$	\overline{y}_p
Total $y_{\cdot j}$	$y_{\cdot 1}$	$y_{\cdot 2}$	\dots	$y_{\cdot j}$	\dots	$y_{\cdot q}$		
Mean $\overline{y}_{\cdot j}$	$\overline{y}_{\cdot 1}$	$\overline{y}_{\cdot 2}$	\dots	$\overline{y}_{\cdot j}$	\dots	$\overline{y}_{\cdot q}$	$G = y_{\cdot\cdot}$	$\overline{y}_{\cdot\cdot}$

If the levels of A are similar, the means $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_p$ are expected to be homogeneous. Similarly, if levels of B are similar, the means $\bar{y}_{.1}, \bar{y}_{.2}, \dots, \bar{y}_{.q}$ are expected to be homogeneous. The differentials in levels of A and in levels of B lead in variations of means of \bar{y}_i and $\bar{y}_{.j}$ ($i = 1, 2, \dots, p; j = 1, 2, \dots, q$). Therefore, the total variation in y_{ij} observations is expected to be due to two identified sources A and B except the variations within the observations of A_i and within the observations of B_j for all values of i and j . Hence, the analysis of variance of the aforesaid pq observations is known as two-way classification.

Model for Two-Way Classification : The linear model for y_{ij} observations is assumed to be

$$\begin{aligned} y_{ij} &= \mu_{ij} + e_{ij} = \mu + (\mu_i - \mu) + (\mu_{.j} - \mu) + (\mu_{ij} - \mu_i - \mu_{.j} + \mu) + e_{ij} \\ &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij}; \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q. \end{aligned}$$

Here μ = general mean, α_i = effect of i th level of A , β_j = effect of j th level of B , $(\alpha\beta)_{ij}$ = interaction of i th level of A in presence of j th level of B and e_{ij} = random error.

Interaction : The term $(\mu_{ij} - \mu_i - \mu_{.j} + \mu)$ is known as interaction.

Here $\mu_{ij} - \mu_i$ is the difference in the mean yield of i th level of A in presence of j th level of B and the mean yield of i th level of A . Again, $\mu_{.j} - \mu$ is the difference in the mean yield of j th level of B and overall mean. The former difference measures the amount of mean yield over mean yield of i th level of A and the latter difference measures the mean yield of j th level of B over grand mean. Therefore, the interaction $(\mu_{ij} - \mu_i - \mu_{.j} + \mu)$ measures the mean yield of i th level of A in presence of j th level of B over the influence of j th level of B .

Since we have only one observation corresponding to i th level of A_i corresponding to j th level of B_j , we cannot estimate the parameter $(\alpha\beta)_{ij}$ from these pq observations. Thus, the model

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij}$$

is not additive. That is y_{ij} observation is not obtained by adding all the impacts and e_{ij} . The model is called non-additive. Here the estimate of $(\alpha\beta)_{ij}$ and e_{ij} are same. Therefore, to analyse the data we need an additive model which is as follows :

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}; \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q.$$

In matrix notation the model is

$$Y = X\beta + U,$$

where, $Y = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1q} \\ \vdots \\ y_{p1} \\ y_{p2} \\ \vdots \\ y_{pq} \end{bmatrix}_{pq \times 1}$, $X = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}_{pq \times p+q+1}$

$$\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix}_{(p+q+1) \times 1}, U = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1q} \\ \vdots \\ e_{p1} \\ e_{p2} \\ \vdots \\ e_{pq} \end{bmatrix}_{pq \times 1}$$

The first column of X matrix is the sum of last q columns or the sum of the p columns after first column. So first column depends on other columns. Second column is the sum of last q columns minus the sum of the $(p - 1)$ columns preceding the last q columns. Therefore, two columns of X matrix are dependent. Hence the rank of X matrix is $(p + q + 1 - 2) = p + q - 1$.

Assumptions : $e_{ij} \sim NID(0, \sigma^2)$

Restriction : $\sum \alpha_i = 0, \sum \beta_j = 0$

The normal equations to estimate the parameter vector are :

$$X'X\hat{\beta} = X'Y.$$

$$\text{It gives } pq\hat{\mu} + q\sum \hat{\alpha}_i + p\sum \hat{\beta}_j = y_{..} \tag{21}$$

$$q\hat{\mu} + q\hat{\alpha}_i + \sum \hat{\beta}_j = y_{i.} \tag{22}$$

$$p\hat{\mu} + \sum \hat{\alpha}_i + p\hat{\beta}_j = y_{.j}. \tag{23}$$

Since the rank of X matrix is $(p + q - 1)$, $(p + q - 1)$ equations out of $(p + q + 1)$ equations are independent. Thus, to get the unique solution of these equations we need to put two restrictions. The restrictions are $\sum \hat{\alpha}_i = 0, \sum \hat{\beta}_j = 0$. Under the restrictions the estimates are

$$\hat{\mu} = \bar{y}_{..}, \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}, \hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$$

These three groups of estimates are independent. It can be shown by partitioning the coefficient matrix of the reduced normal equations as follows : We have

$$\sum \hat{\alpha}_i = 0 \Rightarrow \hat{\alpha}_p = -\sum_{i'=1}^{p-1} \hat{\alpha}_{i'} \quad \text{and} \quad \sum \hat{\beta}_j = 0 \Rightarrow \hat{\beta}_q = -\sum_{j'=1}^{q-1} \hat{\beta}_{j'}$$

$$\text{We have } q\hat{\mu} + q\hat{\alpha}_p = y_{p.}, \quad p\hat{\mu} + p\hat{\beta}_q = y_{.q}, \quad y_{..} = pq\hat{\mu}. \tag{24}$$

$$\text{Also } q\hat{\mu} - q\sum_{i'=1}^{p-1} \hat{\alpha}_{i'} = y_{p.} \tag{25}$$

$$p\hat{\mu} - p\sum_{j'=1}^{q-1} \hat{\beta}_{j'} = y_{.q}. \tag{26}$$

$$(22)-(25) \text{ gives } q\hat{\alpha}_i + q\sum_{i'=1}^{p-1} \hat{\alpha}_{i'} = y_{i.} - y_{p.}. \tag{27}$$

$$(23)-(26) \text{ gives } p\hat{\beta}_j + p\sum_{j'=1}^{q-1} \hat{\beta}_{j'} = y_{.j} - y_{.q}. \tag{28}$$

The coefficient matrix of equations (24), (27) and (28) is written as under :

$$\left[\begin{array}{c|cccc|cccc} pq & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & 2 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 2 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 2 & 1 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 2 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 2 \end{array} \right]$$

Since the coefficient matrix of the reduced normal equations (24), (27) and (28) can be partitioned into 3 parts, the three groups of estimates $\hat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$ are orthogonal. The sum of squares due to these estimates are also independent. The sum of squares due to estimates is

$$\begin{aligned} \hat{\beta}' X' Y &= [\hat{\mu} \hat{\alpha}_1 \hat{\alpha}_2 \cdots \hat{\alpha}_p \hat{\beta}_1 \hat{\beta}_2 \cdots \hat{\beta}_q] \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ \vdots \\ y_{p.} \\ y_{.1} \\ y_{.2} \\ \vdots \\ y_{.q} \end{bmatrix} = \hat{\mu} y_{..} + \sum \hat{\alpha}_i y_{i.} + \sum \hat{\beta}_j y_{.j} \\ &= pq \bar{y}_{..}^2 + q \sum (\bar{y}_{i.} - \bar{y}_{..})^2 + p \sum (\bar{y}_{.j} - \bar{y}_{..})^2 \\ &= S_1 + S_2 + S_3. \end{aligned}$$

This sum of squares has $(p + q - 1)$ d.f. The sum of squares due to error is :

$$\begin{aligned} S_4 = SS(\text{error}) &= Y'Y - \hat{\beta}' X' Y = \sum \sum y_{ij}^2 - pq \bar{y}_{..}^2 - q \sum (\bar{y}_{i.} - \bar{y}_{..})^2 - p \sum (\bar{y}_{.j} - \bar{y}_{..})^2 \\ &= \sum \sum (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2. \end{aligned}$$

This S_4 has $(pq - p - q + 1) = (p - 1)(q - 1)$ d.f.

Under assumption the $SS(\hat{\alpha}_i) = q \sum (\bar{y}_{i.} - \bar{y}_{..})^2$ is distributed as $\chi^2 \sigma^2$ with $(p - 1)$ d.f. This can be shown as follows :

$$\begin{aligned} Eq \sum (\bar{y}_{i.} - \bar{y}_{..})^2 &= Eq \sum (\alpha_i - \bar{\alpha} + \bar{e}_i - \bar{e}_{..})^2 \\ &= q \sum_i -E[\alpha_i^2 - \bar{e}_i^2 + \bar{e}_{..}^2 - 2\bar{e}_i \bar{e}_{..} + 2\alpha_i (\bar{e}_i - \bar{e}_{..})] \quad [\because \sum \alpha_i = 0] \\ &= q \sum \alpha_i^2 + q \sum_i \frac{\sigma^2}{q} + q \sum \frac{\sigma^2}{pq} - \frac{2pq\sigma^2}{pq} \\ &= q \sum \alpha_i^2 + p\sigma^2 + \sigma^2 - 2\sigma^2 = (p - 1)\sigma^2 + q \sum \alpha_i^2. \end{aligned}$$

$$\text{If } \alpha_i = 0, \frac{Eq \sum (\bar{y}_{i.} - \bar{y}_{..})^2}{\sigma^2} = (p - 1).$$

Therefore, $q \sum (\bar{y}_{i.} - \bar{y}_{..})^2$ is $\chi^2 \sigma^2$ with $(p - 1)$ d.f., if $\alpha_i = 0$.

Similarly, $p \sum (\bar{y}_{i.} - \bar{y}_{..})^2$ is $\chi^2 \sigma^2$ with $(q - 1)$ d.f., if $\beta_i = 0$.

Also $\sum \sum (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$ is $\chi^2 \sigma^2$ with $(p - 1)(q - 1)$ d.f.

By Cochran's theorem all these sum of squares are independently distributed as χ^2 . Therefore, to test the null hypothesis $H_0 : \alpha_i = 0$, against $H_A : \alpha_i \neq 0$, the test statistic is :

$$F_1 = \frac{S_2/p - 1}{S_4/(p - 1)(q - 1)}$$

This F is distributed as variance ratio under null hypothesis with $(p - 1)$ and $(p - 1)(q - 1)$ d.f. The non-null distribution of F is noncentral F , where the non-centrality parameter is :

$$\lambda_1 = \frac{q}{2\sigma^2} \sum \alpha_i^2$$

Therefore, if $F_1 \geq F_{\alpha; (p-1), (p-1)(q-1)}$, H_0 is rejected. The test statistic for the null hypothesis $H_0 : \beta_j = 0$, against $H_A : \beta_j \neq 0$ is :

$$F_2 = \frac{S_3/q - 1}{S_4/(p - 1)(q - 1)}$$

Under H_0 this F is distributed as central variance ratio distribution with $(q - 1)$ and $(p - 1)(q - 1)$ d.f. The non-null distribution of F_2 is non-central with noncentrality parameter

$$\lambda_2 = \frac{p}{2\sigma^2} \sum \beta_j^2$$

The null hypothesis is rejected if $F_2 \geq F_{\alpha; (q-1), (p-1)(q-1)}$ or P -value for $F_2 \leq 0.05$.

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$	P -value
A	$p - 1$	S_2	s_2	$F_1 = \frac{s_2}{s_4}$	$\sigma^2 + \frac{q}{p - 1} \sum \alpha_i^2$	$\int_{F_1}^{\infty} f(F) dF$
B	$q - 1$	S_3	s_3	$F_2 = \frac{s_3}{s_4}$	$\sigma^2 + \frac{p}{q - 1} \sum \beta_j^2$	$\int_{F_2}^{\infty} f(F) dF$
Error	$(p - 1)(q - 1)$	S_4	s_4		σ^2	
Total	$pq - 1$					

The rejection of $H_0 : \alpha_i = 0$ leads us to compare any two levels of A. For this, the null hypothesis is

$$H_0 : \alpha_i = \alpha_{i'}, \text{ against } H_A : \alpha_i \neq \alpha_{i'}, i \neq i' = 1, 2, \dots, p$$

Here H_0 implies that $\alpha_i - \alpha_{i'} = 0$. It is a contrast and its estimate is $\bar{y}_{i.} - \bar{y}_{i'.$. The variance of this estimate is :

$$\begin{aligned} V(\bar{y}_{i.} - \bar{y}_{i'..}) &= V(\bar{y}_{i.}) + V(\bar{y}_{i'..}) - 2\text{Cov}(\bar{y}_{i.}, \bar{y}_{i'..}) \\ &= \frac{\sigma^2}{q} + \frac{\sigma^2}{q} - \frac{2\sigma^2}{q}, \end{aligned}$$

where s_4 is the estimate of σ^2 .

Therefore, all pairs of A_i can be compared with

$$\text{C.D.} = t_{\frac{\alpha}{2}, (p-1)(q-1)} \sqrt{\frac{2s_A}{q}}$$

where $t_{\frac{\alpha}{2}, (p-1)(q-1)}$ is the tabulated value of t at $\alpha\%$ level of significance with $(p-1)(q-1)$ d.f. Similarly, CD to compare all pairs of β_j is :

$$\text{C.D.} = t_{\frac{\alpha}{2}, (p-1)(q-1)} \sqrt{\frac{2s_A}{p}}$$

The significance of the contrasts $\sum c_i \alpha_i$ and $\sum d_j \beta_j$, where $\sum c_i = 0$, $\sum d_j = 0$ can also be tested. The test statistics for these two contrasts are :

$$t = \frac{\sum c_i \bar{y}_i}{\sqrt{\frac{s_A}{q} \sum c_i^2}} \quad \text{and} \quad t = \frac{\sum d_j \bar{y}_j}{\sqrt{\frac{s_A}{p} \sum d_j^2}} \quad \text{respectively.}$$

Here $\sum c_i \bar{y}_i$ is the estimate of $\sum c_i \alpha_i$ and $\sum d_j \bar{y}_j$ is the estimate of $\sum d_j \beta_j$. The variances of these two estimates are :

$$V\left(\sum c_i \bar{y}_i\right) = \frac{\sigma^2}{q} \sum c_i^2 \quad \text{and} \quad V\left(\sum d_j \bar{y}_j\right) = \frac{\sigma^2}{p} \sum d_j^2.$$

Example 2.2 : To identify a quality car in respect of minimum petrol consumption 20 cars of 5 companies are kept under investigation. Each company has 4 cars each of which runs for different time period. The distances [kilometre per litre] covered by car are shown below

Time period of running (in year)	The distances covered by cars, y_{ij} (kilometre per litre)						Total $y_{.j}$	Mean $\bar{y}_{.j}$
	Company							
	A	B	C	D	E			
4	10.5	9.2	9.0	9.2	11.6	49.5	9.9	
3	10.0	8.0	8.5	8.5	11.6	46.6	9.32	
2	8.0	7.5	6.0	8.0	10.0	39.5	7.9	
1	8.2	7.0	6.5	7.0	9.0	37.7	7.54	
Total $y_{.j}$	36.7	31.7	30.0	32.7	42.2	173.3	8.665	
Mean $\bar{y}_{.j}$	9.175	7.925	7.5	8.175	10.55			

- Analyse the data and identify the best company in respect of distance covered per litre.
- Does the mean distance increase with the increase in running time compared to the running time of four years?

Solution : (i) Here $p = 4$, $q = 5$, $G = y_{..} = 173.3$, $\text{C.T.} = \frac{\sigma^2}{pq} = \frac{(173.3)^2}{4 \times 5} = 1501.6445$.

$$SS (\text{Total}) = \sum \sum y_{ij}^2 - \text{C.T.} = 1546.89 - 1501.6445 = 45.2455.$$

$$SS (\text{Time period}) = \sum \frac{y_{.j}^2}{q} - \text{C.T.} = \frac{7603.35}{5} - 1501.6445 = 19.0255.$$

$$SS (\text{Company}) = \sum \frac{y_{.j}^2}{p} - \text{C.T.} = \frac{6101.91}{4} - 1501.6445 = 23.833.$$

$$SS (\text{Error}) = SS (\text{Total}) - SS (\text{Time period}) - SS (\text{Company}).$$

$$= 45.2455 - 19.0255 - 23.833 = 2.387.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	$F_{0.01}$
Time period	3	19.0255	6.34183	31.88	3.44	5.95
Company	4	23.833	5.95825	29.96	3.26	5.41
Error	12	2.387	0.1989			
Total	19					

Since $F_1 = 31.88$ is greater than both $F_{0.05}$ and $F_{0.01}$, the differences in petrol consumption in respect of time period are highly significant. Again, $F_2 = 29.96$ is greater than $F_{0.05}$ and $F_{0.01}$. The cars of different companies are highly significantly different in respect of distance covered per litre.

To identify the best company, we can use Duncan's multiple range test, where the test statistic is

$$D_i = d_{\alpha, i, f} \sqrt{\frac{s_4}{p}}, \quad i = 2, 3, 4, 5; \quad f = 12.$$

$$D_2 = 3.08 \sqrt{\frac{0.1989}{4}}, \quad D_3 = 3.23 \sqrt{\frac{0.1989}{4}}, \quad D_4 = 3.33 \sqrt{\frac{0.1989}{4}}, \quad D_5 = 3.36 \sqrt{\frac{0.1989}{4}}$$

$$= 0.687 \qquad \qquad = 0.720 \qquad \qquad = 0.742 \qquad \qquad = 0.749$$

The means related to different companies in ascending order are :

$$\bar{C} = 7.5, \quad \bar{B} = 7.925, \quad \bar{D} = 8.175, \quad \bar{A} = 9.175, \quad \bar{E} = 10.55.$$

$$\bar{E} - \bar{C} = 10.55 - 7.50 = 3.05 > D_5, \quad \therefore \text{the means are different.}$$

$$\bar{A} - \bar{C} = 9.175 - 7.50 = 1.675 > D_4, \quad \therefore A \text{ and } C \text{ are different.}$$

$$\bar{E} - \bar{B} = 10.55 - 7.925 = 2.625 > D_4, \quad \therefore B \text{ and } E \text{ are different.}$$

$$\bar{A} - \bar{B} = 9.175 - 7.925 = 1.25 > D_3, \quad \therefore A \text{ and } B \text{ are different.}$$

$$\bar{D} - \bar{C} = 8.175 - 7.50 = 0.675 < D_3, \quad \therefore D \text{ and } C \text{ are not different.}$$

$$\bar{E} - \bar{D} = 10.55 - 8.175 = 2.375 > D_3, \quad \therefore D \text{ and } E \text{ are different.}$$

$$\bar{A} - \bar{D} = 9.175 - 8.175 = 1.00 > D_2, \quad \therefore A \text{ and } D \text{ are different.}$$

$$\bar{E} - \bar{A} = 10.55 - 9.175 = 1.375 > D_2, \quad \therefore A \text{ and } E \text{ are different.}$$

The underlined means are not significantly different.

$$\underline{\bar{C}}, \underline{\bar{B}}, \underline{\bar{D}}, \underline{\bar{A}}, \underline{\bar{E}}$$

It is observed that the company E is best in preparing car which needs minimum petrol.

(ii) We need to compare \bar{A}_2, \bar{A}_3 and \bar{A}_4 with \bar{A}_1 . This is done by Dunnett's test, where the test statistic is

$$D = d_{\alpha, k-1, f} \sqrt{\frac{2s_4}{q}}, \quad \text{where } k-1 = 3, \quad f = 12$$

$$= 2.72 \sqrt{\frac{2 \times 0.1989}{5}} = 0.767.$$

$|\bar{A}_4 - \bar{A}_3| = |9.9 - 9.32| = 0.58 < D$, A_1 and A_2 are not different.

$|\bar{A}_4 - \bar{A}_2| = |9.9 - 7.90| = 2.00 > D$, A_1 and A_3 are different.

$|\bar{A}_4 - \bar{A}_1| = |9.9 - 7.54| = 2.36 > D$, A_1 and A_4 are different.

There is an increasing trend in petrol consumption with the increase in time period except the time period of three years.

2.3 Two-Way Classification with Several (Equal) Observations Per Cell

We have considered two-way classification with one observation corresponding to i th level of A and j th level of B . Now, let us consider that there are r observations of i th level of A corresponding to j th level of B . Let y_{ijl} ($i = 1, 2, \dots, p$; $j = 1, 2, \dots, q$; $l = 1, 2, \dots, r$) be the yield of l th replication of j th level of B corresponding to i th level of A . The model for this y_{ijl} observation is :

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl},$$

where μ = general mean, α_i = effect of i th level of A , β_j = effect of j th level of B , $(\alpha\beta)_{ij}$ = interaction of i th level of A with j th level of B and e_{ijl} = random error.

$$\text{Restriction for the model : } \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0.$$

Assumption : $e_{ijl} \sim NID(0, \sigma^2)$.

The estimated error sum of squares in analysing the data can be written as

$$\phi = \sum_i \sum_j \sum_l e_{ijl}^2 = \sum_i \sum_j \sum_l (y_{ijl} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - (\hat{\alpha}\hat{\beta})_{ij})^2.$$

The normal equations to find the values of $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$ and $(\hat{\alpha}\hat{\beta})_{ij}$ are :

$$\frac{\partial \phi}{\partial \hat{\mu}} = 0 \Rightarrow y_{...} = pqr\hat{\mu} + qr \sum_i \hat{\alpha}_i + pr \sum_j \hat{\beta}_j + r \sum_i \sum_j (\hat{\alpha}\hat{\beta})_{ij} \quad (29)$$

$$\frac{\partial \phi}{\partial \hat{\alpha}_i} = 0 \Rightarrow y_{i..} = qr\hat{\mu} + qr\hat{\alpha}_i + r \sum_j \hat{\beta}_j + r \sum_j (\hat{\alpha}\hat{\beta})_{ij} \quad (30)$$

$$\frac{\partial \phi}{\partial \hat{\beta}_j} = 0 \Rightarrow y_{.j.} = pr\hat{\mu} + r \sum_i \hat{\alpha}_i + pr\hat{\beta}_j + r \sum_i (\hat{\alpha}\hat{\beta})_{ij} \quad (31)$$

$$\frac{\partial \phi}{\partial (\hat{\alpha}\hat{\beta})_{ij}} = 0 \Rightarrow y_{ij.} = r\hat{\mu} + r\hat{\alpha}_i + r\hat{\beta}_j + r(\hat{\alpha}\hat{\beta})_{ij}. \quad (32)$$

There are pq equations shown by (32). Adding both sides of these equations over suffix j , we get the equations shown in (30). If equations in (32) are added over suffix i , we get the equations shown in (31) and adding both sides of the equations shown in (32) over suffix i and j , we get the equation as shown in (29). There are $(pq + p + q + 1)$ equations. But except last pq equations all other equations are dependent on pq equations shown in (32). Hence, to get the unique solution of these equations we need to put $(p + q + 1)$ restrictions. The restrictions are :

$$\sum_i \hat{\alpha}_i = \sum_j \hat{\beta}_j = \sum_i (\hat{\alpha}\hat{\beta})_{ij} = \sum_j (\hat{\alpha}\hat{\beta})_{ij} = 0.$$

Under the restrictions, the estimates are :

$$\hat{\mu} = \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad (\hat{\alpha}\hat{\beta})_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

These estimates are orthogonal as shown below :

$$\begin{aligned} \text{Cov}(\hat{\alpha}_i, \hat{\beta}_j) &= \text{Cov}(\bar{y}_{i..} - \bar{y}_{...}, \bar{y}_{.j.} - \bar{y}_{...}) \\ &= \text{Cov}(\bar{y}_{i..}, \bar{y}_{.j.}) - \text{Cov}(\bar{y}_{...}, \bar{y}_{.j.}) - \text{Cov}(\bar{y}_{i..}, \bar{y}_{...}) + V(\bar{y}_{...}) \\ &= \frac{r\sigma^2}{qr\ pr} - \frac{pr\sigma^2}{pqr\ pr} - \frac{qr\sigma^2}{qr\ pqr} + \frac{\sigma^2}{pqr} = 0. \end{aligned}$$

Similarly, all other estimates can be shown orthogonal.

The total sum of squares of the observations y_{ijl} can be partitioned as follows :

$$\begin{aligned} \sum \sum \sum (y_{ijl} - \bar{y}_{...})^2 &= \sum \sum \sum [(y_{i..} - \bar{y}_{...}) + (y_{.j.} - \bar{y}_{...}) \\ &\quad + (\bar{y}_{ij.} - \bar{y}_{i..} - y_{.j.} + \bar{y}_{...}) + (y_{ijl} - \bar{y}_{ij.})]^2 \\ &= qr \sum (\bar{y}_{i..} - \bar{y}_{...})^2 + pr \sum (y_{.j.} - \bar{y}_{...})^2 \\ &\quad + r \sum \sum (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum \sum \sum (y_{ijl} - \bar{y}_{ij.})^2. \end{aligned}$$

The other cross-product terms are zero, thus, we have

$$\begin{aligned} SS(\text{total}) &= SS(\hat{\alpha}_i) + SS(\hat{\beta}_j) + SS(\hat{\alpha}\hat{\beta})_{ij} + SS(\text{error}) \\ &= SS(A) + SS(B) + SS(AB) + SS(\text{error}) = S_1 + S_2 + S_3 + S_4. \end{aligned}$$

Under assumption on error variance all these sum of squares are distributed as $\chi^2\sigma^2$ with different d.f. The d.f. of sum of squares are found as follows :

$$\begin{aligned} E(S_1) &= Eqr \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 = Eqr \sum_i (\alpha_i + \bar{e}_{i..} - \bar{e}_{...})^2, \\ &\quad \because \sum_i \alpha_i = \sum_j \beta_j = 0, \sum_i (\alpha\beta)_{ij} = 0 \\ &= qr \sum \alpha_i^2 + qr \sum E\bar{e}_{i..}^2 + qr \sum E\bar{e}_{...}^2 - 2qr \sum E\bar{e}_{i..}\bar{e}_{...} + 2qr \sum \alpha_i(\bar{e}_{i..} - \bar{e}_{...}) \\ &= qr \sum \alpha_i^2 + qr \sum_i \frac{\sigma^2}{qr} + qr \sum_i \frac{\sigma^2}{pqr} - 2pqr \frac{\sigma^2}{pqr} \\ &= qr \sum \alpha_i^2 + p\sigma^2 + \sigma^2 - 2\sigma^2 = (p-1)\sigma^2 + qr \sum \alpha_i^2. \end{aligned}$$

$$\text{If } \alpha_i = 0, E(S_1) = E \frac{qr \sum (\bar{y}_{i..} - \bar{y}_{...})^2}{\sigma^2} = p - 1.$$

Therefore, $qr \sum (\bar{y}_{i..} - \bar{y}_{...})^2$ is central $\chi^2\sigma^2$ with $(p-1)$ d.f., if $\alpha_i = 0$, otherwise S_1 is non-central χ^2 . Similarly, we can show that S_2 and S_3 are distributed as noncentral χ^2 with $(q-1)$ and $(p-1)(q-1)$ d.f., respectively. The sum of squares S_2 follows central χ^2 , if $\beta_j = 0$ and S_3 follows central χ^2 , if $(\alpha\beta)_{ij} = 0$. The error sum of squares (S_4) is distributed as $\chi^2\sigma^2$ with $pq(r-1)$ d.f.

The objectives of the analysis are to test the hypothesis

- (i) $H_0 : \alpha_i = 0$, against $H_A : \alpha_i \neq 0$
- (ii) $H_0 : \beta_j = 0$, against $H_A : \beta_j \neq 0$
- (iii) $H_0 : (\alpha\beta)_{ij} = 0$, against $H_A : (\alpha\beta)_{ij} \neq 0$.

$$\text{The test statistic for } H_0 \text{ (i) is } F_1 = \frac{S_1/(p-1)}{S_4/pq(r-1)}.$$

Under H_0 , F_1 follows central variance ratio distribution with $(p-1)$ and $pq(r-1)$ d.f. Under alternative hypothesis F_1 follows noncentral F -distribution with noncentrality parameter $\lambda_1 = \frac{qr}{2\sigma^2} \sum \alpha_i^2$. If $F_1 \geq F_{\alpha; (p-1), pq(r-1)}$, H_0 is rejected. The test statistics for H_0 (ii) and H_0 (iii) are :

$$F_2 = \frac{S_2/(q-1)}{S_4/pq(r-1)} \quad \text{and} \quad F_3 = \frac{S_3/(p-1)(q-1)}{S_4/pq(r-1)}, \text{ respectively.}$$

The conclusion will be drawn similarly as it is drawn for H_0 (i).

Here also F_2 and F_3 are noncentral F -distribution with noncentrality parameter

$$\lambda_2 = \frac{pr}{2\sigma^2} \sum \beta_j^2 \quad \text{and} \quad \lambda_3 = \frac{r}{2\sigma^2} \sum \sum (\alpha\beta)_{ij}^2, \text{ respectively}$$

under alternative hypothesis.

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{\text{d.f.}}$	F	$E(MS)$
A	$p-1$	S_1	$s_1 = \frac{S_1}{p-1}$	$F_1 = \frac{s_1}{s_4}$	$\sigma^2 + \frac{qr}{p-1} \sum \alpha_i^2$
B	$q-1$	S_2	$s_2 = \frac{S_2}{q-1}$	$F_2 = \frac{s_2}{s_4}$	$\sigma^2 + \frac{pr}{q-1} \sum \beta_j^2$
AB	$(p-1)(q-1)$	S_3	$s_3 = \frac{S_3}{(p-1)(q-1)}$	$F_3 = \frac{s_3}{s_4}$	$\sigma^2 + \frac{r}{(p-1)(q-1)} \sum \sum (\alpha\beta)_{ij}^2$
Error	$pq(r-1)$	S_4	$s_4 = \frac{S_4}{pq(r-1)}$	—	σ^2
Total	$pqr-1$				

If necessary, the pairwise comparison of the levels of A_i , B_j and $(AB)_{ij}$ are performed using Duncan's multiple range test, where the test statistics are :

$$D_i = d_{\alpha, i, f} \sqrt{\frac{s_4}{qr}}, \quad D_j = d_{\alpha, j, f} \sqrt{\frac{s_4}{pr}}, \quad D_k = d_{\alpha, k, f} \sqrt{\frac{s_4}{r}},$$

where $i = 2, 3, \dots, p$; $j = 2, 3, \dots, q$; $k = 2, 3, \dots, pq$, $f = pq(r-1)$.

The significance of the contrast $\sum C_i \alpha_i$, where $\sum C_i = 0$ is tested by the test statistic

$$t = \frac{\sum C_i \bar{y}_{i..}}{\sqrt{s_4 \sum \frac{C_i^2}{qr}}}$$

This t follows Student's t distribution with $pq(r-1)$ d.f. If $|t| \geq t_{\frac{\alpha}{2}, pq(r-1)}$, $H_0 : \sum C_i \alpha_i = 0$ is rejected. The test statistic for $H_0 : \sum d_j \beta_j = 0$, where $\sum d_j = 0$ is

$$t = \frac{\sum d_j \bar{y}_{.j}}{\sqrt{s_4 \sum \frac{d_j^2}{pr}}}$$

This t also is distributed as Student's t distribution with $pq(r-1)$ d.f. The inference will be drawn in similar way as it is done in the previous case.

Example 2.3 : The following data represent the birth-weight (in lb) of some new born babies of different mothers. The babies are classified according to gestation period (in days) and according to pre-natal care of mothers.

Birth-weight (y_{ijl})

Gestation period of mothers A : (in days)	B : pre-natal care of mothers			Total $y_{i..}$	Mean $\bar{y}_{i..}$
	Not at All	Irregular	Regular		
< 240	5.8,6.0,5.7	5.9,5.9, 6.0	6.0,6.1, 6.4	53.8	5.98
240-250	6.0,6.2,5.8	6.4,6.3,5.9	6.8,6.2, 6.0	55.6	6.18
250-260	6.4,5.9,6.0	6.6,5.8,6.2	6.8,6.8,7.0	57.5	6.39
260+	6.0,6.0,6.2	6.8,6.2,6.6	7.2,7.5,7.4	59.9	6.66
Total $y_{.j.}$ mean $\bar{y}_{.j.}$	72.0 6.0	74.6 6.22	80.2 6.68	226.8	6.30

- (i) Analyse the data and comment on the significance of the gestation period and pre-natal care of mothers.
- (ii) Is there any difference in the mean birth-weight of babies due to the variation in pre-natal care of mothers?
- (iii) Is there any difference in the mean birth weights of babies born after gestation period of 260 days compared to those who have been born within 250-260 days?

Solution : (i) Here $p = 4, q = 3, r = 3, G = 226.8$.

$$C.T. = \frac{G^2}{pqr} = \frac{(226.8)^2}{4 \times 3 \times 3} = 1428.84$$

$$SS(\text{Total}) = \sum \sum \sum y_{ijl}^2 - C.T. = 1436.56 - 1428.84 = 7.72$$

The $y_{ij.}$ observations are :

B	B_1	B_2	B_3
A_1	17.5	17.8	18.5
A_2	18.0	18.6	19.0
A_3	18.3	18.6	20.6
A_4	18.2	19.6	22.1

$$SS(A) = \frac{\sum y_{i..}^2}{qr} - C.T. = \frac{12880.06}{3 \times 3} - 1428.84 = 2.2778.$$

$$SS(B) = \frac{\sum y_{.j.}^2}{pr} - C.T. = \frac{17181.2}{4 \times 3} - 1428.84 = 2.9267.$$

$$SS(AB) = \frac{\sum \sum y_{ij.}^2}{r} - C.T. - SS(A) - SS(B) = \frac{4305.32}{3} - 1428.84 - 2.2778 - 2.9267 = 1.0622$$

$$\begin{aligned}
 SS(\text{Error}) &= SS(\text{Total}) - SS(A) - SS(B) - SS(AB) \\
 &= 7.72 - 2.2778 - 2.9267 - 1.0622 = 1.4533.
 \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	$F_{0.01}$
A	3	2.2778	0.7593	12.54	3.01	4.72
B	2	2.9267	1.46335	24.17	3.40	5.61
AB	6	1.0622	0.1770	2.92	2.51	3.67
Error	24	1.4533	0.06055	—	—	—
Total	35					

It is observed that the mean birth-weights of babies vary highly significantly due to the variation in gestation period, since $F_1 = 12.54$ is greater than $F_{0.05}(H_0 : \alpha_i = 0)$ and $F_{0.01}$. Highly significant variation in mean birth-weights is also observed due to the variation in mothers' pre-natal care ($H_0 : \beta_j = 0$), since $F_2 = 24.17$ is greater than both $F_{0.05}$ and $F_{0.01}$. With the increase in gestation period and simultaneously increase in the level of pre-natal care significant increase in mean birth-weight is also observed, since $F_3 = 2.92 > F_{0.05}(H_0 : (\alpha\beta)_{ij} = 0)$.

(ii) We need to compare the mean birth-weights due to pre-natal care of mothers. This can be done by Duncan's multiple range test, where the test statistic is :

$$\begin{aligned}
 D_i &= d_{\alpha, j, f} \sqrt{\frac{s_4}{pr}}, \quad j = 2, 3; \alpha = 0.05, f = 24 \\
 D_2 &= 2.95 \sqrt{\frac{0.06055}{4 \times 3}} = 0.209, \quad D_3 = 3.10 \sqrt{\frac{0.06055}{4 \times 3}} = 0.220.
 \end{aligned}$$

The means in ascending order are $\bar{B}_1 = 6.0$, $\bar{B}_2 = 6.22$, $\bar{B}_3 = 6.68$.

$$\bar{B}_3 - \bar{B}_1 = 6.68 - 6.00 = 0.68 > D_3, \quad \therefore \text{the means } \bar{B}_1 \text{ and } \bar{B}_3 \text{ are different.}$$

$$\bar{B}_2 - \bar{B}_1 = 6.22 - 6.00 = 0.22 \geq D_2, \quad \therefore \bar{B}_1 \text{ and } \bar{B}_2 \text{ are different.}$$

$$\bar{B}_3 - \bar{B}_1 = 6.68 - 6.22 = 0.46 > D_2, \quad \therefore \bar{B}_1 \text{ and } \bar{B}_3 \text{ are different.}$$

All the three means are significantly different.

(iii) We need to test $H_0 : \alpha_3 = \alpha_4$, against $H_A : \alpha_3 \neq \alpha_4$.

$$\text{The test statistic is } t = \frac{\bar{y}_{3..} - \bar{y}_{4..}}{\sqrt{\frac{2s_4}{qr}}} = \frac{6.39 - 6.66}{\sqrt{\frac{2 \times 0.06055}{3 \times 3}}} = -2.33.$$

Since $|t| > t_{0.025, 24} = 2.064$, H_0 is rejected.

2.4 Two-Way Classification with Several (Unequal) Observations Per Cell

The linear model for this analysis is

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl},$$

$$i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q; \quad l = 1, 2, \dots, n_{ij},$$

where μ = general mean, α_i = effect of i -th level of A , β_j = effect of j -th level of B , $(\alpha\beta)_{ij}$ = interaction of j -th level of B with i -th level of A , e_{ijl} = random error. The model in matrix notation ignoring the interaction term $(\alpha\beta)_{ij}$ is :

$$Y = XB + U,$$

where $Y = \begin{bmatrix} y_{111} \\ y_{112} \\ \vdots \\ y_{11n_{11}} \\ y_{121} \\ y_{122} \\ \dots \\ y_{12n_{12}} \\ \vdots \\ y_{pq1} \\ y_{pq2} \\ \vdots \\ y_{pqn_{pq}} \end{bmatrix}$, $X = \begin{bmatrix} 1100 \dots 0100 \dots 0 \\ 1010 \dots 0010 \dots 0 \\ 1001 \dots 0001 \dots 0 \\ \dots \dots \dots \dots \dots \dots \dots \\ 1000 \dots 100 \dots 1 \end{bmatrix}_{n \times (p+q+1)}$

$$B = [\mu \ \alpha_1 \ \alpha_2 \ \dots \ \alpha_p \ \beta_1, \ \beta_2 \ \dots \ \beta_q]',$$

$$U = [e_{111} \ e_{112} \ \dots \ e_{11n_{11}} \ \dots \ e_{pq1} \ e_{pq2} \ \dots \ e_{pqn_{pq}}]'$$

Here $1 = [11 \dots 1]'$, $i = 1, 2, \dots, p$; $j = 1, 2, \dots, q$

There are n_{ij} elements in I corresponding to i -th level of A and j -th level of B . The elements are 1. The model is also written :

$$y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl}. \tag{33}$$

The assumption for analysis of the data is $e_{ijl} \sim \text{NID}(0, \sigma^2)$. The model with interaction will be analysed separately. Here

$$n = \sum_i \sum_j n_{ij} = \sum_i N_i = \sum_j N_j.$$

Theorem : Rank of X is $p + q - 1$.

Proof : The first column of X matrix is equal to the sum of last q columns and second column is the sum of last q columns minus the sum of $(p - 1)$ columns preceding the last q columns. Therefore,

$$r(X) \leq p + q - 1. \tag{34}$$

Let us construct a matrix X_1 taking first element of each of I in different rows, where X_1 matrix is

$$X_1 = \begin{bmatrix} 1 & 1 & 0 \dots 0 & 1 & 0 & 0 \dots 0 \\ 1 & 1 & 0 \dots 0 & 0 & 1 & 0 \dots \\ \vdots & \vdots & \vdots & \vdots & \dots & \dots \\ 1 & 1 & 0 \dots 0 & 0 & 0 & \dots 1 \\ 1 & 0 & 1 \dots 0 & 1 & 0 & \dots 0 \\ 1 & 0 & 1 \dots 0 & 0 & 1 & \dots 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 1 \dots 0 & 0 & 0 & \dots 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & \dots 1 & 1 & 0 & \dots 0 \\ 1 & 0 & \dots 1 & 0 & 1 & \dots 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & \dots 1 \end{bmatrix}_{pq \times (p+q+1)}$$

This X_1 matrix is similar to X matrix in section (1.14). Hence, its rank is $(p + q - 1)$. Since X_1 matrix is formed with rows of X matrix,

$$r(X) \geq p + q - 1. \quad (35)$$

From (34) and (35), it can be concluded that $r(X) = p + q - 1$.

Since the rank of X is $p + q - 1$, there will be $p + q - 1$ linear independent estimable functions. The estimable functions are defined under an assumption on observations.

Assumption : The n_{ij} values are such that for all $i \neq i' = 1, 2, \dots, p$; $j \neq j' = 1, 2, \dots, q$ $\alpha_i - \alpha_{i'}$ and $\beta_j - \beta_{j'}$ are estimable.

The estimates of parameters or parametric functions are obtained by minimizing the estimated error sum of squares,

$$\phi = \sum \sum \sum (y_{ijl} - \hat{\alpha}_i - \hat{\beta}_j)^2.$$

The normal equations are :

$$y_{..} = n\hat{\mu} + \sum N_i \hat{\alpha}_i + \sum N_j \hat{\beta}_j \quad (36)$$

$$y_{i..} = N_i \hat{\mu} + N_i \hat{\alpha}_i + \sum_j n_{ij} \hat{\beta}_j \quad (37)$$

$$y_{.j.} = N_j \hat{\mu} + \sum_i n_{ij} \hat{\alpha}_i + N_j \hat{\beta}_j \quad (38)$$

Since rank of X is $(p + q - 1)$, all the normal equations are not independent. Two of them are dependent. Hence, to get the estimates of the parameters we need to put two restrictions. However, the restrictions is also not sufficient to estimate the parameters.

The estimate of α_i is to be found out eliminating the effect of β_j and the estimate of β_j is to be found out eliminating the effect of α_i . Due to unequal observations the treatment effect, say β_j is entangled with block effect, say α_i . To estimate the adjusted block effect or adjusted treatment effect we need to discuss some theorems :

Theorem : If the n_{ij} values for the model(33) are such that for all $i \neq i' = 1, 2, \dots, p$ and for all $j \neq j' = 1, 2, \dots, q$, $(\alpha_i - \alpha_{i'})$ and $(\beta_j - \beta_{j'})$ are estimable functions, then there are (i) exactly $(p + q - 1)$ estimable functions, and (ii) if $\sum C_i = 0$ and $\sum d_j = 0$, then $\sum C_i \alpha_i$ and $\sum d_j \beta_j$ are estimable functions.

Proof : (i) There are $(p + q + 1)$ normal equations as shown in equations (36), (37) and (38). The estimates of the estimable functions are to be found out from the normal equations. Since rank of X matrix is $(p + q - 1)$, there are exactly $(p + q - 1)$ independent equations and therefore, $(p + q - 1)$ estimable functions. The estimable functions are $\alpha_1 - \alpha_2, \alpha_1 - \alpha_3, \dots, \alpha_1 - \alpha_p; \beta_1 - \beta_2, \beta_1 - \beta_3, \dots, \beta_1 - \beta_q$ and $n\mu + \sum N_i \alpha_i + \sum N_j \beta_j$.

(ii) Under assumption mentioned above, $\alpha_i - \alpha_{i'}$ is estimable and its any linear function is also estimable. Thus,

$$\frac{1}{p} \sum_{i' \neq i=1}^p (\alpha_i - \alpha_{i'})$$

is also estimable. We have

$$\frac{1}{p} \sum_{i' \neq i=1}^p (\alpha_i - \alpha_{i'}) = \frac{1}{p} \sum_{i' \neq i=1}^p \alpha_i - \frac{1}{p} \sum_{i' \neq i=1}^p \alpha_{i'} = \frac{p-1}{p} \alpha_i - \frac{1}{p} \sum_{i=1}^p \alpha_i + \frac{1}{p} \alpha_i = \alpha_i - \bar{\alpha}.$$

Hence, $(\alpha_i - \bar{\alpha})$ is estimable. Therefore, $\sum C_i (\alpha_i - \bar{\alpha}) = \sum C_i \alpha_i$ is also estimable, since $\sum C_i = 0$.

In a similar way, it is possible to show that $\sum d_j \beta_j$ is also estimable.

Let us now discuss the procedure to estimate the contrasts of β_j . We have from equation (37)

$$\hat{\mu} + \hat{\alpha}_i = \frac{y_{i.}}{N_i} - \frac{\sum n_{ij} \hat{\beta}_j}{N_i}.$$

Putting the value of $\hat{\mu} + \hat{\alpha}_i$ in equation (38), we have

$$y_{.j} = N_{.j} \hat{\beta}_j + \sum_i n_{ij} \left(\frac{y_{i.}}{N_i} - \frac{\sum n_{ij} \hat{\beta}_j}{N_i} \right)$$

$$\text{or, } y_{.j} - \sum_i \frac{n_{ij} y_{i.}}{N_i} = N_{.j} \hat{\beta}_j - \sum_i n_{ij} \frac{\sum n_{ij} \hat{\beta}_j}{N_i}$$

$$\text{or, } Q_j = N_{.j} \hat{\beta}_j - \sum_i \frac{n_{ij}^2 \hat{\beta}_j}{N_i} - \sum_i \sum_{j \neq s} \frac{n_{ij} n_{is} \hat{\beta}_s}{N_i}, j \neq s$$

$$\text{or, } Q_j = C_{jj} \hat{\beta}_j - \sum_{s \neq j} C_{js} \hat{\beta}_s, \tag{39}$$

where $Q_j = y_{.j} - \sum_i \frac{n_{ij} y_{i.}}{N_i}$ = adjusted total of j -th treatment,

$$C_{jj} = N_{.j} - \sum_i \frac{n_{ij}^2}{N_i}, C_{js} = \sum_i \frac{n_{ij} n_{is}}{N_i}, j \neq s.$$

The equation (39) is written as

$$C \hat{\beta} = Q. \tag{40}$$

The problem of the analysis is to test the null hypothesis :

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_q.$$

To test the significance of this null hypothesis we need to find the solution of reduced normal equations as shown in (40). The solution of the equations is available if C matrix is of full rank. Let us investigate the rank of C matrix.

Theorem : Rank of C matrix is $(q - 1)$.

Proof : The estimable function is $\beta_j - \beta_j$, and there are $(q - 1)$ such estimable functions. These estimable functions are estimated from the reduced normal equations. Therefore, the rank of C matrix must not be less than $q - 1$. That is

$$r(c) \geq q - 1 \quad (41)$$

Again, adding the columns of C matrix, we get

$$\begin{aligned} C_{jj} + \sum_{s \neq j}^q C_{js} &= N_{.j} - \sum_i \frac{n_{ij}^2}{N_i} + \sum_i \sum_{s \neq j} \frac{n_{ij}n_{is}}{N_i} \\ &= N_{.j} - \sum_i \sum_{s=j+1}^q \frac{n_{ij}n_{is}}{N_i} = N_{.j} - N_{.j} = 0. \end{aligned}$$

This means that at least one column of the C matrix is dependent on other columns. Therefore,

$$r(C) \leq q - 1. \quad (42)$$

From (41) and (42), we can say that $r(c) = q - 1$.

Since rank of C is $(q - 1)$, where C is a $q \times q$ matrix, the unique solution of equation (40) is not available. To get the unique solution we need to put one restriction, where the restriction is $\sum \hat{\beta}_j = 0$ or $I'\hat{\beta} = 0$, when I is a vector with all elements unity. Also, we can write :

$$\begin{pmatrix} C & 1 \\ 1' & 0 \end{pmatrix} \begin{pmatrix} \beta \\ 0 \end{pmatrix} = \begin{pmatrix} Q \\ 0 \end{pmatrix}$$

$$\text{or, } C^* \beta^* = Q^*, \quad (43)$$

where C^* matrix is of order $(q + 1) \times (q + 1)$ and the rank of C^* is $(q + 1)$. Therefore,

$$\beta^* = C^{*-1} Q^*$$

$$\text{Let } C^{*-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

$$\therefore \begin{pmatrix} \hat{\beta} \\ 0 \end{pmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} Q \\ 0 \end{bmatrix}$$

$$\therefore \hat{\beta} = B_{11}Q.$$

Here the elements of C^{*-1} are such that

(i) B_{21} and B_{22} have all elements $\frac{1}{q}$, (ii) $B_{22} = 0$, (iii) $B_{11}CB_{11} = B_{11}$.

(iv) CB_{11} is an idempotent matrix with rank $q - 1$. The diagonal elements of CB_{11} are $(q - 1)/q$ and off-diagonal elements are $-\frac{1}{q}$.

Theorem : (i) $E(Q) = CB$, (ii) $V(\theta) = C\sigma^2$.

Proof : (i) Since $C\hat{\beta} = Q$ is the reduced normal equations and $\hat{\beta}$ is unbiased estimates of β ,

$$E(Q) = E(C\hat{\beta}) = C\beta.$$

$$(ii) Q_j = y_{.j} - \sum_i \frac{n_{ij}y_{i..}}{N_i}$$

We have $V(y_{i..}) = N_i\sigma^2$, $Cov(y_{i..}, y_{i'..}) = 0$, $i \neq i'$

$$Cov(y_{i..}, y_{.j}) = n_{ij}\sigma^2$$

$$Cov(y_{i..}, Q_j) = Cov\left(y_{i..}, y_{.j} - \sum_i \frac{n_{ij}y_{i..}}{N_i}\right) = n_{ij}\sigma^2 - n_{ij}\sigma^2 = 0.$$

$$\begin{aligned} \therefore V(Q_j) &= Cov(\theta_j, \theta_j) = Cov\left(\theta_j, y_{.j} - \sum_i \frac{n_{ij}y_{i..}}{N_i}\right) = Cov(\theta_j, y_{.j}) \\ &= Cov\left(y_{.j} - \frac{n_{1j}y_{1..}}{N_1} - \frac{n_{2j}y_{2..}}{N_2} - \dots - \frac{n_{pj}y_{p..}}{N_p}, y_{.j}\right) \\ &= N_{.j}\sigma^2 - \sum \frac{n_{ij}^2\sigma^2}{N_i} = \sigma^2\left(N_{.j} - \sum \frac{n_{ij}^2}{N_i}\right) = \sigma^2 C_{jj}. \end{aligned}$$

Similarly, it can be shown that

$$Cov(Q_j, Q_s) = \sigma^2 C_{js}, \quad j \neq s, \quad \therefore V(Q) = C\sigma^2.$$

Theorem : $V(\hat{\beta}) = \sigma^2 B_{11}$, where $\hat{\beta} = B_{11}\theta$.

$$\begin{aligned} \text{Proof: } V(\hat{\beta}) &= E[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]' \\ &= E[B_{11}Q - B_{11}E(Q)][B_{11}Q - B_{11}E(Q)]' \\ &= B_{11}E[Q - E(Q)][Q - E(Q)]'B_{11} \\ &= B_{11}V(Q)B_{11} = B_{11}CB_{11}\sigma^2 = B_{11}\sigma^2. \end{aligned}$$

The sum of squares due to estimates is given by

$$\begin{aligned} SS(\text{estimates}) &= \hat{\mu}y_{...} + \sum \hat{\alpha}_i y_{i..} + \sum \hat{\beta}_j y_{.j} \\ &= \sum y_{i..}(\hat{\mu} + \hat{\alpha}_i) + \sum \hat{\beta}_j y_{.j} \\ &= \sum y_{i..} \left(\frac{y_{i..}}{N_i} - \frac{\sum n_{ij}\hat{\beta}_j}{N_i} \right) + \sum \hat{\beta}_j y_{.j} \\ &= \sum \frac{y_{i..}^2}{N_i} + \sum_j \hat{\beta}_j \left(y_{.j} - \sum_i \frac{n_{ij}y_{i..}}{N_i} \right) \\ &= \sum_i \frac{y_{i..}^2}{N_i} + \sum_j \hat{\beta}_j Q_j. \end{aligned} \tag{44}$$

This sum of squares has $(p + q - 1)$ d.f. and it is distributed as χ^2 .

The objective of the analysis is to test the significance of the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_q, \tag{45}$$

which is equivalent to $H_0 : \beta_j = 0$. Under this hypothesis the model is :

$$y_{ijl} = \mu + \alpha_i + e_{ijl}; \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q; \quad l = 1, 2, \dots, n_{ij}.$$

The normal equations to estimate the parameters μ and α_i are

$$y_{...} = n\hat{\mu} + \sum N_i \hat{\alpha}_i \quad (46)$$

$$y_{i..} = N_i \hat{\mu} + N_i \hat{\alpha}_i \quad (47)$$

The equation (46) is equal to the sum of equations in (47) and hence, p out of $p + 1$ normal equations are independent. To get unique solution of these equations we need to put one restriction. The restriction is $\sum N_i \hat{\alpha}_i = 0$. Under this restriction the estimates are :

$$\hat{\mu} = \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$$

The sum of squares due to estimates of the model under H_0 (45) is

$$SS(\text{estimates}) = \hat{\mu}y_{...} + \sum \hat{\alpha}_i y_{i..} = n\bar{y}_{...}^2 + \sum N_i (\bar{y}_{i..} - \bar{y}_{...})\bar{y}_{i..} = \sum \frac{y_{i..}^2}{N_i}. \quad (48)$$

This sum of squares has p d.f. and it follows χ^2 distribution. Therefore, the sum of squares due to $\hat{\beta}_j$ under H_0 is

$$S_2 = SS(\hat{\beta}_j) = (44) - (48) = \sum \hat{\beta}_j Q_j.$$

This sum of squares has $(p + q - 1) - p = (q - 1)$ d.f.

The sum of squares due to error is given by

$$S_3 = SS(\text{error}) = \sum \sum \sum y_{ijl}^2 - \sum \frac{y_{i..}^2}{N_i} - \sum \hat{\beta}_j Q_j.$$

This S_3 is distributed as χ^2 with $(n - p - q + 1)$ d.f., where $E(S_3) = (n - p - q + 1)\sigma^2$. We have

$$S_2 = \sum \sum \sum y_{ijl}^2 - \sum \frac{y_{i..}^2}{N_i} - S_3$$

$$E(S_2) = \sum \sum \sum E(\mu + \alpha_i + \beta_j + e_{ijl})^2 - \sum_i E \frac{1}{N_i} (N_i \hat{\mu} + N_i \hat{\alpha}_i + \sum_j n_{ij} \hat{\beta}_j + e_{i..})^2 - (n - p - q + 1)\sigma^2.$$

$$= n\sigma^2 + \sum \sum n_{ij} (\mu + \alpha_i + \beta_j)^2 - \left[\sum_i \left(\hat{\mu} + \hat{\alpha}_i + \sum_i \sum_j \frac{n_{ij} \hat{\beta}_j}{N_i} \right)^2 + p\sigma^2 \right] - (n - p - q + 1)\sigma^2$$

$$= (q - 1)\sigma^2 + \sum N_{.j} \beta_j^2 - \sum_i \frac{(\sum_i n_{ij} \beta_j)^2}{N_i}$$

$$= (q - 1)\sigma^2, \text{ if } \beta_j = 0.$$

Therefore, $SS(\hat{\beta}_j)$ is distributed as central χ^2 with $(q - 1)$ d.f. if $H_0 : \beta_j = 0$. Under alternative hypothesis, this sum of squares is distributed non-central χ^2 . The test statistic to test the significance of this hypothesis is

$$F_2 = \frac{\frac{S_2}{q-1}}{\frac{S_3}{(n-p-q+1)}}.$$

This F is non-central F -variate with $(q - 1)$ and $(n - p - q + 1)$ d.f. under alternative hypothesis with non-centrality parameter,

$$\lambda_2 = \frac{1}{2\sigma^2} \left[\sum_j N_{.j} \beta_j^2 - \sum_i \frac{1}{N_{.i}} \left(\sum_i n_{ij} \beta_j \right)^2 \right].$$

Therefore, $F_2 \geq F_{\alpha; (q-1), (n-p-q+1)}$ leads to reject the null hypothesis.

The adjusted sum of squares of $\hat{\alpha}_i$ will also be found out in a similar way as it is found for $\hat{\beta}_j$. The sum of squares is :

$$S_1 = \sum_{i=1}^p \hat{\alpha}_i Q_i, \quad \text{where } Q_i = y_{i..} - \sum_j \frac{n_{ij} y_{.j}}{N_{.j}}.$$

This S_1 is found out under $H_0 : \alpha_i = 0$ and $E(S_1) = (p - 1)\sigma^2$, if H_0 is true. Therefore, the test statistic to test the significance of $H_0 : \alpha_i = 0$ is

$$F_1 = \frac{\frac{S_1}{p-1}}{\frac{S_3}{(n-p-q+1)}}.$$

The non-null distribution of F_1 is non-central F with $(p - 1)$ and $(n - p - q + 1)$ d.f. and with non-centrality parameter

$$\lambda_1 = \frac{1}{2\sigma^2} \left[\sum_i N_{.i} \alpha_i^2 - \sum_j \frac{1}{N_{.j}} \left(\sum_i n_{ij} \alpha_i \right)^2 \right].$$

If $F_1 \geq F_{\alpha; (p-1), (n-p-q+1)}$, H_0 is rejected.

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$
A	$p - 1$	S_1	$s_1 = \frac{S_1}{p - 1}$	$F_1 = \frac{s_1}{s_3}$	$\sigma^2 + \sum_i N_{.i} \alpha_i^2 - \sum_j \frac{1}{N_{.j}} \left(\sum_i n_{ij} \alpha_i \right)^2$
B	$q - 1$	S_2	$s_2 = \frac{S_2}{q - 1}$	$F_2 = \frac{s_2}{s_3}$	$\sigma^2 + \sum_j N_{.j} \beta_j^2 - \sum_i \frac{1}{N_{.i}} \left(\sum_j n_{ij} \beta_j \right)^2$
Error	$n - p - q + 1$	S_3	$s_3 = \frac{S_3}{n - p - r + 1}$		σ^2
Total	$n - 1$				

Let us assume that $H_0 : \beta_j = 0$ is rejected. Then we need to test the significance of the null hypothesis

$$H_0 : \sum d_j \beta_j = 0, \quad \text{against } H_A : \sum d_j \beta_j \neq 0.$$

The test statistic for this hypothesis is

$$t = \frac{\sum d_j \hat{\beta}_j}{\sqrt{s_3 \sum_i \sum_j d_i d_j b_{ij}}},$$

where b_{ij} is the j -th element of i -th row of B_{11} matrix.

If $|t| \geq t_{\frac{\alpha}{2}, (n-p-q+1)}$, H_0 is rejected.

Example 2.4 : To reduce the systolic blood pressure (in mm of Hg) of patients three different types of medicine are given to different patients of different age groups. After the experiment the blood pressure level is measured for each patient. The collected information are shown below :

The blood pressure level (y_{ijl})

Age group ~ A	B ~ Medicine			Total $y_{i.}$
	M_1	M_2	M_3	
A_1	130, 142, 136	130, 132, 126	130, 134, 126	1060
A_2	140	150, 145, 140	138, 140	853
A_3	146, 148	141	138, 144	717
A_4	158, 146, 160	142, 140	140	886
Total $y_{.j}$	1306	1120	1090	3516

Matrix of observations n_{ij}

B	B_1	B_2	B_3	Total N_i
A				
A_1	3	2	3	8
A_2	1	3	2	6
A_3	2	1	2	5
A_4	3	2	1	6
Total N_j	9	8	8	25

Which medicine, in your consideration, is better to reduce blood pressure?

Solution : We know, C.T. = $\frac{G^2}{n} = \frac{(3516)^2}{25} = 494490.24$

$$C_{jj} = N_{.j} - \sum_i \frac{n_{ij}^2}{N_i}, \quad C_{js} = - \sum_i \frac{n_{ij}n_{is}}{N_i} \quad \text{and} \quad Q_j = y_{.j} - \sum_i \frac{n_{ij}y_i}{N_i}.$$

Therefore, the matrix C and the vector Q are as follows :

$$C = \begin{bmatrix} 5.41 & -2.65 & -2.76 \\ -2.65 & 5.13 & -2.48 \\ -2.76 & -2.48 & 5.24 \end{bmatrix}, \quad Q = \begin{bmatrix} 36.53 \\ -10.23 \\ -26.30 \end{bmatrix}$$

Under the restriction $\hat{\beta}_3 = 0$, the normal equations are :

$$\begin{bmatrix} 5.41 & -2.65 \\ -2.65 & 5.13 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 36.53 \\ -10.23 \end{bmatrix}.$$

$$\therefore \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 5.41 & -2.65 \\ -2.65 & 5.13 \end{bmatrix}^{-1} \begin{bmatrix} 36.53 \\ -10.23 \end{bmatrix} = \begin{bmatrix} 7.73815 \\ 1.97809 \end{bmatrix}.$$

$$\text{Adjusted } SS(\hat{\beta}_j) = S_2 = \sum \hat{\beta}_j Q_j = 262.44.$$

$$SS(\text{Total}) = \sum \sum \sum y_{ijl}^2 - \text{C.T.} = 496110 - 494190.24 = 1619.76.$$

$$SS(\hat{\alpha}_i) = \sum \frac{y_{i..}^2}{N_i} - C.T. = 495368.63 - 494490.24 = 878.39$$

$$SS(\hat{\text{Error}}) = SS(\text{Total}) - SS(\hat{\alpha}_i) - SS(\hat{\beta}_j) = 1619.76 - 878.39 - 262.44 = 478.93.$$

ANOVA Table

Sources of variation	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	$F_{0.01}$	P-value
Age group (A)	3	292.7967	11.62	3.13	5.01	0.00
Medicine (B)	2	131.22	5.21	3.52	5.93	< 0.01
Error	19	25.2068				
Total	24					

Since $F_2 = 5.21 > F_{0.05}$, $H_0 : \beta_j = 0$ is rejected. This indicates that the average blood pressure level of patients using different medicines are not similar. The averages are $\bar{M}_1 = 145.11$, $\bar{M}_2 = 140.00$ and $\bar{M}_3 = 136.25$. Since the average blood pressure level of patients using M_3 is lower, M_3 can be considered better.

The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimates of parametric function $\beta_1 - \beta_3$ and $\beta_2 - \beta_3$ respectively, where estimates of $V(\hat{\beta}_1)$, $V(\hat{\beta}_2)$ and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ are respectively.

$$v(\hat{\beta}_1) = 0.24746 \times 25.2068 = 6.23767, \quad v(\hat{\beta}_2) = 0.26096 \times 25.2068 = 6.57797$$

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = 0.12723 \times 25.2068 = 3.20706.$$

To test the significance of $H_0 : \beta_1 = \beta_3$, the test statistic is

$$t = \frac{\hat{\beta}_1}{s.c(\hat{\beta}_1)} = \frac{7.73815}{\sqrt{6.23767}} = 3.098.$$

Since $|t| > t_{0.025,19} = 2.093$, H_0 is rejected. M_3 differs significantly from M .

To test the significance of $H_0 : \beta_2 = \beta_3$, the test statistic is

$$t = \frac{\hat{\beta}_2}{s.c(\hat{\beta}_2)} = \frac{1.97809}{\sqrt{6.57797}} = 0.77.$$

Since $|t| < t_{0.025,19}$, H_0 is accepted. M_2 and M_3 does not differ significantly. M_2 and M_3 are similar.

2.5 Two-Way Classification with Several (Unequal) Observations Per Cell with Interaction

The model assumed for this analysis is

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl}. \tag{49}$$

The meanings of μ, α_i, β_j and $(\alpha\beta)_{ij}$ are discussed in section (2.4).

$$i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q; \quad l = 1, 2, \dots, n_{ij}.$$

$$n = \sum \sum n_{ij}, \quad N_i = \sum_j n_{ij}, \quad N_j = \sum_i n_{ij}.$$

The assumption to analyse the data is $e_{ijl} \sim \text{NID}(0, \sigma^2)$.

The objectives of the analysis are to test the significance of the hypotheses,

$$H_0 : \alpha_i = 0, H_0 : \beta_j = 0 \text{ and } H_0 : (\alpha\beta)_{ij} = 0.$$

The normal equations to estimate the parameters are :

$$y_{..} = n\hat{\mu} + \sum N_i \hat{\alpha}_i + \sum N_{.j} \hat{\beta}_j + \sum \sum n_{ij} (\alpha\hat{\beta})_{ij}$$

$$y_{i.} = N_i \hat{\mu} + N_i \hat{\alpha}_i + \sum_j n_{ij} \hat{\beta}_j + \sum_j n_{ij} (\alpha\hat{\beta})_{ij}$$

$$y_{.j} = N_{.j} \hat{\mu} + \sum_i n_{ij} \hat{\alpha}_i + N_{.j} \hat{\beta}_j + \sum_i n_{ij} (\alpha\hat{\beta})_{ij}$$

$$y_{ij.} = n_{ij} (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + (\alpha\hat{\beta})_{ij}).$$

Here the last pq equations are independent. Other equations depend on these last pq equations. Thus, the number of independent estimable functions are pq .

The sum of squares due to estimates is

$$\begin{aligned} SS(\text{estimates}) &= \hat{\mu}y_{..} + \sum \hat{\alpha}_i y_{i.} + \sum \hat{\beta}_j y_{.j} + \sum \sum (\alpha\hat{\beta})_{ij} y_{ij.} \\ &= \sum \sum y_{ij.} (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + (\alpha\hat{\beta})_{ij}) \\ &= \sum_{i=1}^p \sum_{j=1}^q \frac{y_{ij.}^2}{n_{ij}}. \end{aligned}$$

This sum of squares has pq d.f. The sum of squares due to error is

$$S = \sum \sum \sum y_{ijl}^2 - \sum_i \sum_j \frac{y_{ij.}^2}{n_{ij}}.$$

It has $(n - pq)$ d.f., where $E(s) = (n - pq)\sigma^2$.

Under the hypothesis $(\alpha\beta)_{ij} = 0$, the model stands $y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl}$.

The sum of squares due to error for this model is

$$S_1 = \sum \sum \sum y_{ijl}^2 - \sum \frac{y_{i.}^2}{N_i} - \sum \hat{\beta}_j Q_j \quad [\text{section 2.4}]$$

This S_1 has $(n - p - q + 1)$ d.f. Therefore, the sum of squares under H_0 is

$$SS(\alpha\hat{\beta})_{ij} = S_1 - S \quad [S_1 = \sum \sum \sum y_{ijl}^2 - \sum \frac{y_{i.}^2}{N_i} - \sum \hat{\beta}_j y_{.j}, \text{ if } \hat{\beta}_j \text{ is not adjusted}]$$

This sum of squares has $(n - p - q + 1 - n + pq) = (p - 1)(q - 1)$ d.f. Therefore, the test statistic to test the significance of $H_0 : (\alpha\beta)_{ij} = 0$ is

$$F_3 = \frac{(S_1 - S)/(p - 1)(q - 1)}{S/(n - pq)}.$$

If $F_3 \geq F_{\alpha; (p-1)(q-1), n-pq}$, H_0 is rejected.

If $H_0 : (\alpha\beta)_{ij} = 0$ is true, the interaction term from model (49) can be dropped. Even if it is not true, let us assume that $(\alpha\beta)_{ij} = 0$. Under H_0 or under assumption the model stands

$$y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl}. \quad (50)$$

The analysis of this model is presented in section (2.4).

Our objective is to test the significance of $H_0 : \beta_j = 0$. Under this H_0 the model becomes

$$y_{ijl} = \mu + \alpha_i + e_{ijl}.$$

For this model the sum of squares due to error is

$$S_2 = \sum \sum \sum \sum y_{ijl}^2 - \sum_i \frac{y_{i..}^2}{N_i}.$$

This S_2 has $(n - p)$ d.f. Now, $(S_2 - S_1)$ gives the adjusted sum of squares due to $\hat{\beta}_j$, where

$$SS(\hat{\beta}_j) = S_2 - S_1.$$

It has $(q - 1)$ d.f. Therefore, the test statistic to test the significance of $H_0 : \beta_j = 0$ is

$$F_2 = \frac{(S_2 - S_1)/(q - 1)}{S/(n - pq)}.$$

$F_2 \geq F_{\alpha;(q-1),(n-pq)}$ leads us to reject the concerned null hypothesis.

Under $H_0 : \alpha_i = 0$, the model takes the shape

$$y_{ijl} = \mu + \beta_j + e_{ijl}.$$

The sum of squares due to error of this model is

$$S_3 = \sum \sum \sum \sum y_{ijl}^2 - \sum_j \frac{y_{.j.}^2}{N_{.j}}.$$

It has $(n - q)$ d.f. Now, $(S_3 - S_1)$ gives the sum of squares due to $\hat{\alpha}_i$ under $H_0 : \alpha_i = 0$. This sum of squares has $(p - 1)$ d.f. Therefore, the test statistic related to the hypothesis is

$$F_1 = \frac{(S_3 - S_1)/(p - 1)}{S/(n - pq)}$$

and $F_1 \geq F_{\alpha;(p-1),(n-pq)}$ leads us to reject the null hypothesis.

Example 2.5 : An experiment is conducted to study the productivity of four varieties of cotton seed using four different levels of urea. The experiment is conducted in 10×7 m² plots. The production of cotton (in kg/plot) are recorded for analysis. The data on production are given below :

Production of cotton (y_{ijl} kg/plot)

Level of urea	Varieties of cotton seed				Total $y_{i..}$
	B_1	B_2	B_3	B_4	
A_1	7.6, 7.2, 7.0	6.4, 6.8	7.0, 7.1	6.8	55.9
A_2	8.0, 8.2, 8.1, 8.3	7.6, 7.7, 7.0	6.2, 6.6, 6.0	7.0, 7.2	87.9
A_3	8.5, 8.2	6.6, 6.1	6.5, 6.2, 6.6	7.5, 7.2	63.4
A_4	6.0, 6.1, 5.8	6.2, 6.0	5.8, 5.9	6.01, 6.01	53.8
Total $y_{.j.}$	89.0	60.4	63.9	47.7	261.0
Mean $\bar{y}_{.j.}$	7.42	6.71	6.39	6.81	

The incidence matrix of n_{ij} elements

$B \backslash A$	B_1	B_2	B_3	B_4	Total N_{i0}
A_1	3	2	2	1	8
A_2	4	3	3	2	12
A_3	2	2	3	2	9
A_4	3	2	2	2	9
Total $N_{.j}$	12	9	10	7	38

Are the varieties of cotton seed similar?

$$\text{Solution : C.T.} = \frac{G^2}{n} = \frac{(261)^2}{38} = 1792.6579.$$

$$SS(\text{Total}) = \sum \sum \sum y_{ijl}^2 - \text{C.T.} = 1816.42 - 1792.6579 = 23.7621.$$

$$SS(\hat{\mu} + \hat{\alpha}_i) = \sum \frac{y_{i..}^2}{N_i} = 1802.691$$

$$S = \sum \sum \sum y_{ijl}^2 - \sum \sum \frac{y_{ij.}^2}{n_{ij}} = 1816.42 - 1815.232 = 1.188.$$

$$\text{We know } C_{jj} = N_j - \sum \frac{n_{ij}^2}{N_i} \text{ and } C_{js} = - \sum \frac{n_{ij}n_{is}}{N_i}, \quad Q_j = y_{.j} - \sum \frac{n_{ij}y_{i.}}{N_i}.$$

The normal equations to estimate β_j after adjusting the effect A are given by

$$C\beta = Q.$$

We have under restriction $\hat{\beta}_4 = 0$.

$$\begin{bmatrix} 8.0972 & -2.8611 & -3.0833 \\ -2.8611 & 6.8611 & -2.3611 \\ -3.0833 & -2.3611 & 7.3056 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 6.7153 \\ -1.5944 \\ -5.1389 \end{bmatrix}.$$

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0.229589 & 0.145238 & 0.143837 \\ 0.145238 & 0.255865 & 0.143990 \\ 0.143837 & 0.143990 & 0.244123 \end{bmatrix} \begin{bmatrix} 6.7153 \\ -1.5944 \\ -5.1389 \end{bmatrix} = \begin{bmatrix} 0.5710 \\ -0.1726 \\ -0.5182 \end{bmatrix}.$$

$$SS(\hat{\beta}_j)_{\text{adjusted}} = \sum \hat{\beta}_j Q_j = 6.7726.$$

$$S_1 = \sum \sum \sum y_{ijl}^2 - \sum \frac{y_{i.}^2}{N_i} - \sum \hat{\beta}_j Q_j = 1816.42 - 1802.69 - 6.7726 = 6.9514.$$

$$SS(\hat{\alpha}\beta)_{ij} = S_1 - S = 6.9564 - 1.188 = 5.7684.$$

$$\therefore F_3 = \frac{(S_1 - S)/(p-1)(q-1)}{S/n - pq} = \frac{5.7684/9}{1.188/22} = 11.87.$$

Since $F_3 > F_{0.05,9,22} = 2.36$, $H_0(\beta)_{ij} = 0$ is rejected indicating the significant impact of interaction of seed variety with level of urea.

$$F_2 = \frac{SS(\hat{\beta}_j)/q-1}{S/n - pq} = \frac{6.7726/3}{1.188/22} = 41.81.$$

$F_2 > F_{0.05,3,22} = 3.05$, $H_0 : \beta_j = 0$ is rejected. The seed varieties are significantly different.

$$S_3 = \sum \sum \sum y_{ijl}^2 - \sum \frac{y_{.j}^2}{N_j} = 1816.42 - 1798.8509 = 17.5691.$$

$$SS(\hat{\alpha}_i) = S_3 - S_1 = 17.5691 - 6.9564 = 10.6127.$$

$$\therefore F_1 = \frac{SS(\hat{\alpha}_i)/p-1}{S/n - pq} = \frac{10.6127/3}{1.188/22} = 65.51.$$

Since $F_1 > F_{0.05,3,22} = 3.05$, $H_0 : \alpha_i = 0$ is rejected. The levels of urea are significantly different.

Here the estimate $\hat{\beta}_1$ is the estimate of contrast $\beta_1 - \beta_4$. We can test the significance of the null hypothesis, $H_0 : \beta_1 - \beta_4 = 0$, against $H_A : \beta_1 - \beta_4 \neq 0$.

The test statistic to test the significance of the above hypothesis is $t = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$.

We have $v(\hat{\beta}_1) = B_{11}s^2$, where $B_{11} = 0.229584$, $s^2 = \frac{S^2}{n - pq} = 0.054$
 $= 0.229584 \times 0.054 = 0.012397$.

$\therefore s.e(\hat{\beta}_1) = \sqrt{v(\hat{\beta}_1)} = \sqrt{0.012397} = 0.11134$.

$\therefore t = \frac{0.571}{0.11134} = 5.13$.

Since $|t| > t_{0.025,22} = 2.074$, H_0 is rejected. Seed-1 is better than seed-4.

The $v(\hat{\beta}_2) = 0.01382$ and $v(\hat{\beta}_3) = 0.01318$.

The test statistic to test the significance of $H_0 : \beta_2 - \beta_4 = 0$ is

$$|t| = \frac{|\hat{\beta}_2|}{s.e(\hat{\beta}_2)} = \frac{|-0.1726|}{0.1176} = 1.47 < t_{0.025,22}$$

The test statistic to test the significance of $H_0 : \beta_3 - \beta_4 = 0$ is

$$|t| = \frac{|\hat{\beta}_3|}{s.e(\hat{\beta}_3)} = \frac{0.5182}{0.1148} = 4.51 > t_{0.025,22}$$

Therefore, seed-3 is also better than seed-4.

2.6 Three-Way Classification

Let there be three factors A , B and C having levels p , q and r respectively. Let us consider that the experimental result corresponding to i -th level of A , j -th level of B and l -th level of C ($i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, r$) be y_{ijl} . The total variation in the y_{ijl} observations can be partitioned into three main identified sources of variation namely, due to factor A , factor B and factor C and hence, the analysis of variance of these pqr observations is known as three-way classification.

The model for this analysis is

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{il} + (\beta\gamma)_{jl} + e_{ijl}, \tag{51}$$

where μ = general mean, α_i = effect of i -th level of A , β_j = effect of j -th level of B , γ_l = effect of l -th level of C , $(\alpha\beta)_{ij}$ = interaction of i -th level of A with j -th level of B , $(\alpha\gamma)_{il}$ = interaction of i -th level of A with l -th level of C , $(\beta\gamma)_{jl}$ = interaction of j -th level of B with l -th level of C and e_{ijl} = random error.

The restrictions to analysis the data are :

$$\begin{aligned} \sum \alpha_i &= \sum \beta_j = \sum \gamma_l = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = \sum_i (\alpha\gamma)_{il} = \sum_l (\alpha\gamma)_{il} \\ &= \sum_j (\beta\gamma)_{jl} = \sum_l (\beta\gamma)_{jl} = 0. \end{aligned}$$

Assumption : $e_{ijl} \sim \text{NID}(0, \sigma^2)$.

The estimated error sum of squares in analysing the model is :

$$\phi = \sum \sum \sum [y_{ijl} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_l - (\hat{\alpha\beta})_{ij} - (\hat{\alpha\gamma})_{il} - (\hat{\beta\gamma})_{jl}]^2.$$

The normal equations to estimate the parameters are :

$$\frac{\partial \phi}{\partial \hat{\mu}} = 0, \quad \frac{\partial \phi}{\partial \hat{\alpha}_i} = 0, \quad \frac{\partial \phi}{\partial \hat{\beta}_j} = 0, \quad \frac{\partial \phi}{\partial \hat{\gamma}_l} = 0, \quad \frac{\partial \phi}{\partial (\hat{\alpha}\hat{\beta})_{ij}} = 0, \quad \frac{\partial \phi}{\partial (\hat{\beta}\hat{\gamma})_{jl}}, \quad \frac{\partial \phi}{\partial (\hat{\alpha}\hat{\gamma})_{il}} = 0.$$

Thus, we have

$$\begin{aligned} y_{...} &= pqr\hat{\mu} + qr \sum \hat{\alpha}_i + pr \sum \hat{\beta}_j + pq \sum \hat{\gamma}_l + r \sum \sum (\hat{\alpha}\hat{\beta})_{ij} \\ &\quad + p \sum \sum (\hat{\beta}\hat{\gamma})_{jl} + q \sum \sum (\hat{\alpha}\hat{\gamma})_{il} \\ y_{i..} &= qr\hat{\mu} + qr\hat{\alpha}_i + r \sum \hat{\beta}_j + q \sum \hat{\gamma}_l + r \sum_j (\hat{\alpha}\hat{\beta})_{ij} + \sum \sum (\hat{\beta}\hat{\gamma})_{jl} + q \sum_l (\hat{\alpha}\hat{\gamma})_{il} \\ y_{.j.} &= pr\hat{\mu} + r \sum \hat{\alpha}_i + pr\hat{\beta}_j + p \sum \hat{\gamma}_l + r \sum_i (\hat{\alpha}\hat{\beta})_{ij} + p \sum_i (\hat{\beta}\hat{\gamma})_{jl} + \sum \sum (\hat{\alpha}\hat{\gamma})_{il} \\ &= y_{.l} = pq\hat{\mu} + q \sum \hat{\alpha}_i + p \sum \hat{\beta}_j + pq\hat{\gamma}_l + \sum \sum (\hat{\alpha}\hat{\beta})_{ij} + p \sum_j (\hat{\beta}\hat{\gamma})_{jl} + q \sum_i (\hat{\alpha}\hat{\gamma})_{il} \\ y_{ij.} &= r\hat{\mu} + \hat{\alpha}_i + r\hat{\beta}_j + \sum \hat{\gamma}_l + r(\hat{\alpha}\hat{\beta})_{ij} + \sum_l (\hat{\beta}\hat{\gamma})_{jl} + \sum_l (\hat{\alpha}\hat{\gamma})_{il} \\ y_{i.l} &= q\hat{\mu} + q\hat{\alpha}_i + \sum \hat{\beta}_j + q\hat{\gamma}_l + \sum_j (\hat{\alpha}\hat{\beta})_{ij} + q(\hat{\alpha}\hat{\gamma})_{il} + \sum_j (\hat{\beta}\hat{\gamma})_{jl} \\ y_{.jl} &= p\hat{\mu} + \sum \hat{\alpha}_i + p\hat{\beta}_j + p\hat{\gamma}_l + \sum_i (\hat{\alpha}\hat{\beta})_{ij} + \sum_i (\hat{\alpha}\hat{\gamma})_{il} + p(\hat{\beta}\hat{\gamma})_{jl}. \end{aligned}$$

There are $(pq + pr + qr + p + q + r + 1)$ normal equations. Among these $2(p + q + r)$ equations are dependent on each other. Hence, to get the unique solution of these equations, we need to put $2(p + q + r)$ restrictions. The restrictions are

$$\begin{aligned} \sum \hat{\alpha}_i &= \sum \hat{\beta}_j = \sum \hat{\gamma}_l = \sum_i (\hat{\alpha}\hat{\beta})_{ij} = \sum_j (\hat{\alpha}\hat{\beta})_{ij} = \sum_j (\hat{\alpha}\hat{\gamma})_{il} = \sum_l (\hat{\alpha}\hat{\gamma})_{il} \\ &= \sum_j (\hat{\beta}\hat{\gamma})_{jl} = \sum_l (\hat{\beta}\hat{\gamma})_{jl} = 0. \end{aligned}$$

Under the restrictions the estimates are

$$\begin{aligned} \hat{\mu} &= \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad \hat{\gamma}_l = \bar{y}_{.l} - \bar{y}_{...} \\ (\hat{\alpha}\hat{\beta})_{ij} &= \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}, \quad (\hat{\alpha}\hat{\gamma})_{il} = \bar{y}_{i.l} - \bar{y}_{i..} - \bar{y}_{.l} + \bar{y}_{...} \\ (\hat{\beta}\hat{\gamma})_{jl} &= \bar{y}_{.jl} - \bar{y}_{.j.} - \bar{y}_{.l} + \bar{y}_{...} \end{aligned}$$

These estimates are independent since

$$\begin{aligned} \text{Cov}(\hat{\alpha}_i, \hat{\beta}_j) &= \text{Cov}[\bar{y}_{i..} - \bar{y}_{...}, \bar{y}_{.j.} - \bar{y}_{...}] \\ &= \text{Cov}(\bar{y}_{i..}, \bar{y}_{.j.}) - \text{Cov}(\bar{y}_{i..}, \bar{y}_{...}) - \text{Cov}(\bar{y}_{.j.}, \bar{y}_{...}) + V(\bar{y}_{...}) \\ &= \frac{r\sigma^2}{qrpr} - \frac{qr\sigma^2}{qrpr} - \frac{pr\sigma^2}{prpr} + \frac{\sigma^2}{pqr} = 0 \end{aligned}$$

Similarly, other covariances of estimates can be shown zero.

The total sum of squares of observations can be partitioned as follows :

$$\begin{aligned}
 \sum \sum \sum (y_{ijl} - \bar{y} \dots)^2 &= \sum^p \sum^q \sum^r [(\bar{y}_{i..} - \bar{y} \dots) + (\bar{y}_{.j.} - \bar{y} \dots) + (\bar{y}_{..l} - \bar{y} \dots) \\
 &\quad + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y} \dots) + (\bar{y}_{i.l} - \bar{y}_{i..} - \bar{y}_{..l} + \bar{y} \dots) \\
 &\quad + (\bar{y}_{.jl} - \bar{y}_{.j.} - \bar{y}_{..l} + \bar{y} \dots) \\
 &\quad + (y_{ijl} - \bar{y}_{ij.} - \bar{y}_{.jl} - \bar{y}_{i.l} + \bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{..l} - \bar{y} \dots)]^2 \\
 &= qr \sum (\bar{y}_{i..} - \bar{y} \dots)^2 + pr \sum (\bar{y}_{.j.} - \bar{y} \dots)^2 + pq \sum (\bar{y}_{..l} - \bar{y} \dots)^2 \\
 &\quad + r \sum \sum (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y} \dots)^2 + q \sum \sum (\bar{y}_{i.l} - \bar{y}_{i..} - \bar{y}_{..l} + \bar{y} \dots)^2 \\
 &\quad + p \sum \sum (\bar{y}_{.jl} - \bar{y}_{.j.} - \bar{y}_{..l} + \bar{y} \dots)^2 \\
 &\quad + \sum \sum \sum (y_{ijl} - \bar{y}_{ij.} - \bar{y}_{.jl} - \bar{y}_{i.l} + \bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{..l} - \bar{y} \dots)^2 \\
 &\quad + \text{cross-product terms.}
 \end{aligned}$$

The cross-product terms are zero. Therefore, we have

$$\begin{aligned}
 SS(\text{Total}) &= SS(\hat{\alpha}_i) + SS(\hat{\beta}_j) + SS(\hat{\gamma}_l) + SS(\hat{\alpha}\hat{\beta})_{ij} + SS(\hat{\alpha}\hat{\gamma})_{il} + SS(\hat{\beta}\hat{\gamma})_{jl} + SS(\text{error}). \\
 &= SS(A) + SS(B) + SS(C) + SS(AB) + SS(AC) + SS(BC) + SS(\text{error}) \\
 &= S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7.
 \end{aligned}$$

The objectives of the present analysis are to test the significance of the hypotheses :

- (i) $H_0 : \alpha_i = 0$, against $H_A : \alpha_i \neq 0$
- (ii) $H_0 : \beta_j = 0$, against $H_A : \beta_j \neq 0$.
- (iii) $H_0 : \gamma_l = 0$, against $H_A : \gamma_l \neq 0$
- (iv) $H_0 : (\alpha\beta)_{ij} = 0$, against $H_A : (\alpha\beta)_{ij} \neq 0$
- (v) $H_0 : (\alpha\gamma)_{il} = 0$, against $H_A : (\alpha\gamma)_{il} \neq 0$
- (vi) $H_0 : (\beta\gamma)_{jl} = 0$, against $H_A : (\beta\gamma)_{jl} \neq 0$.

Under the null hypotheses and under the assumption all the sum of squares due to effects and interactions are distributed as central $\chi^2\sigma^2$. The d.f. of χ^2 is shown below :

$$\begin{aligned}
 Eqr \sum (\bar{y}_{i..} - \bar{y} \dots)^2 &= Eqr \sum [\alpha_i + \bar{e}_{i..} - \bar{e} \dots]^2, \quad \because \sum \alpha_i = 0 \\
 &= qr \sum \alpha_i^2 + qr \sum E(\bar{e}_{i..}^2) + qr \sum E(\bar{e} \dots^2) \\
 &\quad + 2qr \sum \alpha_i E(\bar{e}_{i..} - \bar{e} \dots) - 2qr \sum E(\bar{e}_{i..} \bar{e} \dots) \\
 &= qr \sum \alpha_i^2 + \frac{pqr\sigma^2}{qr} + \frac{pqr\sigma^2}{pqr\sigma^2} - \frac{2pqr\sigma^2}{pqr} \\
 &= (p-1)\sigma^2 + qr \sum \alpha_i^2.
 \end{aligned}$$

$$\therefore E \frac{qr \sum (\bar{y}_{i..} - \bar{y} \dots)^2}{\sigma^2} = p-1, \quad \text{if } \alpha_i = 0.$$

Therefore, $SS(A)$ is distributed as $\chi^2\sigma^2$ with $(p-1)$ d.f., if $\alpha_i = 0$.

Again, $E \frac{qr \sum (\bar{y}_{i..} - \bar{y}_{...})^2}{p-1} = \sigma^2$, if $\alpha_i = 0$.

$$E[MS(A)] = \sigma^2 + \frac{qr}{p-1} \sum \alpha_i^2.$$

Similarly, the d.f. of other sum of squares due to effects and interactions can be shown. Again,

$$E \sum \sum \sum (y_{ijl} - \bar{y}_{ij.} - \bar{y}_{i.l} - \bar{y}_{.jl} + \bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{.l} - \bar{y}_{...})^2 = (p-1)(q-1)(r-1)\sigma^2.$$

The sum of squares due to error is distributed as $\chi^2\sigma^2$ without any restriction.

ANOVA Table

Sources	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$
A	$p-1$	S_1	s_1	$F_1 = \frac{s_1}{s_7}$	$\sigma^2 + \frac{qr}{p-1} \sum \alpha_i^2$
B	$q-1$	S_2	s_2	$F_2 = \frac{s_2}{s_7}$	$\sigma^2 + \frac{pr}{q-1} \sum \beta_j^2$
C	$r-1$	S_3	s_3	$F_3 = \frac{s_3}{s_7}$	$\sigma^2 + \frac{pq}{r-1} \sum \gamma_l^2$
AB	$(p-1)(q-1)$	S_4	s_4	$F_4 = \frac{s_4}{s_7}$	$\sigma^2 + \frac{r}{(p-1)(q-1)} \sum \sum (\alpha\beta)_{ij}^2$
AC	$(p-1)(r-1)$	S_5	s_5	$F_5 = \frac{s_5}{s_7}$	$\sigma^2 + \frac{q}{(p-1)(r-1)} \sum \sum (\alpha\gamma)_{il}^2$
BC	$(q-1)(r-1)$	S_6	s_6	$F_6 = \frac{s_6}{s_7}$	$\sigma^2 + \frac{p}{(q-1)(r-1)} \sum \sum (\beta\gamma)_{jl}^2$
Error.	$(p-1)(q-1)(r-1)$	S_7	s_7		σ^2
Total	$pqr-1$				

To test the significance of $H_0 : \alpha_i = 0$, the test statistic is F_1 . This F_1 is distributed as central variance ratio with $(p-1)$ and $(p-1)(q-1)(r-1)$ d.f. Therefore, $F_1 \geq F_{\alpha; p-1, (p-1)(q-1)(r-1)}$ leads us to reject the null hypothesis. The non-null distribution of F_1 is non-central F with non-centrality parameter.

$$\lambda_1 = \frac{qr}{2\sigma^2} \sum \alpha_i^2.$$

The conclusion regarding other hypotheses will be made in a similar way.

If the hypotheses (i), (ii) and (iii) are rejected, it needs to compare the different levels of A_i , B_j and C_l . The Duncan's multiple range statistics for these pairwise comparisons are :

$$\text{Levels of } A_i : D_i = d_{\alpha, i, f} \sqrt{\frac{s_7}{qr}}, \quad \text{where } f = (p-1)(q-1)(r-1) \text{ and } i = 2, 3, \dots, p$$

$$\text{Levels of } B_j : D_j = d_{\alpha, j, f} \sqrt{\frac{s_7}{pr}}, \quad j = 2, 3, \dots, q$$

$$\text{Levels of } C_l : D_l = d_{\alpha, l, f} \sqrt{\frac{s_7}{pq}}, \quad l = 2, 3, \dots, r.$$

Example 2.6 : A pharmaceutical company produced 4 varieties of protein and these were given to guinea pigs of different ages and different body condition scores. The guinea pigs were classified into 3 classes according to body condition score and 3 classes according to age groups. The body weight of each guinea pig was recorded before and after the experiment. The increased body weights of each guinea pig are shown below :

Increased body weights (y_{ijl} in gm)

Age group (A)	Body condition score (B)	Protein (C)				Total y_{ij}	Mean $\bar{y}_{i..}$
		p_1	p_2	p_3	p_4		
A_1	B_1	8.0	8.5	4.6	5.0	26.1	
	B_2	7.0	7.0	6.2	5.2	25.4	
	B_3	8.5	9.0	7.6	6.0	31.1	
	Sub-total, $y_{1..}$	23.5	24.5	18.4	16.2	82.6	6.88
A_2	B_1	9.0	8.2	6.6	5.6	29.4	
	B_2	9.5	8.0	7.2	6.2	30.9	
	B_3	8.5	7.5	6.0	5.8	27.8	
	Sub-total, $y_{2..}$	27.0	23.7	19.8	17.6	88.1	7.34
A_3	B_1	6.2	8.2	5.2	4.8	24.4	
	B_2	6.0	8.0	5.6	4.0	23.6	
	B_3	7.4	8.0	4.8	4.2	24.4	
	Sub-total, $y_{3..}$	19.6	24.2	15.6	13.0	72.4	6.03

- (i) Analyse the data and group the varieties of protein.
- (ii) Is there any difference in the average increased body-weight due to A_1 and A_2 ?
- (iii) Is there any difference in B_2 and B_3 ?

Solution : (i) We have $p = 3, q = 3, r = 4, G = 243.1$.

$$C.T = \frac{G^2}{pqr} = \frac{(243.1)^2}{3 \times 3 \times 4} = 1641.6003.$$

$$SS(\text{Total}) = \sum \sum \sum y_{ijl}^2 - C.T. = 1721.29 - 1641.6003 = 79.6897.$$

$$SS(A) = \sum \frac{y_{i..}^2}{qr} - C.T. = \frac{19826.13}{3 \times 4} - 1641.6003 = 10.5772.$$

Observations of B and C [$y_{.jl}$]

B	C				Total $y_{.j}$	Mean $\bar{y}_{.j}$
	P_1	P_2	P_3	P_4		
B_1	23.2	24.9	16.4	15.4	79.9	6.66
B_2	22.5	23.0	19.0	15.4	79.9	6.66
B_3	24.4	24.5	18.4	16.0	83.3	6.94
Total $y_{.1}$	70.1	72.4	53.8	46.8	243.1	—
Mean $\bar{y}_{.1}$	7.79	8.04	5.98	5.20	—	6.75

$$SS(B) = \sum \frac{y_{.j}^2}{pr} - \text{C.T.} = \frac{19706.91}{3 \times 4} - 1641.6003 = 0.6422.$$

$$SS(C) = \sum \frac{y_{.l}^2}{2pq} - \text{C.T.} = \frac{15240.45}{3 \times 3} - 1641.6003 = 51.783.$$

$$\begin{aligned} SS(AB) &= \sum \sum \frac{y_{ij}^2}{r} - \text{C.T.} - SS(A) - SS(B) \\ &= \frac{6633.27}{4} - 1641.6003 - 10.5772 - 0.6422 = 5.5028. \end{aligned}$$

$$\begin{aligned} SS(AC) &= \sum \sum \frac{y_{i.l}^2}{q} - \text{C.T.} - SS(A) - SS(C) \\ &= \frac{5128.15}{3} - 1641.6003 - 10.5722 - 51.783 = 5.4278. \end{aligned}$$

$$\begin{aligned} SS(BC) &= \sum \sum \frac{y_{.jl}^2}{p} - \text{C.T.} - SS(B) - SS(C) \\ &= \frac{5087.95}{3} - 1641.6003 - 0.6422 - 51.783 = 1.9578. \end{aligned}$$

$$\begin{aligned} SS(\text{error}) &= SS(\text{Total}) - SS(A) - SS(B) - SS(C) - SS(AB) - SS(AC) - SS(BC) \\ &= 79.6897 - 10.5722 - 0.6422 - 51.783 - 5.5028 - 5.4278 - 1.9578 = 3.8039. \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{\text{d.f.}}$	F	$F_{0.05}$	$F_{0.01}$	P -value
A	2	10.5722	5.2861	16.68	3.88	6.93	0.00
B	2	0.6422	0.3211	1.01	"	"	> 0.05
C	3	51.783	27.261	86.00	3.49	5.95	< 0.00
AB	4	5.5028	1.3757	4.34	3.26	5.41	< 0.05
AC	6	5.4278	0.9046	2.85	3.00	4.82	> 0.05
BC	6	1.9578	0.3263	1.03	"	"	> 0.05
Error.	12	3.8039	0.31699	—	—	—	—
Total	35						

Here $F_1 = 16.68$ which is greater than $F_{0.05}$ and $F_{0.01}$. Hence, the impact of age group on increased body weight is highly significant. $F_3 = 86.00$ is greater than $F_{0.05}$ and $F_{0.01}$. Hence, the varieties of protein are highly significantly different. The effect of body condition score is insignificant ($F_2 = 1.01 < F_{0.05}$), but its interaction with age group is significant ($F_4 = 4.34 > F_{0.05}$).

Since varieties of protein are significantly different, these can be grouped according to their homogeneity by Duncan's multiple range test, where the test statistic is :

$$D_1 = d_{\alpha, l, f} \sqrt{\frac{s_7^2}{pq}}, \quad \text{where } \alpha = 0.05; l = 2, 3, 4; f = 12 \text{ and } s_7^2 = 0.31699,$$

$$D_2 = 3.08\sqrt{\frac{0.31699}{3 \times 3}} = 0.58, \quad D_3 = 3.23\sqrt{\frac{0.31699}{3 \times 3}} = 0.61, \quad D_4 = 3.33\sqrt{\frac{0.31699}{3 \times 3}} = 0.62.$$

The means of protein varieties in ascending order are :

$$\bar{P}_4 = 5.20, \quad \bar{P}_3 = 5.98, \quad \bar{P}_1 = 7.79, \quad \bar{P}_2 = 8.04$$

$$\bar{P}_2 - \bar{P}_4 = 8.04 - 5.20 = 2.84 > D_4, \quad \therefore \text{the means are different.}$$

$$\bar{P}_2 - \bar{P}_3 = 8.04 - 5.98 = 2.06 > D_3, \quad \therefore P_2 \text{ and } P_3 \text{ are different.}$$

$$\bar{P}_1 - \bar{P}_4 = 7.79 - 5.20 = 2.59 > D_3, \quad \therefore P_1 \text{ and } P_4 \text{ are different.}$$

$$\bar{P}_3 - \bar{P}_4 = 5.98 - 5.20 = 0.78 > D_2, \quad \therefore P_3 \text{ and } P_4 \text{ are different.}$$

$$\bar{P}_1 - \bar{P}_3 = 7.79 - 5.98 = 1.81 > D_2, \quad \therefore P_1 \text{ and } P_3 \text{ are different.}$$

$$\bar{P}_2 - \bar{P}_1 = 8.04 - 7.79 = 0.25 < D_2, \quad \therefore P_1 \text{ and } P_2 \text{ are similar.}$$

Therefore, P_1 and P_2 are in one group, other varieties of protein are different and they are also different from this group.

(ii) We need to test $H_0 : \alpha_1 = \alpha_2$ against $H_A : \alpha_1 \neq \alpha_2$.

$$\text{The test statistic is } t = \frac{\bar{y}_{1..} - \bar{y}_{2..}}{\sqrt{\frac{2s^2}{qr}}} = \frac{6.88 - 7.34}{\sqrt{\frac{2 \times 0.31699}{3 \times 4}}} = -2.00.$$

Since $|t| < t_{0.05,9} = 2.179$, H_0 is accepted. The levels A_1 and A_2 are not significantly different.

(iii) We need to test $H_0 : \beta_2 = \beta_3$, against $H_A : \beta_2 \neq \beta_3$.

$$\text{The test statistic is } t = \frac{\bar{y}_{.2.} - \bar{y}_{.3.}}{\sqrt{\frac{2s^2}{pr}}}.$$

However, since all levels of B are homogeneous as is observed by F -test ($F_2 = 1.01 < F_{0.05}$), no need to find the value of 't'. The levels B_2 and B_3 are similar.

2.7 Three-Way Classification with Several (Equal) Observations Per Cell

Let y_{ijkl} be the experimental result observed in an experiment conducted with i -th level of A , j -th level of B , l -th level of C in k -th replication; $i = 1, 2, \dots, p$; $j = 1, 2, \dots, q$; $l = 1, 2, \dots, r$ and $k = 1, 2, \dots, m$. The model for these $pqrm$ observations is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{il} + (\beta\gamma)_{jl} + (\alpha\beta\gamma)_{ijl} + e_{ijkl}. \quad (52)$$

Here μ = general mean, α_i = effect of i -th level of A , β_j = effect of j -th level of B , γ_l = effect of l -th level of C , $(\alpha\beta)_{ij}$ = interaction of i -th level of A with j -th level of B , $(\alpha\gamma)_{il}$ = interaction of i -th level of A with l -th level of C , $(\beta\gamma)_{jl}$ = interaction of j -th level of B with l -th level of C , $(\alpha\beta\gamma)_{ijl}$ = interaction of i -th level of A with j -th level of B and l -th level of C and e_{ijkl} = random error.

The restrictions in analysing the model (52) are

$$\begin{aligned} \sum \alpha_i &= \sum \beta_j = \sum \gamma_l = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = \sum_i (\alpha\gamma)_{il} = \sum_l (\alpha\gamma)_{il} = \sum_j (\beta\gamma)_{jl} \\ &= \sum_l (\beta\gamma)_{jl} = \sum_i (\alpha\beta\gamma)_{ijl} = \sum_j (\alpha\beta\gamma)_{ijl} = \sum_l (\alpha\beta\gamma)_{ijl} = 0. \end{aligned}$$

Assumption : $e_{ijkl} \sim \text{NID}(0, \sigma^2)$.

The estimated error sum of squares in analysing the model is

$$\phi = \sum \sum \sum \sum [y_{ijkl} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_l - (\hat{\alpha\beta})_{ij} - (\hat{\alpha\gamma})_{il} - (\hat{\beta\gamma})_{jl} - (\hat{\alpha\beta\gamma})_{ijl}]^2.$$

The normal equations to estimate different effects and interactions are given by

$$\frac{\partial \phi}{\partial \hat{\mu}} = 0, \quad \frac{\partial \phi}{\partial \hat{\alpha}_i} = 0, \quad \frac{\partial \phi}{\partial \hat{\beta}_j} = 0, \quad \frac{\partial \phi}{\partial \hat{\gamma}_l} = 0, \quad \frac{\partial \phi}{\partial (\hat{\alpha\beta})_{ij}} = 0, \quad \frac{\partial \phi}{\partial (\hat{\alpha\gamma})_{il}} = 0, \quad \frac{\partial \phi}{\partial (\hat{\beta\gamma})_{jl}} = 0$$

and $\frac{\partial \phi}{\partial (\hat{\alpha\beta\gamma})_{ijl}} = 0$.

On simplification the normal equations are found as below :

$$\begin{aligned} y_{\dots} &= pqr m \hat{\mu} + qrm \sum \hat{\alpha}_i + pr m \sum \hat{\beta}_j + pqm \sum \hat{\gamma}_l + rm \sum \sum (\hat{\alpha\beta})_{ij} \\ &\quad + qm \sum \sum (\hat{\alpha\gamma})_{il} + pm \sum \sum (\hat{\beta\gamma})_{jl} + m \sum \sum \sum (\hat{\alpha\beta\gamma})_{ijl} \\ y_{i\dots} &= qrm \hat{\mu} + qrm \hat{\alpha}_i + rm \sum \hat{\beta}_j + qm \sum \hat{\gamma}_l + rm \sum_j (\hat{\alpha\beta})_{ij} \\ &\quad + qm \sum_l (\hat{\alpha\gamma})_{il} + m \sum_j \sum_l (\hat{\beta\gamma})_{jl} + m \sum_j \sum_l (\hat{\alpha\beta\gamma})_{ijl} \\ y_{\cdot j \cdot} &= pr m \hat{\mu} + rm \sum \hat{\alpha}_i + pr m \hat{\beta}_j + pm \sum \hat{\gamma}_l + rm \sum_i (\hat{\alpha\beta})_{ij} \\ &\quad + m \sum \sum (\hat{\alpha\gamma})_{il} + pm \sum_l (\hat{\beta\gamma})_{jl} + m \sum_i \sum_l (\hat{\alpha\beta\gamma})_{ijl} \\ y_{\cdot \cdot l} &= pqm \hat{\mu} + qm \sum \hat{\alpha}_i + pm \sum \hat{\beta}_j + pqm \hat{\gamma}_l + m \sum \sum (\hat{\alpha\beta})_{ij} \\ &\quad + qm \sum_i (\hat{\alpha\gamma})_{il} + pm \sum_j (\hat{\beta\gamma})_{jl} + m \sum_i \sum_j (\hat{\alpha\beta\gamma})_{ijl} \\ y_{ij\cdot} &= rm \hat{\mu} + rm \hat{\alpha}_i + rm \hat{\beta}_j + m \sum \hat{\gamma}_l + rm (\hat{\alpha\beta})_{ij} + m \sum_l (\hat{\alpha\gamma})_{il} \\ &\quad + m \sum_l (\hat{\beta\gamma})_{jl} + m \sum_l (\hat{\alpha\beta\gamma})_{ijl} \\ y_{i\cdot l} &= qm \hat{\mu} + qm \hat{\alpha}_i + m \sum \hat{\beta}_j + qm \hat{\gamma}_l + m \sum (\hat{\alpha\beta})_{ij} + qm (\hat{\alpha\gamma})_{il} \\ &\quad + m \sum_j (\hat{\beta\gamma})_{ij} + m \sum_j (\hat{\alpha\beta\gamma})_{ijl} \\ y_{\cdot j l} &= pm \hat{\mu} + m \sum \hat{\alpha}_i + pm \hat{\beta}_j + pm \hat{\gamma}_l + m \sum_i (\hat{\alpha\beta})_{ij} + m \sum (\hat{\alpha\gamma})_{il} \\ &\quad + pm (\hat{\beta\gamma})_{jl} + m \sum_i (\hat{\alpha\beta\gamma})_{ijl} \\ y_{ijl} &= m [\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_l + (\hat{\alpha\beta})_{ij} + (\hat{\alpha\gamma})_{il} + (\hat{\beta\gamma})_{jl} + (\hat{\alpha\beta\gamma})_{ijl}]. \end{aligned}$$

There are $(pqr + pr + qr + pq + r + q + p + 1)$ normal equations and pqr of these are independent. The number of dependent equations are $(pq + pr + qr + p + q + r + 1)$. Therefore,

to get the unique solution of these normal equations we need to put $(pq + pr + qr + p + q + r + 1)$ restrictions and the restrictions are

$$\begin{aligned} \sum \hat{\alpha}_i &= \sum \hat{\beta}_j = \sum \hat{\gamma}_l = \sum_i (\hat{\alpha\beta})_{ij} = \sum_j (\hat{\beta\gamma})_{jl} = \sum_l (\hat{\beta\gamma})_{jl} \\ &= \sum_i (\hat{\alpha\gamma})_{il} = \sum_l (\hat{\alpha\gamma})_{il} = \sum_i (\hat{\alpha\beta\gamma})_{ijl} = \sum_j (\hat{\alpha\beta\gamma})_{ijl} = \sum_l (\hat{\alpha\beta\gamma})_{ijl} = 0. \end{aligned}$$

Under the restrictions the estimates are :

$$\begin{aligned} \hat{\mu} &= \bar{y}_{....}, \hat{\alpha}_i = \bar{y}_{i...} - \bar{y}_{....}, \hat{\beta}_j = \bar{y}_{.j..} - \bar{y}_{....}, \hat{\gamma}_l = \bar{y}_{..l.} - \bar{y}_{....} \\ (\hat{\alpha\beta})_{ij} &= \bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....}, (\hat{\alpha\gamma})_{il} = \bar{y}_{i.l.} - \bar{y}_{i...} - \bar{y}_{..l.} + \bar{y}_{....} \\ (\hat{\beta\gamma})_{jl} &= \bar{y}_{.j.l.} - \bar{y}_{.j..} - \bar{y}_{..l.} + \bar{y}_{....} \\ (\hat{\alpha\beta\gamma})_{ijl} &= \bar{y}_{ijl.} - \bar{y}_{ij..} - \bar{y}_{i.l.} - \bar{y}_{.j.l.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..l.} - \bar{y}_{....} \end{aligned}$$

These estimates are independent since

$$\begin{aligned} \text{Cov}(\hat{\mu}, (\hat{\alpha\beta})_{ij}) &= \text{Cov}(\bar{y}_{....}, \bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....}) \\ &= \text{Cov}(\bar{y}_{....}, \bar{y}_{ij..}) - \text{Cov}(\bar{y}_{....}, \bar{y}_{i...}) - \text{Cov}(\bar{y}_{....}, \bar{y}_{.j..}) + V(\bar{y}_{....}) \\ &= \frac{rm\sigma^2}{pqrm} - \frac{qrm\sigma^2}{pqrm} - \frac{pr m\sigma^2}{pqrm} + \frac{\sigma^2}{pqrm} = 0. \end{aligned}$$

Similarly, other covariances are also zero.

The total sum of squares of all observations is partitioned as follows :

$$\begin{aligned} \sum_i \sum_j \sum_l \sum_k (y_{ijkl} - \bar{y}_{....})^2 &= \sum_i \sum_j \sum_l \sum_k [(\bar{y}_{i...} - \bar{y}_{....}) + (\bar{y}_{.j..} - \bar{y}_{....}) \\ &\quad + (\bar{y}_{..l.} - \bar{y}_{....}) + (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....}) \\ &\quad + (\bar{y}_{i.l.} - \bar{y}_{i...} - \bar{y}_{..l.} + \bar{y}_{....}) + (\bar{y}_{.j.l.} - \bar{y}_{.j..} - \bar{y}_{..l.} + \bar{y}_{....}) \\ &\quad + (\bar{y}_{ijl.} - \bar{y}_{ij..} - \bar{y}_{i.l.} - \bar{y}_{.j.l.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..l.} - \bar{y}_{....}) \\ &\quad + \sum \sum \sum \sum (y_{ijkl} - \bar{y}_{ijl.})]^2 \\ &= qrm \sum (\bar{y}_{i...} - \bar{y}_{....})^2 + pr m \sum (\bar{y}_{.j..} - \bar{y}_{....})^2 + pqm \sum (\bar{y}_{..l.} - \bar{y}_{....})^2 \\ &\quad + rm \sum \sum (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....})^2 + qm \sum \sum (\bar{y}_{i.l.} - \bar{y}_{i...} - \bar{y}_{..l.} + \bar{y}_{....})^2 \\ &\quad + pm \sum \sum (\bar{y}_{.j.l.} - \bar{y}_{.j..} - \bar{y}_{..l.} + \bar{y}_{....})^2 \\ &\quad + m \sum \sum \sum (\bar{y}_{ijl.} - \bar{y}_{ij..} - \bar{y}_{i.l.} - \bar{y}_{.j.l.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..l.} - \bar{y}_{....})^2 \\ &\quad + \sum \sum \sum \sum (y_{ijkl} - \bar{y}_{ijl.})^2 + \text{cross-product terms.} \end{aligned}$$

The cross-product terms are zero. Therefore,

$$\begin{aligned} SS(\text{Total}) &= SS(\hat{\alpha}_i) + SS(\hat{\beta}_j) + SS(\hat{\gamma}_l) + SS(\hat{\alpha\beta})_{ij} + SS(\hat{\alpha\gamma})_{il} + SS(\hat{\beta\gamma})_{jl} \\ &\quad + SS(\hat{\alpha\beta\gamma})_{ijl} + SS(\text{error}) \\ &= SS(A) + SS(B) + SS(C) + SS(AB) + SS(AC) + SS(BC) + SS(ABC) + SS(\text{error}) \\ &= S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7 + S_8. \end{aligned}$$

The objectives of the analysis are to test the significance of the hypotheses :

- (i) $H_0 : \alpha_i = 0$, against $H_A : \alpha_i \neq 0$,
- (ii) $H_0 : \beta_j = 0$, against $H_A : \beta_j \neq 0$,
- (iii) $H_0 : \gamma_l = 0$, against $H_A : \gamma_l \neq 0$,
- (iv) $H_0 : (\alpha\beta)_{ij} = 0$, against $H_A : (\alpha\beta)_{ij} \neq 0$,
- (v) $H_0 : (\alpha\gamma)_{il} = 0$, against $H_A : (\alpha\gamma)_{il} \neq 0$,
- (vi) $H_0 : (\beta\gamma)_{jl} = 0$, against $H_A : (\beta\gamma)_{jl} \neq 0$,
- (vii) $H_0 : (\alpha\beta\gamma)_{ijl} = 0$, against $H_A : (\alpha\beta\gamma)_{ijl} \neq 0$.

All the sum of squares are independently distributed as central chi-square under null hypotheses. The d.f. of each chi-square variate can be shown as follows :

$$\begin{aligned}
 E[SS(A)] &= Eqrm \sum (\bar{y}_{i\dots} - \bar{y}\dots)^2 = Eqrm \sum (\alpha_i + \bar{y}_{i\dots} - \bar{y}\dots)^2 \\
 &= qrm \sum \alpha_i^2 + qrm \sum E(\bar{y}_{i\dots}^2) + qrm \sum E(\bar{y}\dots^2) \\
 &\quad - 2qrm E \sum (\bar{y}_{i\dots}, \bar{y}\dots) + 2qrm \sum \alpha_i E(\bar{y}_{i\dots} - \bar{y}\dots) \\
 &= qrm \sum \alpha_i^2 + \frac{pqrm\sigma^2}{qrm} + \frac{pqrm\sigma^2}{pqrm} - \frac{2pqrm\sigma^2}{pqrm} \\
 &= (p-1)\sigma^2 + qrm \sum \alpha_i^2 = (p-1)\sigma^2, \text{ if } \alpha_i = 0.
 \end{aligned}$$

$$\therefore E \left[\frac{qrm \sum (\bar{y}_{i\dots} - \bar{y}\dots)^2}{\sigma^2} \right] = p-1, \text{ if } \alpha_i = 0.$$

Therefore, $SS(A)/\sigma^2$ is distributed as χ^2 with $(p-1)$ d.f., if $\alpha_i = 0$. The d.f. of other sum of squares can be found out in a similar way. The d.f. of SS (error) is found out as follows :

$$\begin{aligned}
 E &= \sum \sum \sum \sum (y_{ijkl} - \bar{y}_{ijl})^2 = E \sum \sum \sum \sum (e_{ijkl} - \bar{e}_{ijl})^2 \\
 &= \sum \sum \sum \sum E(e_{ijkl}^2) + \sum \sum \sum \sum E(\bar{e}_{ijl}^2) - 2 \sum \sum \sum \sum E(e_{ijkl} \bar{e}_{ijl}) \\
 &= pqrm\sigma^2 + pqrm \frac{\sigma^2}{m} - \frac{2pqrm\sigma^2}{m} = \sigma^2 pqr(m-1).
 \end{aligned}$$

$$\therefore \frac{SS(\text{error})}{\sigma^2} \text{ is distributed as chi-square with } pqr(m-1) \text{ d.f. without any restriction.}$$

$$\text{Therefore, } E \frac{SS(\text{error})}{\sigma^2} = pqr(m-1) \text{ or, } EMS(\text{error}) = \sigma^2.$$

The objectives of the analysis are to test the null hypotheses :

- (i) $H_0 : \alpha_i = 0$, against $H_A : \alpha_i \neq 0$,
- (ii) $H_0 : \beta_j = 0$, against $H_A : \beta_j \neq 0$,
- (iii) $H_0 : \gamma_l = 0$, against $H_A : \gamma_l \neq 0$,
- (iv) $H_0 : (\alpha\beta)_{ij} = 0$, against $H_A : (\alpha\beta)_{ij} \neq 0$,
- (v) $H_0 : (\alpha\gamma)_{il} = 0$, against $H_A : (\alpha\gamma)_{il} \neq 0$,
- (vi) $H_0 : (\beta\gamma)_{jl} = 0$, against $H_A : (\beta\gamma)_{jl} \neq 0$,
- (vii) $H_0 : (\alpha\beta\gamma)_{ijl} = 0$, against $H_A : (\alpha\beta\gamma)_{ijl} \neq 0$.

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$
A	$p - 1$	S_1	$s_1 = \frac{S_1}{p - 1}$	$F_1 = \frac{s_1}{s_8}$	$\sigma^2 + \frac{qrm}{p - 1} \sum \alpha_i^2$
B	$q - 1$	S_2	$s_2 = \frac{S_2}{q - 1}$	$F_2 = \frac{s_2}{s_8}$	$\sigma^2 + \frac{prm}{q - 1} \sum \beta_j^2$
C	$r - 1$	S_3	$s_3 = \frac{S_3}{r - 1}$	$F_3 = \frac{s_3}{s_8}$	$\sigma^2 + \frac{pqm}{r - 1} \sum \gamma_l^2$
AB	$(p - 1)(q - 1)$	S_4	$s_4 = \frac{S_4}{(p - 1)(q - 1)}$	$F_4 = \frac{s_4}{s_8}$	$\sigma^2 + \frac{rm}{(p - 1)(q - 1)} \sum \sum (\alpha\beta)_{ij}^2$
AC	$(p - 1)(r - 1)$	S_5	$s_5 = \frac{S_5}{(p - 1)(r - 1)}$	$F_5 = \frac{s_5}{s_8}$	$\sigma^2 + \frac{qm}{(p - 1)(r - 1)} \sum \sum (\alpha\gamma)_{il}^2$
BC	$(q - 1)(r - 1)$	S_6	$s_6 = \frac{S_6}{(q - 1)(r - 1)}$	$F_6 = \frac{s_6}{s_8}$	$\sigma^2 + \frac{pm}{(q - 1)(r - 1)} \sum \sum (\beta\gamma)_{jl}^2$
ABC	$(p - 1)(r - 1)(q - 1)$	S_7	$s_7 = \frac{S_7}{(p - 1)(r - 1)(q - 1)}$	$F_7 = \frac{s_7}{s_8}$	$\sigma^2 + \frac{m \sum \sum \sum (\alpha\beta\gamma)_{ijl}^2}{(p - 1)(r - 1)(q - 1)}$
Error	$pqr(m - 1)$	S_8	$s_8 = \frac{S_8}{pqr(m - 1)}$	—	σ^2
Total	$pqrm - 1$				

The null hypothesis (i) is rejected if $F_1 \geq F_{0.05; (p-1), (p-1)(q-1)(r-1)}$, where F_1 is distributed as central variance ratio distribution with $(p - 1)$ and $(p - 1)(q - 1)(r - 1)$ d.f. The non-null distribution of F_1 is non-central F with non-centrality parameter,

$$\lambda_1 = \frac{qrm}{2\sigma^2} \sum \alpha_i^2.$$

The rejection of hypothesis (i) leads to test the significance of null hypothesis,

$$H_0 : \alpha_i = \alpha_{i'}, \text{ against } H_A : \alpha_i \neq \alpha_{i'}, i \neq i' = 1, 2, \dots, p.$$

The comparison of all pairs of α_i and $\alpha_{i'}$ is done by Duncan's multiple range test, where the test statistic is

$$D_i = d_{\alpha, i, f} \sqrt{\frac{s_8}{qrm}}, \quad i = 2, 3, \dots, p.$$

For comparison of any particular pair α_i and $\alpha_{i'}$, the test statistic is

$$t = \frac{\bar{y}_{i\dots} - \bar{y}_{i'\dots}}{\sqrt{\frac{2s_8}{qrm}}},$$

where t is distributed as Student's t with $(p - 1)(q - 1)(r - 1)$ d.f. Hence, $|t| \geq t_{\frac{\alpha}{2}, (p-1)(q-1)(r-1)}$ leads to reject the corresponding null hypothesis.

The test statistics for hypotheses (ii) to (vii) are respectively F_2, F_3, \dots, F_7 . The conclusion will be made similarly as it is done for hypothesis (i). The Duncan's multiple range test statistics for $H_0 : \beta_j = \beta_{j'} (j \neq j' = 1, 2, \dots, q)$ and for $H_0 : \gamma_l = \gamma_{l'} (l \neq l' = 1, 2, \dots, r)$ are respectively.

$$D_j = d_{\alpha, j, f} \sqrt{\frac{s_8}{prm}}, \quad j = 2, 3, \dots, q; \quad f = (p - 1)(q - 1)(r - 1)$$

and $D_l = d_{\alpha, l, f} \sqrt{\frac{s_8}{pqm}}, \quad l = 2, 3, \dots, r.$

Example 2.7 : The number of ever born children to mothers of child bearing age below 40 years are recorded from 72 mothers, where mothers are classified according to their education, working condition and socioeconomic status. The recorded data are shown below :

Number of ever born children (y_{ijk})

Level of education A	Working condition B	Socioeconomic condition (C)			Total $y_{ij..}$	Total $y_{i...}$
		C_1 Low	C_2 Medium	C_3 High		
Illiterate A_1	B_1 housewife	4, 3, 5	4, 3, 3	3, 3, 3	31	
	work outside = B_2	3, 3, 2	3, 3, 4	2, 3, 2	25	
	Total $y_{1..}$	20	20	16		56
Primary A_2	B_1	4, 4, 4	4, 3, 4	4, 3, 2	32	
	B_2	3, 3, 2	3, 2, 2	3, 3, 3	24	
	Total $y_{2..}$	20	18	18		56
Secondary A_3	B_1	4, 3, 3	4, 2, 3	4, 1, 2	26	
	B_2	2, 1, 1	2, 2, 1	1, 2, 1	13	
	Total $y_{3..}$	14	14	11		39
Higher A_4	B_1	2, 1, 4	2, 3, 1	1, 1, 3	18	
	B_2	2, 1, 1	2, 2, 2	2, 1, 2	15	
	Total $y_{4..}$	11	12	10		33
Total $y_{..l}$		65	64	55	184	
Mean $\bar{y}_{..l}$		2.71	2.67	2.29	2.56	

- Analyse the data and compare the averages of ever born children for different levels of education.
- Is there any difference in the averages of ever born children of outside working mothers and house-wife mothers?
- How does the average children of mothers of medium and high socioeconomic status differ from that of mothers of low socioeconomic status?

Solution : (i) We have $p = 4$, $q = 2$, $r = 3$, $m = 3$, $G = 184$.

$$C.T. = \frac{G^2}{pqrm} = \frac{(184)^2}{72} = 470.2222.$$

$$SS(\text{Total}) = \sum \sum \sum \sum y_{ijkl}^2 - C.T. = 546 - 470.2222 = 75.7778.$$

$$SS(A) = \frac{\sum y_{i...}^2}{qrm} - C.T. = \frac{8882}{18} - 470.2222 = 23.2222.$$

$$SS(B) = \frac{\sum y_{.j..}^2}{prm} - C.T. = \frac{17378}{46} - 470.2222 = 12.50.$$

$$SS(C) = \frac{\sum y_{..l}^2}{pqm} - C.T. = \frac{11346}{24} - 470.2222 = 2.5278.$$

Number of ever born children of three classes of mothers (y_{ijl})

A	B	C			A	B	C		
		C ₁	C ₂	C ₃			C ₁	C ₂	C ₃
A ₁	B ₁	12	10	9	A ₃	B ₁	10	9	7
	B ₂	8	10	7		B ₂	4	5	4
A ₂	B ₁	12	11	9	A ₄	B ₁	7	6	5
	B ₂	8	7	9		B ₂	4	6	5

Number of ever born children of mothers according to working condition and socioeconomic condition ($y_{j.l}$)

B	C			Total $y_{j..}$
	C ₁	C ₂	C ₃	
B ₁	41	36	30	107
B ₂	24	28	25	77

$$SS(AB) = \frac{\sum \sum y_{ij.}^2}{rm} - \text{C.T.} - SS(A) - SS(B)$$

$$= \frac{4580}{9} - 470.2222 - 23.2222 - 12.50 = 2.9445.$$

$$SS(AC) = \frac{\sum \sum y_{i.l}^2}{rm} - \text{C.T.} - SS(A) - SS(C)$$

$$= \frac{2982}{6} - 470.2222 - 23.2222 - 2.5278 = 1.0278.$$

$$SS(BC) = \frac{\sum \sum y_{.jl}^2}{pm} - \text{C.T.} - SS(B) - SS(C)$$

$$= \frac{5862}{12} - 470.2222 - 12.50 - 2.5278 = 3.25.$$

$$SS(ABC) = \frac{\sum \sum \sum y_{ijl}^2}{m} - \text{C.T.} - SS(A) - SS(B) - SS(C) - SS(AB) - SS(AC) - SS(BC)$$

$$= \frac{1552}{3} - 470.2222 - 23.2222 - 12.50 - 2.5278 - 2.9445 - 1.0278 - 3.25$$

$$= 1.6388.$$

$$SS(\text{error}) = SS(\text{Total}) - SS(A) - SS(B) - SS(C) - SS(AB) - SS(AC) - SS(BC) - SS(ABC)$$

$$= 75.7778 - 23.2222 - 12.50 - 2.5278 - 2.9445 - 1.0278 - 3.25 - 1.6388$$

$$= 28.6667.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{.05}$	$F_{.09}$
A	3	23.2222	7.7407	12.96**	2.81	4.24
B	1	12.50	12.50	20.93**	4.05	7.22
C	2	2.5278	1.2639	2.12	3.20	5.10
AB	3	2.9445	1.3148	2.20	2.81	4.24
AC	6	1.0278	0.1713	0.29	2.30	3.22
BC	2	3.25	1.625	2.72	3.20	5.10
ABC	6	1.6388	0.2731	0.46	2.30	3.22
Error	48	28.6667	0.5972	—	—	—
Total	71					

**Indicates highly significant effect.

It is observed that due to the changes in the level of education of mothers average ever born children varies significantly [$F_1 = 12.96 > F_{0.05}$ and $F_{0.01}$]. The difference in the averages of house-wife mothers and outside working mothers is also highly significant [$F_2 = 20.93 > F_{0.05}$ and $F_{0.01}$].

To compare the pairwise means of ever born children for mothers of different levels of education the test statistic is

$$D_i = d_{0.05, i, f} \sqrt{\frac{s_8}{qrm}}, \quad i = 2, 3, 4; \quad f = 48$$

$$D_2 = d_{0.05, 2, 48} \sqrt{\frac{s_8}{qrm}} \approx 2.77 \sqrt{\frac{0.5972}{18}} = 0.50$$

$$D_3 = d_{0.05, 3, 48} \sqrt{\frac{s_8}{qrm}} \approx 2.92 \sqrt{\frac{0.5972}{18}} = 0.53$$

$$D_4 = d_{0.05, 4, 48} \sqrt{\frac{s_8}{qrm}} \approx 3.02 \sqrt{\frac{0.5972}{18}} = 0.55.$$

The means in ascending order are :

$$\bar{A}_4 = 1.83, \quad \bar{A}_3 = 2.17, \quad \bar{A}_1 = 3.11, \quad \bar{A}_2 = 3.11$$

$$\bar{A}_2 - \bar{A}_4 = 3.11 - 1.83 = 1.28 > D_4, \quad \therefore \text{means are different.}$$

$$\bar{A}_2 - \bar{A}_3 = 3.11 - 2.17 = 0.94 > D_3, \quad \therefore A_2 \text{ and } A_3 \text{ are different.}$$

$$\bar{A}_1 - \bar{A}_4 = 3.11 - 1.83 = 1.28 > D_3, \quad \therefore A_1 \text{ and } A_4 \text{ are different.}$$

$$\bar{A}_3 - \bar{A}_4 = 2.17 - 1.83 = 0.34 < D_2, \quad \therefore A_3 \text{ and } A_4 \text{ are similar.}$$

$$\bar{A}_1 - \bar{A}_3 = 3.11 - 2.17 = 0.94 > D_2, \quad \therefore A_1 \text{ and } A_3 \text{ are different.}$$

The underlined means are similar.

$$\underline{\bar{A}_4, \bar{A}_3} \quad \underline{\bar{A}_1, \bar{A}_2}$$

(ii) We need to test the significance of $H_0 : \beta_1 = \beta_2$, against $H_A : \beta_1 \neq \beta_2$.

The test statistic is $t = \frac{\bar{y}_{.1.} - \bar{y}_{.2.}}{\sqrt{\frac{2s_8}{pm}}}$.

This t is distributed as Student's t with $f = 48$ d.f. However, we do not need to calculate the value of t , since there are only two levels of B and by F -test these two levels are found significantly different.

(iii) We need to compare \bar{C}_2 and \bar{C}_3 with \bar{C}_1 . Considering C_1 as control treatment, the comparison is made by Dunnett's test, where the test statistic is

$$\begin{aligned} D &= d_{0.05, k-1, f} \sqrt{\frac{2s_8}{pqm}}, \quad k-1 = 2, \quad f = 48 \\ &= 2.282 \sqrt{\frac{2 \times 0.5972}{4 \times 2 \times 3}} = 0.50. \end{aligned}$$

However, the comparison is not needed, since by F -test the levels of C are found insignificant.

Chapter 3

Basic Design

3.1 Introduction

The analytical procedures of experimental data have been presented in the previous chapter. The data used in different analysis are not of the same type. The sources of variation in data are different. These sources are related to different factors. For example, in agricultural experiment if the objective is to identify a best variety of crop among a group of crop varieties, the experiment with varieties of crop can be performed in homogeneous experimental plots or the experiment can be conducted in homogeneous plots using different levels of fertilizers and different levels of irrigation. In the first case, experiment is conducted allocating the crop varieties completely randomly in the plots. The identified source of variation in the data is due to crop variety. The analysis of data of such experiment is known as *one-way classification*. In the second case a group of plots are selected to use a particular level of fertilizer and in that group crop varieties are randomly allocated. Several groups of plots are selected for several levels of fertilizer. In each group the crop varieties are randomly allocated. The analysis of data collected from such experiment is known as *two-way classification*.

From the above discussion, it is clear that the analysis of data depends on method of data collection by conducting experiments. The method of conducting the experiment is known as design of experiment. The basic designs are (i) Completely Randomized Design, (ii) Randomized Block Design and (iii) Latin Square Design. Whatever be the design used in any experiment, the experiment is conducted allocating the treatments to the plots or plots of a block by a random process.

3.2 Completely Randomized Design (CRD)

The simplest design of experiment is completely randomized design, where the treatments are allocated completely randomly in the experimental plots. Let us assume that we have k treatments for an experiment. These treatments are to be allocated in n homogeneous plots. The homogeneity in plots is considered in size, shape and in soil fertility in case of agricultural experiment. For experiment in medical research, the homogeneous plots may be guinea pigs of same age, same weight and same body condition score. Let us consider that the i th treatment ($i = 1, 2, \dots, k$) is to be replicated in n_i plots such that $\sum n_i = n$. The method of allocation of i th treatment to n_i plots at random for all values of $i = 1, 2, \dots, k$ is known as completely randomized design.

To allocate the treatment, first select n_1 plots from n plots by any of the random procedure and allocate first treatment to these selected n_1 plots. From the remaining $(n - n_1)$ plots select n_2 plots by a random process and allocate second treatment to these n_2 plots. In this way the k th treatment is allocated to the last selected n_k plots. This procedure of allocation of treatment to the plots is known as completely randomized design.

The treatments may be replicated equally or unequal times. Here n_i is the number of replication of i th treatment. However, if all treatments are equally replicated, the efficiency in estimating treatment effect is increased.

Analysis of Data : Since all the experimental plots used in the experiment are homogeneous, the only identified source of variation in the data obtained from such an experiment is due to the variation in the treatment. Let y_{ij} be the result of j th replication of i th treatment ($i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$). The linear model for y_{ij} observation is :

$$y_{ij} = \mu + \alpha_i + e_{ij},$$

where μ = general mean, α_i = effect of i th treatment and e_{ij} is random error arises in j th replication of i th treatment. It is assumed that $e_{ij} \sim NID(0, \sigma^2)$. The data are analysed under the restriction $\sum n_i \alpha_i = 0$. The analysis of data is similar to that of one-way classification.

ANOVA Table

Sources	d.f.	SS	$MS = \frac{SS}{df}$	F	$E(MS)$	P-value
Treatment	$k - 1$	S_1	$s_1 = \frac{S_1}{k - 1}$	$\frac{s_1}{s_2}$	$\sigma^2 + \frac{1}{k - 1} \sum n_i \alpha_i^2$	$\int_F^\infty f(F) dF$
Error	$n - k$	S_2	$s_2 = \frac{S_2}{n - k}$	—	σ^2	
Total	$n - 1$					

The objective of the analysis is to test the significance of the null hypothesis :

$$H_0 : \alpha_i, \text{ against } H_A : \alpha_i \neq 0$$

The null hypothesis is rejected if $F \geq F_{\alpha; k-1, n-k}$. The rejection of null hypothesis leads to compare the treatments in pairs by Duncan's multiple range test, where

$$D_i = d_{\alpha, i, f} \sqrt{\frac{s_2}{n_H}},$$

where $i = 2, 3, \dots, k$ and n_H = harmonic mean of n_i .

Example 3.1 : To observe the heart beat/minute of *Milux rusticus* slug under different types of pesticide an experiment is conducted in a laboratory. The slugs are kept in pesticide for a week. After a week the heart rates are measured.

Treatment : Pesticide, T_i	Heart beat of slugs per minute (y_{ij})	Total y_i	Mean \bar{y}_i
Control	11, 11, 12, 10, 12, 10	66	11.00
Supracide	9, 9, 9, 8, 9, 9	53	8.83
Pomex	10, 10, 8, 9, 8, 9	54	9.00
Sumicidin	7, 7, 7, 7, 8, 6	43	7.17

- (i) Analyse the data and comment on the impacts of pesticide
- (ii) Group the homogeneous pesticides.
- (iii) Do the pesticides differ from control?

Solution : (i) We have $n = 6, k = 4, G = 216, C.T. = \frac{G^2}{nk} = \frac{(216)^2}{6 \times 4} = 1944.00$.

$$SS (\text{Total}) = \sum \sum y_{ij}^2 - C.T. = 2000 - 1944.00 = 56.00.$$

$$SS (\text{Treatment}) = \sum \frac{y_i^2}{n_i} - \text{C.T.} = \frac{11930}{6} - 1944.00 = 44.33.$$

$$SS (\text{Error}) = SS (\text{Total}) - SS (\text{Treatment}) = 56.00 - 44.33 = 11.67.$$

ANOVA Table

Sources of Variation	d.f.	SS	MS = $\frac{SS}{df}$	F	$F_{0.05}$	$F_{0.01}$
Treatment	3	44.33	14.7767	25.32	3.10	4.94
Error	20	11.67	0.5835			
Total	23					

The hypothesis is $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ or, $H_0 : \alpha_i = 0$, $H_A : \alpha_i \neq 0$.

Since $F = 25.32 > F_{0.05}$ and $F_{0.01}$, H_0 is rejected. The pesticides are highly significantly different (P -value < 0.00).

(ii) The grouping of pesticides is done by Duncan's multiple range test, where the test statistic is :

$$D_i = d_{0.05, i, f} \sqrt{\frac{s_2}{n}}, \text{ where } f = 20; i = 2, 3, 4.$$

$$D_2 = 2.97 \sqrt{\frac{0.5835}{6}} = 0.93, \quad D_3 = 3.12 \sqrt{\frac{0.5835}{6}} = 0.97, \quad D_4 = 3.21 \sqrt{\frac{0.5835}{6}} = 1.00.$$

This means in ascending order are : $\bar{T}_4 = 7.17$, $\bar{T}_2 = 8.83$, $\bar{T}_3 = 9.00$, $\bar{T}_1 = 11.00$.

$$\bar{T}_1 - \bar{T}_4 = 11.00 - 7.17 = 3.83 > D_4, \quad \therefore \text{the pesticides are different.}$$

$$\bar{T}_1 - \bar{T}_2 = 11.00 - 8.83 = 2.17 > D_3, \quad \therefore T_1 \text{ and } T_2 \text{ are different.}$$

$$\bar{T}_3 - \bar{T}_4 = 9.00 - 7.17 = 1.83 > D_3, \quad \therefore T_3 \text{ and } T_4 \text{ are different.}$$

$$\bar{T}_2 - \bar{T}_4 = 8.83 - 7.17 = 1.66 > D_2, \quad \therefore T_2 \text{ and } T_4 \text{ are different.}$$

$$\bar{T}_3 - \bar{T}_2 = 9.00 - 8.83 = 0.17 < D_2, \quad \therefore T_2 \text{ and } T_3 \text{ are similar.}$$

$$\bar{T}_1 - \bar{T}_3 = 11.00 - 9.00 = 2.00 > D_2, \quad \therefore T_1 \text{ and } T_3 \text{ are different.}$$

The underlined means are in one group \bar{T}_4 , \bar{T}_2 , \bar{T}_3 , \bar{T}_1 .

(iii) To test the significance in the difference between control and other pesticides we need to use Dunnett's test, where the test statistic is :

$$\begin{aligned} D &= d_{0.05, k-1, f} \sqrt{\frac{2s_2}{n}}, \quad n = 6, \quad f = 20, \quad k - 1 = 3 \\ &= 2.57 \sqrt{\frac{2 \times 0.5835}{6}} = 1.13. \end{aligned}$$

Now, $|\bar{T}_1 - \bar{T}_2| = |11.00 - 8.83| = 2.17$, \therefore supracide (T_2) is better than control in reducing heart beat.

$$|\bar{T}_1 - \bar{T}_3| = |11.00 - 9.00| = 2.00, \quad \therefore \text{pomex is better than control.}$$

$$|\bar{T}_1 - \bar{T}_4| = |11.00 - 7.17| = 3.83, \quad \therefore \text{sumicidin is better than control in reducing heart beat.}$$

Advantages and Disadvantages of CRD

Advantages :

- (i) It is easy to use completely randomized design in practice.
- (ii) The d.f. of error of this design becomes large since total d.f. is divided into only two parts, one for treatment and one for error.
- (iii) If homogeneous plots are available, the design can be used for any number of treatments.
- (iv) The analysis of data of this design is simple and easy. Even in case of one or more missing observations the analysis is not complicated.

Disadvantages :

- (i) If the plots used in the experiment are not homogeneous, the design is not suitable.
- (ii) If number of treatments are large, large number of homogeneous plots are needed to conduct the experiment. As a result, there is chance to lose the homogeneity of the plots.
- (iii) The design is not suitable for agricultural experiment in the field.

Uses of CRD :

- (i) The design is very much suitable in conducting experiment in the laboratory.
- (ii) It is suitable in conducting experiment in dairy science research.
- (iii) The design is also used in green-house experiment.

3.3 Randomized Block Design (RBD)

The experimental plots used in any experiment may not always be homogeneous. For example, in dairy science research on feeding trial the different types of food are to be given to different cows of different lactation periods or different types. If the objective of the experiment is to identify the best animal food for increased milk production, the different types of food must be given to cows of same lactation period, since the milk production varies with the variation in lactation period. However, in practice varieties of food are given to cows of different lactation periods. A group of cows of a particular lactation period constitute a block. Different blocks may be used in an experiment. Due to blocks, one-directional external source of variation prevails in the experimental plots and the source is block. Here blocks are formed perpendicular to the source of variation. Treatments are allocated to the plots of a block by a random process. The resultant design is known as randomized block design. In feeding trial experiment varieties of food are given randomly to the cows of any one lactation period.

Let us consider that we have q treatments under investigation. These are to be replicated in p blocks of q plots each. The q treatments are to be allocated in q plots of a block. Similar allocation procedure is followed in all p blocks if there are p blocks for the experiment. The resultant design is known as randomized block design.

Analysis of Data Collected from RBD : Let y_{ij} be the experimental result of j th treatment in i th block. The model for y_{ij} ($i = 1, 2, \dots, p; j = 1, 2, \dots, q$) observations is

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

where μ = general mean, α_i = effect of i th block, β_j = effect of j th treatment, and e_{ij} = random error. The assumption to analyse the data is $e_{ij} \sim NID(0, \sigma^2)$.

The main objective of the analysis is to test the significance of the hypothesis :

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_q \quad \text{or} \quad H_0 : \beta_j = 0, \quad j = 1, 2, \dots, q.$$

The analytical procedure has been discussed in two-way classification. The analysis of variance table is shown below :

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$	P-value
Block	$p - 1$	S_1	s_1	$F_1 = \frac{s_1}{s_2}$	$\sigma^2 + \frac{q}{p-1} \sum \alpha_i^2$	$\int_{F_1}^{\infty} f(F) dF$
Treatment	$q - 1$	S_2	s_2	$F_2 = \frac{s_2}{s_3}$	$\sigma^2 + \frac{p}{q-1} \sum \beta_i^2$	$\int_{F_2}^{\infty} f(F) dF$
Error	$(p-1)(q-1)$	S_3	s_3			
Total	$pq - 1$					

If $F_2 \geq F_{\alpha; q-1, (p-1)(q-1)}$, H_0 is rejected at $100\alpha\%$ level of significance.

The rejection of null hypothesis leads to compare the treatments in pairs. This can be done by Duncan's multiple range test, where the test statistic is :

$$D_j = d_{\alpha, j, f} \sqrt{\frac{s_3}{p}}, \quad j = 2, 3, \dots, q-1; \quad f = (p-1)(q-1).$$

Example 3.2 : In an agricultural research station an experiment is conducted to identify the best variety of wheat. Three varieties of newly discovered wheat along with an old variety (W_1) are used in the experiment. The wheat varieties are cultivated using 5 doses of nitrogen as urea. The design used in the experiment is RBD. The production (kg/hectare) data are given below :

Levels of urea	Production of wheat varieties (y_{ij} , kg/hectare)					
	W_1	W_2	W_3	W_4	Total $y_{.j}$	Mean $\bar{y}_{.j}$
N_1	2620	2840	2625	2600	10785	2696.25
N_2	2715	2912	2715	2718	11060	2765.00
N_3	2600	2818	2810	2750	10978	2744.50
N_4	2735	2950	2615	2610	10910	2727.50
N_5	2848	3020	2870	2700	11438	2859.50
Total $y_{.j}$	13618	14540	13635	13078	54871	
Mean $\bar{y}_{.j}$	2723.60	2908.00	2727.00	2615.60	2743.55	

- (i) Analyse the data and group the wheat varieties.
- (ii) Is there any difference in N_4 and N_5 ?
- (iii) Are the new wheat varieties better than the old variety W_1 ?

Solution : (i) We have $p = 5$, $q = 4$, $G = 54871$, C.T. = $\frac{G^2}{pq} = \frac{(5487)^2}{5 \times 4} = 150541332.05$

$$SS (\text{Total}) = \sum \sum y_{ij}^2 - \text{C.T.} = 152475521 - 150541332.05 = 1934188.95.$$

$$SS (\text{Block}) = \sum \frac{y_{.j}^2}{q} - \text{C.T.} = \frac{609012253}{4} - 150541332.05 = 1711731.2$$

$$SS (\text{Wheat}) = \sum \frac{y_j^2}{p} - \text{C.T.} = \frac{753808833}{5} - 150541332.05 = 220434.55$$

$$SS (\text{Error}) = SS (\text{Total}) - SS (\text{Block}) - SS (\text{Wheat}) \\ = 1934188.95 - 1711731.20 - 220434.55 = 2023.20$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{\text{d.f.}}$	F	F _{0.05}	F _{0.01}	P-value
Block (Urea)	4	1711731.20	427932.8	2538.15	3.26	5.41	0.00
Treatment	3	220434.55	73478.18	435.81	3.49	5.95	0.00
Error	12	2023.20	168.6				
Total	19						

Since $F_2 = 435.81 > F_{0.05}$ and $F_{0.01}$, the wheat varieties are highly significantly different ($P\text{-value} < 0.00$). The wheat varieties can be grouped as follows :

$$D_j = d_{0.05, j, f} \sqrt{\frac{s_3}{p}}, \quad j = 2, 3, 4; \quad f = 12.$$

$$D_2 = 3.08 \sqrt{\frac{168.6}{5}} = 17.88, \quad D_3 = 3.23 \sqrt{\frac{168.6}{5}} = 18.76, \quad D_4 = 3.33 \sqrt{\frac{168.6}{5}} = 19.34.$$

The means in ascending order are :

$$\bar{W}_4 = 2615.60, \quad \bar{W}_1 = 2723.60, \quad \bar{W}_3 = 2727.00, \quad \bar{W}_2 = 2908.00$$

$$\bar{W}_2 - \bar{W}_4 = 2908.00 - 2615.60 = 292.40 > D_4, \quad \therefore \text{the wheat varieties are different}$$

$$\bar{W}_4 - \bar{W}_1 = 2908.00 - 2723.60 = 184.40 > D_3, \quad \therefore W_1 \text{ and } W_4 \text{ are different.}$$

$$\bar{W}_3 - \bar{W}_4 = 2727.00 - 2615.60 = 111.40 > D_3, \quad \therefore W_3 \text{ and } W_4 \text{ are different.}$$

$$\bar{W}_1 - \bar{W}_4 = 2723.60 - 2615.60 = 108.00 > D_2, \quad \therefore W_1 \text{ and } W_4 \text{ are different.}$$

$$\bar{W}_3 - \bar{W}_1 = 2727.00 - 2723.60 = 3.40 < D_2, \quad \therefore W_1 \text{ and } W_3 \text{ are similar.}$$

$$\bar{W}_2 - \bar{W}_3 = 2908.00 - 2727.00 = 181.00 > D_2, \quad \therefore W_2 \text{ and } W_3 \text{ are different.}$$

The underlined means are in one group : $\bar{W}_4, \bar{W}_1, \bar{W}_3, \bar{W}_2$.

(ii) We need to test the significance of the hypothesis $H_0 : \alpha_4 = \alpha_5$, against $H_A : \alpha_4 \neq \alpha_5$.

$$\text{The test statistic is : } t = \frac{\bar{y}_4 - \bar{y}_5}{\sqrt{\frac{2s_3}{q}}} = \frac{2725.50 - 2859.50}{\sqrt{\frac{2 \times 168.6}{4}}} = -14.59.$$

Since $|t| > t_{0.025, 12} = 2.179$, H_0 is rejected. The levels of urea 4 and 5 are significantly different.

(iii) To compare W_1 with other wheat varieties we use Dunnett's test, where the test statistic is :

$$D = d_{0.05, k-1, f} \sqrt{\frac{2s_3}{p}}, \quad k-1 = 3, \quad f = 12, \quad p = 5 \\ = 2.72 \sqrt{\frac{2 \times 168.6}{5}} = 28.19.$$

Now, $|\bar{W}_1 - \bar{W}_2| = |2723.60 - 2908.00| = 184.4 > D$, W_2 is better than W_1 .

$|\bar{W}_1 - \bar{W}_3| = |2723.60 - 2727.00| = 3.4 < D$, W_1 and W_3 are similar.

$|\bar{W}_1 - \bar{W}_4| = |2723.60 - 2615.60| = 108.0 > D$, W_1 and W_4 are different.

It is observed that W_2 is significantly better than W_1 , but W_1 is significantly better than W_4 . The varieties W_1 and W_3 are similar.

3.4 Randomized Block Design with More than One (Equal) Observations Per Cell

The randomized block design what we have considered in the previous section contains q plots to allocate q treatments. Let us consider that each animal food is fed to several cows of same lactation period and different types of food are fed to several cows (number of cows are equal for each food) of same lactation period.

Consider that we have cows of p lactation periods and total number of cows are pqr , where qr cows are available in each lactation period. Also consider that, we have q different types of food. Each food is fed to r cows of any lactation period. The types of food are randomly allocated to the cows of each lactation period. The resultant design is randomized block design with r observations per cell.

Let y_{ijl} be the l th result of j th treatment in i th block ($i = 1, 2, \dots, p$; $j = 1, 2, \dots, q$; $l = 1, 2, \dots, r$). The model for y_{ijl} observations is :

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl}$$

where μ = general mean, α_i = effect of i th level of A , β_j = effect j th level of B (treatment), $(\alpha\beta)_{ij}$ = interaction of j th treatment with i th block, e_{ijl} = random error.

The assumption to analyse the data is $e_{ijl} \sim NID(0, \sigma)$. The different steps of analysis are presented in section (2.3). The analysis of variance table is shown below :

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$	P-value
Block (A)	$p - 1$	S_1	s_1	$F_1 = \frac{s_1}{s_4}$	$\sigma^2 + \frac{qr}{p-1} \sum \alpha_i^2$	P_1
Treatment	$q - 1$	S_2	s_2	$F_2 = \frac{s_2}{s_4}$	$\sigma^2 + \frac{pr}{q-1} \sum \beta_j^2$	P_2
Block \times treatment (AB)	$(p-1)(q-1)$	S_3	s_3	$F_3 = \frac{s_3}{s_4}$	$\sigma^2 + \frac{r \sum \sum (\alpha\beta)_{ij}^2}{(p-1)(q-1)}$	P_3
Error	$pq(r-1)$	S_4	s_4	—	σ^2	—
Total	$pqr - 1$					

Here $P_k = \int_{F_k}^{\infty} f(F) dF$. If $P(F_k) \leq \alpha$ ($k = 1, 2, 3$), the corresponding hypothesis is rejected at $100\alpha\%$ level of significance, or if $F_k \leq F_\alpha$ with corresponding d.f., H_0 is rejected at $100\alpha\%$ level of significance. The multiple comparisons are also to be performed in a similar way as these are discussed in section (2.3).

Example 3.3 : Nair et al. (1993) have conducted an experiment through RBD to study the heart beat/minute of 4 different types of slug, where slugs are kept under 6 varieties of pesticide. The heart beat/minute are shown below :

Slugs	Pesticides						Total $y_{i..}$	Mean $\bar{y}_{i..}$
	Control (P_1)	Supracid (P_2)	Pomex (P_3)	Sumicidin (P_4)	Glyphosphate (P_5)	Milkrup (P_6)		
S_1	14, 11	9, 8	9, 10	9, 9	7, 8	9, 10	113	9.42
S_2	13, 13	8, 9	9, 9	9, 9	7, 8	9, 8	111	9.25
S_3	12, 14	8, 8	11, 10	10, 10	11, 12	13, 13	132	11.00
S_4	12, 13	9, 8	11, 11	10, 9	12, 12	12, 13	132	11.00
Total $y_{.j.}$	102	67	80	75	77	87	488	
Mean $\bar{y}_{.j.}$	12.75	8.375	10.00	9.375	9.625	10.875	10.17	

- (i) Analyse the data and group the pesticides
- (ii) Is there any difference between S_1 and S_3 ?
- (iii) Do you think that the pesticides are useful in reducing the heart rate?

Solution : (i) We have $r = 2, p = 4, q = 6, G = 488, C.T. = \frac{G^2}{pqr} = \frac{(488)^2}{48} = 4961.3333.$

$$SS \text{ (Total)} = \sum \sum \sum y_{ijl}^2 - C.T. = 5138 - 4961.3333 = 176.6667 .$$

$$SS \text{ (Slugs)} = \sum \frac{y_{i..}^2}{qr} - C.T. = \frac{59938}{6 \times 2} - 4961.3333 = 33.5$$

$$SS \text{ (Pesticide)} = \sum \frac{y_{.j.}^2}{pr} - C.T. = \frac{40416}{4 \times 2} - 4961.3333 = 90.6667$$

The observations of Slugs and Pesticides ($y_{ij.}$) :

Slugs	Pesticides					
	P_1	P_2	P_3	P_4	P_5	P_6
S_1	25	17	19	18	15	19
S_2	26	17	18	18	15	17
S_3	26	16	21	20	23	26
S_4	25	17	22	19	24	25

$$SS \text{ (Slugs} \times \text{Pesticides)} = \sum \sum \frac{y_{ij.}^2}{r} - C.T. - SS \text{ (Slugs)} - SS \text{ (Pesticides)}$$

$$= \frac{10250}{2} - 4961.3333 - 33.5 - 90.6667 = 39.5.$$

$$SS \text{ (Error)} = SS \text{ (Total)} - SS \text{ (Pesticides)} - SS \text{ (Slugs)} - SS \text{ (Slugs} \times \text{Pesticides)}$$

$$= 176.6667 - 33.5 - 90.6667 - 39.5 = 13.00.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	P-value
Slugs	3	33.50	11.1667	20.61	3.01	0.00
Pesticides	5	90.6667	18.1333	33.47	2.62	0.00
Slugs × Pesticides	15	39.50	2.6333	4.86	2.13	< 0.01
Error	24	13.00	0.5417	—	—	—
Total	47					

Since $P(F_1) < 0.01$, the slugs are highly significantly different in respect of heart rate under pesticides. $P(F_2) < 0.01$ also indicates highly significant differences in pesticides in reducing heart rates of slugs. Different types of slug behave differently in presence of different pesticides. This is concluded from $P(F_3) < 0.01$.

The pesticides are different, but some of these may be of the similar type in behaviour in reducing heart rate of slugs. This can be investigated by grouping the mean heart rates recorded under different pesticides. The test statistic is :

$$D_j = d_{0.05, j, f} \sqrt{\frac{s_4}{pr}}, \quad j = 2, 3, 4, 5; \quad f = 24$$

$$D_2 = 2.95 \sqrt{\frac{0.5417}{4 \times 2}} = 0.77, \quad D_3 = 3.10 \sqrt{\frac{0.5417}{4 \times 2}} = 0.81, \quad D_4 = 3.18 \sqrt{\frac{0.5417}{4 \times 2}} = 0.83,$$

$$D_5 = 3.25 \sqrt{\frac{0.5417}{4 \times 2}} = 0.85.$$

The means in ascending order are :

$$\bar{P}_2 = 8.375, \bar{P}_4 = 9.375, \bar{P}_5 = 9.625, \bar{P}_3 = 10.00, \bar{P}_6 = 10.875, \bar{P}_1 = 12.75.$$

$$\bar{P}_1 - \bar{P}_2 = 12.75 - 8.375 = 4.375 > D_5, \quad \therefore \text{means are significantly different.}$$

$$\bar{P}_1 - \bar{P}_4 = 12.75 - 9.375 = 3.375 > D_4, \quad \therefore P_1 \text{ and } P_4 \text{ are different.}$$

$$\bar{P}_6 - \bar{P}_2 = 10.875 - 8.375 = 2.50 > D_4, \quad \therefore P_2 \text{ and } P_6 \text{ are different.}$$

$$\bar{P}_1 - \bar{P}_3 = 12.75 - 10.00 = 2.75 > D_3, \quad \therefore P_1 \text{ and } P_3 \text{ are different.}$$

$$\bar{P}_6 - \bar{P}_5 = 10.875 - 9.625 = 1.25 > D_3, \quad \therefore P_5 \text{ and } P_6 \text{ are different.}$$

$$\bar{P}_3 - \bar{P}_4 = 10.00 - 9.375 = 0.625 < D_3, \quad \therefore P_3, P_4 \text{ and } P_5 \text{ are similar.}$$

$$\bar{P}_5 - \bar{P}_2 = 9.625 - 8.375 = 1.25 > D_3, \quad \therefore P_2 \text{ and } P_5 \text{ are different.}$$

$$\bar{P}_4 - \bar{P}_2 = 9.375 - 8.375 = 1.00 > D_2, \quad \therefore P_2 \text{ and } P_4 \text{ are different.}$$

$$\bar{P}_5 - \bar{P}_4 = 9.625 - 9.375 = 0.25 < D_2, \quad \therefore P_4 \text{ and } P_5 \text{ are similar.}$$

$$\bar{P}_6 - \bar{P}_3 = 10.875 - 10.00 = 0.875 > D_2, \quad \therefore P_3 \text{ and } P_6 \text{ are different.}$$

$$\bar{P}_1 - \bar{P}_6 = 12.75 - 10.875 = 1.875 > D_2, \quad \therefore P_1 \text{ and } P_6 \text{ are different.}$$

Similar pesticides are shown by underline : $\bar{P}_2, \underline{\bar{P}_4}, \underline{\bar{P}_5}, \bar{P}_3, \bar{P}_6, \bar{P}_1$.

(ii) We need to test the significance of $H_0 : \alpha_1 = \alpha_3$, against $H_A : \alpha_1 \neq \alpha_3$.

$$\text{The test statistic is } t = \frac{\bar{y}_{1..} - \bar{y}_{3..}}{\sqrt{\frac{2s_4}{qr}}} = \frac{9.42 - 11.00}{\sqrt{\frac{2 \times 0.5417}{6 \times 2}}} = -5.26.$$

Since $|t| > t_{0.025, 24} = 2.064$, H_0 is rejected. Slug S_1 and slug S_3 are significantly different.

(iii) Pesticides will be useful if heart rates are reduced when it is used. So any pesticide is to be compared with control. This can be done by Dunnett's test, where the test statistic is :

$$D = d_{0.05, k-1, f} \sqrt{\frac{2s_4}{pr}}, \quad k - 1 = 5, \quad f = 24.$$

$$D = 2.36 \sqrt{\frac{2 \times 0.5417}{4 \times 2}} = 0.87.$$

$$|\bar{P}_1 - \bar{P}_2| = |12.75 - 8.375| = 4.375, \quad \therefore P_2 \text{ is better than } P_1.$$

$$|\bar{P}_1 - \bar{P}_3| = |12.75 - 10.00| = 2.75, \quad \therefore P_3 \text{ is better than } P_1.$$

$$|\bar{P}_1 - \bar{P}_4| = |12.75 - 9.375| = 3.375, \quad \therefore P_4 \text{ is better than } P_1.$$

$$|\bar{P}_1 - \bar{P}_5| = |12.75 - 9.625| = 3.125, \quad \therefore P_5 \text{ is better than } P_1.$$

$$|\bar{P}_1 - \bar{P}_6| = |12.75 - 10.875| = 1.875, \quad \therefore P_6 \text{ is better than } P_1.$$

Therefore, pesticides are useful in reducing heart beat of slug.

3.5 Efficiency of Randomized Block Design

It has already been mentioned in section (3.2) that the CRD is simple to apply if all experimental units are homogeneous. But the design is less used in field experiment since the chance of heterogeneity in field plots is more. To avoid the problem, blocks are formed with the plots of homogeneous type, and treatments are easily allocated randomly to the plots of blocks. However, if homogeneous plots are available, CRD can also be used in field experiment instead of RBD. Therefore, the efficiency of randomized block design needs to be studied compared to the latter design, if the efficiency of the former design is not sufficiently large compared to the latter design there is no use of RBD utilising more experimental resources in terms of money and time.

Let us consider that we have pq plots which are grouped into p blocks each of q plots. Separate randomisation is done in allocating q treatments in q plots of any block. the analysis of variance table of data collected from such an experiment is shown below :

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$
Block	$p - 1$	S_1	s_1
Treatment	$q - 1$	S_2	s_2
Error	$(p - 1)(q - 1)$	S_3	s_3
Total	$pq - 1$		

Let us consider that the treatment effect is insignificant and the treatment variance is, on an average, equal to the error variance, then the analysis of variance table takes the following shape :

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$
Block	$p - 1$	S_1	s_1
Error	$p(q - 1)$	$p(q - 1)s_3$	s_3
Total	$pq - 1$		

In such analysis, if block comparison is not made, the error mean square stands as

$$\frac{p(q - 1)s_3 + (p - 1)s_1}{pq - 1}$$

If block comparison is not made, the variance of treatment effect will be proportional to the above variance and if block comparison is made it will be proportional to s_3 . The above variance is the error variance of CRD if treatment effects are assumed to be insignificant. Therefore, the efficiency of RBD compared to CRD is :

$$\frac{(p - 1)s_1 + p(q - 1)s_3}{(pq - 1)s_3}$$

If the impact of heterogeneity in plots is removed by blocking, the block variance will be more than error variance and efficiency of RBD will be 100 per cent or more compared to that of CRD. The CRD will be more efficient if plot heterogeneity is not removed by blocking.

Example 3.4 : In example 3.2, it is observed that $p = 5$, $q = 4$, $s_3 = 168.6$, $s_1 = 427932.8$. Find the efficiency of randomized block design compared to completely random design.

Solution : The efficiency of RBD compared to CRD is

$$\frac{(p - 1)s_1 + p(q - 1)s_3}{(pq - 1)s_3} = \frac{(5 - 1)427932.8 + 5(4 - 1)168.6}{(20 - 1)168.6} = 535.14.$$

Here RBD is 53514% efficient compared to CRD.

3.6 Advantages, Disadvantages and Uses of Randomized Block Design

Advantages :

- (i) Since RBD controls one-directional external source of variation in experimental plots, the design is more efficient than completely randomized design.
- (ii) Since block sum of squares is subtracted in calculating error sum of squares, the error sum of squares and hence the estimate of error variance are reduced.
- (iii) No restriction is needed in block numbers and treatment numbers.
- (iv) A particular treatment may be replicated more times than other by allocating it in a block. However, the analysis of data of such experiment is not complicated much.
- (v) In field experiment the blocks are not needed to be adjacent.
- (vi) Since treatments are randomly allocated to plots of a block, separate randomisation for different blocks is not harmful.

(vii) The analysis of data is easy and simple. Even it is not complicated if one or two observations are missing.

Disadvantages :

- (i) If the number of treatments is too large, homogeneity of plots within a block may be lost due to large number of plots of a block. In that case design will not be suitable.
- (ii) The design is not advantages if block heterogeneity is too much.

Uses :

- (i) The design is used profitably in agricultural experiment.
- (ii) The design is also used in many laboratory experiments.
- (iii) It is used in a situation where one directional external source of variation is needed to be controlled by design.

3.7 Randomized Block Design with Missing Observation(s)

Let us consider that the result of j th treatment in i th block is missing ($i = 1, 2, \dots, p$; $j = 1, 2, \dots, q$). Let this observation be x . In analysing data collected from randomized block design we need to estimate x so that the estimated error sum of squares is minimum, where the usual sum of squares due to error is

$$\phi = SS \text{ (error)} = \sum_{i=1}^p \sum_{j=1}^q y_{ij}^2 - \sum_{i=1}^p \frac{y_i^2}{q} - \sum_{j=1}^q \frac{y_j^2}{p} + \frac{G^2}{pq}$$

Let $(B_i + x)$ be the total of i th block in which y_{ij} is missing. The total of j th treatment is $(T_j + x)$ since it has one missing observation. The grand total of all observations can be written as $(G + x)$, where G is the total without y_{ij} observation. Now

$$\phi = \sum_{i' \neq i}^p \sum_{j' \neq j}^q y_{i'j'}^2 + x^2 - \sum_{i' \neq i}^p \frac{y_{i'}^2}{q} - \frac{(B_i + x)^2}{q} - \sum_{j' \neq j}^q \frac{y_{j'}^2}{p} - \frac{(T_j + x)^2}{p} + \frac{(G + x)^2}{pq}$$

We need to find the value of x so that ϕ is minimum. The value of x is found out such that

$$\frac{\partial \phi}{\partial x} = 0 \quad \text{or,} \quad 2x - \frac{2(B_i + x)}{q} - \frac{2(T_j + x)}{p} + \frac{2(G + x)}{pq} = 0 \quad \text{or,} \quad x = \frac{PB_i + qT_j - G}{(p - 1)(q - 1)}$$

The missing observation is to be replaced by its estimate x and then the analysis is done as usual except that 1 is subtracted from total d.f. and hence, from d.f. of error for one missing observation. The analysis of variance table takes the following shape :

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F
Block	$p - 1$	S_1	s_1	$F_1 = \frac{s_1}{s_3}$
Treatment	$q - 1$	S_2	s_2	$F_2 = \frac{s_2}{s_3}$
Error	$(p - 1)(q - 1) - 1$	S_3	s_3	—
Total	$pq - 2$			

The analytical conclusion is to be drawn in a similar way as it is done in the analysis without missing observation. However, the comparison of j th and j' th ($j' \neq j = 1, 2, \dots, q$) treatment is done by t -test in modified form, since $V(\hat{\beta}_j - \hat{\beta}_{j'}) \neq \frac{2\sigma^2}{p}$. Here

$$V(\hat{\beta}_j - \hat{\beta}_{j'}) = V(\bar{y}_{.j} - \bar{y}_{.j'}) = \frac{\sigma^2}{p} \left[2 + \frac{q}{(p-1)(q-1)} \right].$$

This is shown below :

$$\text{We have } V(\bar{y}_{.j'}) = \frac{\sigma^2}{p}.$$

$$\text{But } V(\bar{y}_{.j}) = \frac{1}{p^2} V \left[\frac{pB_i + qT_j - G}{(p-1)(q-1)} + T_j \right].$$

$$\text{Here } \bar{y}_{.j} = \frac{B_i + x}{p}.$$

$$\text{Now } V(\bar{y}_{.j}) = \frac{1}{p^2} \left[\frac{V(pB_i + qT_j - G)}{(p-1)^2(q-1)^2} + V(T_j) + 2\text{Cov} \left\{ T_j, \frac{pB_i + qT_j - G}{(p-1)(q-1)} \right\} \right].$$

$$\begin{aligned} \text{Again, } V[pB_i + qT_j - G] &= p^2V(B_i) + q^2V(T_j) + V(G) + 2pq\text{Cov}(B_j, T_j) \\ &\quad - 2p\text{Cov}(B_i, G) - 2q\text{Cov}(T_j, G) \\ &= [p^2(q-1) + q^2(p-1) + (pq-1) - 2p(q-1) - 2q(p-1)]\sigma^2 \\ &= (p-1)(q-1)(p+q-1)\sigma^2. \end{aligned}$$

$$V(T_j) = (p-1)\sigma^2$$

$$\begin{aligned} \text{Cov} \left[T_j, \frac{pB_i + qT_j - G}{(p-1)(q-1)} \right] &= \frac{1}{(p-1)(q-1)} [\text{Cov}(T_j, B_i) + qV(T_j) - \text{Cov}(T_j, G)] \\ &= \frac{\sigma^2}{(p-1)(q-1)} [q(p-1) - (p-1)] = \sigma^2. \end{aligned}$$

$$\therefore V(\bar{y}_{.j}) = \frac{\sigma^2}{p^2} \left[\frac{p+q-1}{(p-1)(q-1)} + (p-1) + 2 \right] = \frac{\sigma^2}{p} \left[1 + \frac{q}{(p-1)(q-1)} \right].$$

$$\begin{aligned} \text{Now } V(\bar{y}_{.j} - \bar{y}_{.j'}) &= V(\bar{y}_{.j}) + V(\bar{y}_{.j'}) - 2\text{Cov}(\bar{y}_{.j} - \bar{y}_{.j'}) \\ &= \frac{\sigma^2}{p} \left[1 + \frac{q}{(p-1)(q-1)} \right] + \frac{\sigma^2}{p} = \frac{\sigma^2}{p} \left[2 + \frac{q}{(p-1)(q-1)} \right]. \end{aligned}$$

Here in estimating $V(\bar{y}_{.j} - \bar{y}_{.j'})$ the variance σ^2 is to be replaced by mean square error s_3 (say).

Therefore, the test statistic to test the significance of $H_0 : \beta_j = \beta_{j'}$, against $H_A : \beta_j \neq \beta_{j'}$, $j \neq j' = 1, 2, \dots, q$ is :

$$t = \frac{\bar{y}_{.j} - \bar{y}_{.j'}}{\sqrt{\frac{s_3}{p} \left[2 + \frac{q}{(p-1)(q-1)} \right]}}$$

This ' t ' follows Student's ' t ' distribution with $(p-1)(q-1) - 1$ d.f. The conclusion will be drawn as usual.

Analysis with Two Missing Observations : Let the result of j th treatment in i th block be missing. Let us denote this observation by x . Consider also that the result of j' th treatment in i' th block ($i \neq i', j \neq j'$) is missing and this observation is y . The value of x and y are to be estimated in such a way that the estimated error sum of squares is minimum. The estimated error sum of squares can be written as under :

$$\phi = SS \text{ (error)} = \sum_{i''=1}^p \sum_{j''=1}^q y_{i''j''}^2 + x^2 + y^2 - \sum_{i''=1}^p \frac{y_{i''}^2}{q} - \sum_{j''=1}^q \frac{y_{j''}^2}{p} - \frac{(B_i + x)^2}{q} - \frac{(B_{i'} + y)^2}{q} - \frac{(T_j + x)^2}{p} - \frac{(T_{j'} + y)^2}{p} + \frac{(G + x + y)^2}{pq}$$

where B_i = total of i th block without x , $B_{i'}$ = total of i' th block without y , T_j = total j th treatment, without x , $T_{j'}$ = total of j' th treatment without y , G = grand total without x and y .

The values of x and y are to be found out from the equations : $\frac{\partial \phi}{\partial x} = 0$ and $\frac{\partial \phi}{\partial y} = 0$.

$$\text{Here } \frac{\partial \phi}{\partial x} = 2x - \frac{2(B_i + x)}{q} - \frac{2(T_j + x)}{p} + \frac{2(G + x + y)}{pq} = 0$$

$$\frac{\partial \phi}{\partial y} = 2y - \frac{2(B_{i'} + y)}{q} - \frac{2(T_{j'} + y)}{p} + \frac{2(G + x + y)}{pq} = 0.$$

On the simplification we get

$$x = \frac{(p-1)(q-1)[PB_i + qT_j - G] - (PB_{i'} + qT_{j'} - G)}{(p-1)^2(q-1)^2 - 1}$$

and $y = \frac{(p-1)(q-1)[PB_{i'} + qT_{j'} - G] - (PB_i + qT_j - G)}{(p-1)^2(q-1)^2 - 1}$

The analysis of data will be performed as usual replacing the missing observations by the estimated values of x and y . However, 2 is to be subtracted from total d.f. and hence, from d.f. of error.

Due to missing observations orthogonality of effects is lost. The sum of squares due to treatment is no longer orthogonal to block sum of squares. Thus adjusted sum of squares (eliminating the effect of block) due to treatment is to be calculated, where

$$SS \text{ (Treatment) adjusted} = SS \text{ (Total) of original data} - SS \text{ (Error) using } x \text{ and } y - SS \text{ (Block) of original data}$$

$$= SS \text{ (Block) + } SS \text{ (Treatment) - } SS \text{ (Block) of original data.}$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F
Block	$p - 1$	S_1	s_1	$F_1 = \frac{s_1}{s_3}$
Treatment (adjusted)	$q - 1$	S_2	s_2	$F_2 = \frac{s_2}{s_3}$
Error	$(p - 1)(q - 1) - 2$	S_3	s_3	—
Total	$pq - 3$			

The comparison of two treatment means without missing observations is to be done as usual. The comparison of a treatment mean with a mean having one missing observation is done by t -test as proposed before. The comparison of two treatment means both containing missing observations is done by t -test, where

$$t = \frac{\bar{y}_{.j} - \bar{y}_{.j'}}{\sqrt{\frac{s_a^2}{p} \left(\frac{1}{r_j} + \frac{1}{r_{j'}} \right)}}$$

Here r_j is known as effective number of replicate of j th treatment having one missing observation, where r_j value is measured from all blocks and all measured values are added for a single r_j value. Here

$$\begin{aligned} \text{a component of } r_j &= 1, \text{ if observation of } j'\text{th treatment is present in a block,} \\ &= \frac{1}{2}, \text{ if observation of } j'\text{th treatment is absent in a block,} \\ &= 0, \text{ if } j\text{th treatment is missing in a block.} \end{aligned}$$

The value of $r_{j'}$ is also calculated similarly, where $r_{j'}$ is the effective number of replicate of j' th treatment.

We have discussed the analytical procedure of data collected from randomized block design with one or two missing observations. There may be more missing observations of different treatments in different blocks. If the experiment is conducted for one observation of one treatment per block, the model can be formulated as

$$y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl}, \quad i = 1, 2, \dots, p, \quad j = 1, 2, \dots, q;$$

$$\begin{aligned} l = n_{ij} &= 0, \text{ if observation of } j\text{th treatment in } i\text{th block is missing} \\ &= 1, \text{ if observation is not missing.} \end{aligned}$$

Again, the missing observation may be experienced in experiment with several replications of a particular treatment in any block. Then, replications per treatment are unequal. The replications of j th treatment in i th block are n_{ij} .

Due to unequal number of observations per cell, the treatment effect and block effect are not independently estimated. However, adjusted treatment effect and adjusted sum of squares due to treatment can be found out. The analytical procedure will be similar as it is done for two-way classification with unequal number of observations per cell [section 2.4].

Example 3.5 : To identify the best dose of nitrogen for potato cultivation an experiment is conducted using 4 doses of nitrogen as urea. Each dose of nitrogen is replicated 5 times. The design used is randomized block design. The production data of potato (kg/plot) are shown below :

Production of potato (y_{ij} kg/plot)

Block	Dose of nitrogen				Total y_i
	N_1	N_2	N_3	N_4	
B_1	15.6	17.2	14.2	18.2	65.2
B_2	16.0	x	15.6	18.8	$50.4 + x$
B_3	14.2	16.8	16.0	17.2	64.2
B_4	15.0	17.0	15.7	18.0	65.7
B_5	16.0	16.8	16.0	17.5	66.3
Total y_j	76.8	$67.8 + x$	77.5	89.7	$311.8 + x$

It is observed that the production of second treatment in,second block is missing.

(i) Analyse the data, (ii) Compare N_2 with N_4 .

Solution : (i) We have, $p = 5, q = 4, B_2 = 50.4, T_2 = 67.8, G = 311.8$.

$$\therefore x = \frac{PB_2 + qT_2 - G}{(p - 1)(q - 1)} = \frac{5 \times 50.4 + 4 \times 67.8 - 311.8}{(5 - 1)(4 - 1)} = 17.6.$$

	N_1	N_2	N_3	N_4	Total y_i .	Mean \bar{y}_i .
B_1					65.2	16.30
B_2		17.6			68.0	17.00
B_3					64.2	16.05
B_4					65.7	16.42
B_5					66.3	16.57
Total $y_{.j}$	76.8	85.4	77.5	89.7	329.4 = G_1	16.47
Mean $y_{.j}$	15.36	17.08	15.50	17.94		

The analyses from original data are as follows :

$$C.T. = \frac{G^2}{19} = \frac{(311.8)^2}{19} = 5116.80.$$

$$SS \text{ (Total)} = \sum_{i'=1}^4 \sum_{\substack{j'=1 \\ i' \neq i', j' \neq j'}}^5 y_{i'j'}^2 - C.T. = 5175.58 - 5116.80 = 58.78$$

$$SS \text{ (Block)} = \frac{\sum_{i' \neq i}^5 y_{i'}^2}{q_{i'}} - C.T.$$

$$= \frac{(65.2)^2}{4} + \frac{(50.4)^2}{3} + \frac{(64.2)^2}{4} + \frac{(65.7)^2}{4} + \frac{(66.3)^2}{4} - 5116.80 = 1.135.$$

Analysis after estimating missing observation :

$$C.T_1 = \frac{G_1^2}{pq} = \frac{(329.4)^2}{20} = 5425.218.$$

$$SS \text{ (Total)}_1 = \sum \sum y_{ij}^2 - C.T. = 5485.34 - 5425.218 = 60.122.$$

$$SS \text{ (Block)}_1 = \frac{\sum y_{i.}^2}{q} - C.T. = \frac{21708.86}{4} - 5425.218 = 1.997.$$

$$SS \text{ (Treatment)}_1 = \frac{\sum y_{.j}^2}{p} - C.T. = \frac{27243.74}{5} - 5425.218 = 23.53.$$

$$SS \text{ (Error)}_1 = SS \text{ (Total)}_1 - SS \text{ (Block)}_1 - SS \text{ (Treatment)}_1$$

$$= 60.122 - 1.997 - 23.53 = 34.595.$$

$$SS \text{ (Treatment) adjusted} = SS \text{ (Total)} - SS \text{ (Error)}_1 - SS \text{ (Block)}$$

$$= 58.78 - 34.595 - 1.135 = 23.05.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	P-value
Block	4	1.997	0.499	0.16	3.36	> 0.05
Treatment (adjusted)	3	23.05	7.683	2.44	3.59	> 0.05
Error	11	34.595	3.145	—	—	—
Total	18					

No factor effect is found significant since $F_1 = 0.16 < F_{0.05}$ and $F_2 = 2.56 < F_{0.05}$.

(ii) Since treatment effects are not found significant, N_2 and N_4 are considered similar.

Example 3.6 : An experiment is conducted in a nursery to observe the production of rose of one variety in presence of different doses of nitrogen and potash. The production of roses per plant within 15 days are recorded for analysis. The land is prepared giving 5 doses of potash in different plots of $(1 \times 1) \text{ m}^2$. Each dose of potash is given in block of 4 plots. In each block 4 doses of nitrogen are randomly allocated. The plan of the experiment is randomized block design, where nitrogen is used as treatment.

Production of roses (y_{ij}) within 15 days

Levels of potash	Levels of nitrogen as Urea				Total y_i
	N_1	N_2	N_3	N_4	
P_1	12	16	18	19	65
P_2	14	18	17	x	$49 + x$
P_3	14	19	20	23	76
P_4	16	20	y	23	$59 + y$
P_5	15	22	25	24	86
Total y_j	71	95	$80 + y$	$89 + x$	$335 + x + y$

It is observed that the result of N_4 in P_2 and the result of N_3 in P_4 are missing.

(i) Analyse the data and comment.

(ii) Is there any difference in the levels of N_3 and N_4 ?

(iii) Is there any difference in the levels of N_2 and N_3 ?

Solution : (i) We have $B_2 = 49$, $B_4 = 59$, $T_3 = 80$, $T_4 = 89$, $p = 5$, $q = 4$, $G = 335$. The value of x and y are estimated by

$$x = \frac{(p-1)(q-1)[pB_2 + qT_4 - G] - [pB_4 + qT_3 - G]}{(p-1)^2(q-1)^2 - 1}$$

$$= \frac{4 \times 3[5 \times 49 + 4 \times 89 - 335] - [5 \times 59 + 4 \times 80 - 335]}{16 \times 9 - 1} = 20.4.$$

$$y = \frac{(p-1)(q-1)[pB_4 + qT_3 - G] - [pB_2 + qT_4 - G]}{(p-1)^2(q-1)^2 - 1}$$

$$= \frac{4 \times 3[5 \times 59 + 4 \times 80 - 335] - [5 \times 49 + 4 \times 89 - 335]}{16 \times 9 - 1} = 21.6.$$

The observations after estimating missing values are as follows :

Levels of potash	Levels of nitrogen as Urea				Total y_i	Mean \bar{y}_i
	N_1	N_2	N_3	N_4		
P_1					65	16.25
P_2				20.4	69.4	17.35
P_3					76	19.00
P_4			21.6		80.6	20.15
P_5					86	21.50
Total y_j	71	95	101.6	109.4	$377 = G_1$	18.85
Total \bar{y}_j	14.20	19.00	20.32	21.88		

Analysis from original data :

$$C.T. = \frac{G^2}{18} = \frac{(335)^2}{18} = 6234.72.$$

$$SS (\text{Total}) = \sum_{\substack{i \neq i' \neq i''=1 \\ j \neq j' \neq j''=1}}^4 \sum^5 y_{i'j''}^2 - C.T. = 6475 - 6234.72 = 240.28.$$

$$SS (\text{Block}) = \sum_{q_i} y_i^2 - C.T. = 75.20.$$

Analysis after estimating missing observations :

$$C.T._1 = \frac{G_1^2}{pq} = \frac{(377)^2}{20} = 7106.45.$$

$$SS (\text{Total})_1 = \sum \sum y_{ij}^2 - C.T. = 7357.72 - 7106.45 = 251.27.$$

$$SS (\text{Block})_1 = \frac{\sum y_i^2}{q} - C.T. = \frac{28709.72}{4} - 7106.45 = 70.98.$$

$$SS (\text{Treatment})_1 = \frac{\sum y_j^2}{p} - C.T. = \frac{36356.92}{5} - 7106.45 = 164.934.$$

$$SS (\text{Error})_1 = SS (\text{Total})_1 - SS (\text{Block})_1 - SS (\text{Treatment})_1 \\ = 251.27 - 70.98 - 164.934 = 15.356.$$

$$SS (\text{Treatment}) \text{ adjusted}_1 = SS (\text{Total}) - SS (\text{Error})_1 - SS (\text{Block}) \\ = 240.28 - 15.356 - 75.20 = 149.724.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	P-value
Block (Potash)	4	70.79	17.6975	11.52	3.48	0.00
Treatment (Urea) adjusted	3	149.724	49.908	32.50	3.71	0.00
Error	10	15.356	1.5356	—	—	—
Total	17					

Since $F_2 = 32.50 > F_{0.05}[p(F_2) < 0.01]$, the doses of nitrogen are highly significantly different.

(ii) We need to test the significance of $H_0 : \beta_3 = \beta_4$, against $H_A : \beta_3 \neq \beta_4$. Both treatments N_3 and N_4 have missing observations. The test statistic is

$$t = \frac{\bar{y}_3 - \bar{y}_4}{\sqrt{s_3 \left(\frac{1}{r_3} + \frac{1}{r_4} \right)}}$$

The components of r_3 are as follows :

$$\begin{aligned} r_3 &= 1, \text{ since } N_4 \text{ is present in block-1} \\ &= \frac{1}{2}, \text{ since } N_4 \text{ is absent in block-2} \\ &= 1, \text{ since } N_4 \text{ is present in block-3} \\ &= 0, \text{ since } N_3 \text{ is itself absent in block-4} \\ &= 1, \text{ since } N_4 \text{ is present in block-5.} \end{aligned}$$

$$\therefore r_3 = 1 + \frac{1}{2} + 1 + 0 + 1 = 3.5.$$

Similarly, $r_4 = 3.5$.

$$\therefore t = \frac{20.32 - 21.88}{\sqrt{1.5356 \left(\frac{1}{3.5} + \frac{1}{3.5} \right)}} = -1.66.$$

Since $|t| < t_{0.025,10} = 2.228$, N_3 and N_4 are not significantly different.

(iii) We need to test the significance of $H_0 : \beta_2 = \beta_3$, against $H_A : \beta_2 \neq \beta_3$. Since N_3 has one missing observation, the test statistic is

$$t = \frac{\bar{y}_2 - \bar{y}_3}{\sqrt{\frac{s_3}{p} \left[2 + \frac{q}{(p-1)(q-1)} \right]}} = \frac{19.00 - 20.32}{\sqrt{\frac{1.5356}{5} \left[2 + \frac{4}{4 \times 3} \right]}} = -1.56.$$

Since $|t| < t_{0.025,10} = 2.228$, H_0 is accepted. N_2 does not differ from N_3 .

3.8 Latin Square Design (LSD)

To control two directional external sources of variation in experimental plots Latin square design is used as a mode of data collection. For example, let us consider that a company produces 4 varieties of protein and it needs to identify the best variety. The company decides to do one experiment with 4 varieties of protein where proteins are to be fed to different guinea pigs. To do the experiment the company needs guinea pigs of multiple of 4 so that each variety of protein can be fed to several guinea pigs. In practice, it is difficult to select more guinea pigs of same age and same body weight. Even if it is possible to select guinea pigs of same age, their body condition scores or body weights may not be same. These latter two characters are important to study the impact of protein. If the guinea pigs vary in ages and in body weights or body condition scores, two external sources of variation prevail in the experimental units. Therefore, experimental plan is to be formulated so that two external sources of variation are controlled.

Consider that there are 16 guinea pigs of 4 different ages. The body weights of different guinea pigs of each age are also different. Here 4 rows with 4 guinea pigs in each row are to be

formed. Four columns of guinea pigs are also to be formed, where in each column there are 4 guinea pigs each of same body weight or body condition score. The rows are perpendicular to the variation of body condition scores and columns are perpendicular to the variation in ages. Now, four varieties of protein are to be given to guinea pigs of rows and columns in such a way that a variety of protein is given only once to a guinea pig in each row and in each column.

Let P_1, P_2, P_3 and P_4 be the four varieties of protein. The varieties are to be allocated according to the following plan to get a 4×4 Latin square design.

Ages	Body condition scores			
	C_1	C_2	C_3	C_4
R_1	P_1	P_2	P_3	P_4
R_2	P_2	P_1	P_4	P_3
R_3	P_3	P_4	P_2	P_1
R_4	P_4	P_3	P_1	P_2

In the above plan, there are three factors, viz., rows, columns and treatments. Each factor has 4 levels. But the experiment is conducted in 4^2 plots instead of $4 \times 4 \times 4 = 64$ plots. Thus, the design is considered as an incomplete three-way layout.

Let there be k treatments to be allocated in k^2 plots, where the plots are divided into k rows and k columns so that each row or each column contains k plots. The rows are taken perpendicular to one source of variation and columns are taken perpendicular to another source of variation. The treatments are allocated to the plots in such a way that each treatment is allocated once and only once in a row and in a column. The resultant design is known as $k \times k$ Latin square design. This design controls two-directional external sources of variation.

Methods of Construction of Latin Square Design

One of the plan of allocation of treatments is shown above. The treatments may be allocated differently and accordingly the names of the design are different. The construction of Latin square design has been discussed by Fisher (1926), Yates (1933), Fisher and Yates (1934), Wilks (1944), Mann (1949) and Kempthorne (1952). A short description of different types of Latin square designs is presented below :

Standard Square : A Latin square design is said to be standard square if the treatments in first row and first column are arranged in alphabetic order or in numerical order. For example, let there be 4 treatments A, B, C and D . These 4 treatments can be allocated in 4 standard squares as follows :

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

1

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

2

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

3

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

4

For three treatments A, B and C the standard square is

A	B	C
B	C	A
C	A	B

5

The latter square can be arranged in other squares as follows :

<table border="1"><tr><td>A</td><td>C</td><td>B</td></tr><tr><td>B</td><td>A</td><td>C</td></tr><tr><td>C</td><td>B</td><td>A</td></tr></table> 6	A	C	B	B	A	C	C	B	A	<table border="1"><tr><td>B</td><td>C</td><td>A</td></tr><tr><td>C</td><td>A</td><td>B</td></tr><tr><td>A</td><td>B</td><td>C</td></tr></table> 7	B	C	A	C	A	B	A	B	C	<table border="1"><tr><td>B</td><td>A</td><td>C</td></tr><tr><td>C</td><td>B</td><td>A</td></tr><tr><td>A</td><td>C</td><td>B</td></tr></table> 8	B	A	C	C	B	A	A	C	B	<table border="1"><tr><td>C</td><td>B</td><td>A</td></tr><tr><td>A</td><td>C</td><td>B</td></tr><tr><td>B</td><td>A</td><td>C</td></tr></table> 9	C	B	A	A	C	B	B	A	C	<table border="1"><tr><td>C</td><td>A</td><td>B</td></tr><tr><td>A</td><td>B</td><td>C</td></tr><tr><td>B</td><td>C</td><td>A</td></tr></table> 10	C	A	B	A	B	C	B	C	A	<table border="1"><tr><td>A</td><td>B</td><td>C</td></tr><tr><td>C</td><td>A</td><td>B</td></tr><tr><td>B</td><td>C</td><td>A</td></tr></table> 11	A	B	C	C	A	B	B	C	A
A	C	B																																																									
B	A	C																																																									
C	B	A																																																									
B	C	A																																																									
C	A	B																																																									
A	B	C																																																									
B	A	C																																																									
C	B	A																																																									
A	C	B																																																									
C	B	A																																																									
A	C	B																																																									
B	A	C																																																									
C	A	B																																																									
A	B	C																																																									
B	C	A																																																									
A	B	C																																																									
C	A	B																																																									
B	C	A																																																									
<table border="1"><tr><td>A</td><td>C</td><td>B</td></tr><tr><td>C</td><td>B</td><td>A</td></tr><tr><td>B</td><td>A</td><td>C</td></tr></table> 12	A	C	B	C	B	A	B	A	C	<table border="1"><tr><td>B</td><td>C</td><td>A</td></tr><tr><td>A</td><td>B</td><td>C</td></tr><tr><td>C</td><td>A</td><td>B</td></tr></table> 13	B	C	A	A	B	C	C	A	B	<table border="1"><tr><td>B</td><td>A</td><td>C</td></tr><tr><td>A</td><td>C</td><td>B</td></tr><tr><td>C</td><td>B</td><td>A</td></tr></table> 14	B	A	C	A	C	B	C	B	A	<table border="1"><tr><td>C</td><td>B</td><td>A</td></tr><tr><td>B</td><td>A</td><td>C</td></tr><tr><td>A</td><td>C</td><td>B</td></tr></table> 15	C	B	A	B	A	C	A	C	B	<table border="1"><tr><td>C</td><td>A</td><td>B</td></tr><tr><td>B</td><td>C</td><td>A</td></tr><tr><td>A</td><td>B</td><td>C</td></tr></table> 16	C	A	B	B	C	A	A	B	C										
A	C	B																																																									
C	B	A																																																									
B	A	C																																																									
B	C	A																																																									
A	B	C																																																									
C	A	B																																																									
B	A	C																																																									
A	C	B																																																									
C	B	A																																																									
C	B	A																																																									
B	A	C																																																									
A	C	B																																																									
C	A	B																																																									
B	C	A																																																									
A	B	C																																																									

It is observed that with 3 treatments there are $3!(3-1)! = 12$ arrangements of squares. One of the 12 squares is standard one and the remaining 11 squares are non-standard. There are 4 standard squares of four treatments A, B, C and D . These 4 letters of each standard square can be arranged in rows and columns in $4!(4-1)!$ ways. Thus, there are in total $144 \times 4 = 576$ squares of treatments. However, only 4 of them are standard squares. It is observed that in $k \times k$ Latin square design each standard square can be arranged in $k!(k-1)!$ squares.

Conjugate squares : If the arrangement of treatments in rows of one square is same as the arrangement of treatments in columns of another square of two standard squares, then the squares are called conjugate squares. For example, the square-1 and square-2 of four treatments A, B, C and D given above are conjugate squares.

Self-conjugate square : If the arrangement of treatments in rows and columns are similar in a square, it is called self-conjugate square. For example, all the standard squares given above are self-conjugate squares. Moreover, square numbers 7, 10, 12, 14 and 15 are also self-conjugate squares.

Adjugate-set : If the rows, columns and treatments are combined with each other, there will be 6 combinations. These six squares of combinations are called adjugate-set.

Self-Adjugate-set : If rows, columns and treatments are combined and the combinations give the same set, then the set is called self-adjugate-set. The square 1, 2, 3 and 4 are self-adjugate-set.

For any practical purpose, the plan of treatment can be selected from the table of Latin square arrangement as suggested by Fisher and Yates (1934). In practice, the Latin square design is used for maximum 10 treatments.

Analysis of Data : Let there be k treatments which are allocated in rows and columns according to a given $k \times k$ Latin square plan. Let y_{ijl} be the result of l th treatment in j th column of i th rows ($i = j = l = 1, 2, \dots, k$). The linear model for y_{ijl} observation is :

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l + e_{ijl}$$

where μ = general mean, α_i = effect of i th row, β_j = effect of j th column, γ_l = effect and l th treatment and e_{ijl} = random error.

Assumption : $e_{ijl} \sim NID(0, \sigma^2)$.

The estimated error sum of squares related to the model is :

$$\phi = \sum \sum \sum (y_{ijl} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_l)^2.$$

The values of $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$ and $\hat{\gamma}_l$ are to be found out solving the following equations :

$$\frac{\partial \phi}{\partial \hat{\mu}} = 0, \quad \frac{\partial \phi}{\partial \hat{\alpha}_i} = 0, \quad \frac{\partial \phi}{\partial \hat{\beta}_j} = 0 \quad \text{and} \quad \frac{\partial \phi}{\partial \hat{\gamma}_l} = 0.$$

The equations are also written as

$$\begin{aligned} y_{...} &= k^2 \hat{\mu} + k \sum \hat{\alpha}_i + k \sum \hat{\beta}_j + k \sum \hat{\gamma}_l \\ y_{i..} &= k \hat{\mu} + k \hat{\alpha}_i + \sum \hat{\beta}_j + \sum \hat{\gamma}_l \\ y_{.j.} &= k \hat{\mu} + \sum \hat{\alpha}_i + k \hat{\beta}_j + \sum \hat{\gamma}_l \\ y_{.l.} &= k \hat{\mu} + \sum \hat{\alpha}_i + \sum \hat{\beta}_j + k \hat{\gamma}_l. \end{aligned}$$

There are $(3k + 1)$ normal equations, but 3 of them are dependent and $(3k - 2)$ are independent. To get the unique solution of these equations we need to put 3 restrictions. The restrictions are $\sum \hat{\alpha}_i = \sum \hat{\beta}_j = \sum \hat{\gamma}_l = 0$. Under the restrictions, the estimates are

$$\hat{\mu} = \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad \hat{\gamma}_l = \bar{y}_{.l.} - \bar{y}_{...}$$

These elements are independent. It can be shown as follows :

$$\begin{aligned} \text{Cov}(\hat{\alpha}_i, \hat{\gamma}_l) &= \text{Cov}(\bar{y}_{i..} - \bar{y}_{...}, \bar{y}_{.l.} - \bar{y}_{...}) \\ &= \text{Cov}(\bar{y}_{i..}, \bar{y}_{.l.}) - \text{Cov}(\bar{y}_{i..}, \bar{y}_{...}) - \text{Cov}(\bar{y}_{...}, \bar{y}_{.l.}) + V(\bar{y}_{...}) \\ &= \frac{\sigma^2}{k^2} - \frac{k\sigma^2}{k \cdot k^2} - \frac{k\sigma^2}{k^2 \cdot k} + \frac{\sigma^2}{k^2} = 0. \end{aligned}$$

Similarly, all other covariances can be shown as zero.

The total variation of observations can be partitioned as follows :

$$\begin{aligned} SS \text{ (Total)} &= \sum_i^k \sum_j^k \sum_l^k (y_{ijl} - \bar{y}_{...})^2 = \sum \sum \sum [(y_{i..} - \bar{y}_{...}) + (y_{.j.} - \bar{y}_{...}) \\ &\quad + (y_{.l.} - \bar{y}_{...}) + (y_{ijl} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{.l.} + 2\bar{y}_{...})]^2 \\ &= k \sum (\bar{y}_{i..} - \bar{y}_{...})^2 + k \sum (\bar{y}_{.j.} - \bar{y}_{...})^2 + k \sum (\bar{y}_{.l.} - \bar{y}_{...})^2 \\ &\quad + \sum \sum \sum (y_{ijl} - \bar{y}_{i..} - y_{.j.} - \bar{y}_{.l.} + 2\bar{y}_{...})^2 + \text{cross-product terms} \end{aligned}$$

All cross-product terms are zero. Thus, we have

$$\begin{aligned} \sum_i^k \sum_j^k \sum_l^k (y_{ijl} - \bar{y}_{...})^2 &= S_1 + S_2 + S_3 + S_4 = SS(\hat{\alpha}_i) + SS(\hat{\beta}_j) + SS(\hat{\gamma}_l) + SS \text{ (Error)} \\ &= SS \text{ (Row)} + SS \text{ (Column)} + SS \text{ (Treatment)} + SS \text{ (Error)}. \end{aligned}$$

Under assumption all sum of squares are independently distributed as $\chi^2\sigma^2$. The d.f. of these χ^2 are $(k-1)$, $(k-1)$, $(k-1)$ and $(k-1)(k-2)$, respectively. The d.f. can be shown as follows :

$$Ek \sum_l (y_{..l} - \bar{y}_{...})^2 = kE \sum (\bar{e}_{..l} - \bar{e}_{...} + \gamma_l)^2, \text{ under restrictions } \sum \alpha_i = 0 = \sum \beta_j = \sum \gamma_l.$$

Therefore, $kE \sum_l (y_{..l} - \bar{y}_{...})^2 = k \sum E(\bar{e}_{..l}^2) + k \sum E(\bar{e}_{...}^2) - 2kE \sum \bar{e}_{...} \bar{e}_{..l}$

$$+ k \sum E\gamma_l^2 + 2kE \sum \gamma_l(\bar{e}_{..l} - \bar{e}_{...})$$

$$= \frac{kk\sigma^2}{k} + \frac{kk\sigma^2}{k^2} - \frac{2k^2\sigma^2}{k^2} + k \sum \gamma_l^2 = (k-1)\sigma^2 + k \sum \gamma_l^2$$

$$= (k-1)\sigma^2, \text{ if } \gamma_l = 0$$

$$\frac{kE \sum_l (y_{..l} - \bar{y}_{...})^2}{\sigma^2} = k-1, \text{ if } \gamma_l = 0.$$

Therefore, SS (Treatment) is distributed as central $\chi^2\sigma^2$ with $(k-1)$ d.f. under the assumption $\gamma_l = 0$. Now, we need to test the significance of the null hypothesis $H_0 : \gamma_l = 0$, against $H_A : \gamma_l \neq 0$.

The non-null distribution of SS (Treatment) is non-central χ^2 with non-centrality parameter.

$$\lambda_3 = \frac{k}{2\sigma^2} \sum \gamma_l^2.$$

In a similar way, the d.f. of other sum of squares can be shown. The sum of squares due to error is distributed as $\chi^2\sigma^2$ with $(k-1)(k-2)$ d.f. without any restriction, since

$$E \sum \sum \sum (y_{ijl} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..l} + 2\bar{y}_{...})^2$$

$$= \sum \sum \sum E(e_{ijl} - \bar{e}_{i..} - \bar{e}_{.j.} - \bar{e}_{..l} + 2\bar{e}_{...})^2 = (k-1)(k-2)\sigma^2.$$

The test statistic to test the significance of $H_0 : \gamma_l = 0$ is

$$F_3 = \frac{S_3/(k-1)}{S_4/(k-1)(k-2)}.$$

If $F_3 \geq F_{\alpha; k-1, (k-1)(k-2)}$, H_0 is rejected.

The test statistics for $H_0 : \alpha_i = 0$ and $H_0 : \beta_j = 0$ are found out similarly and conclusion is also made in a similar way.

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{\text{d.f.}}$	F	$E(MS)$	P -value
Row	$k-1$	S_1	s_1	$F_1 = \frac{s_1}{s_4}$	$\sigma^2 = \frac{k}{k-1} \sum \alpha_i^2$	
Column	$k-1$	S_2	s_2	$F_2 = \frac{s_2}{s_4}$	$\sigma^2 = \frac{k}{k-1} \sum \beta_j^2$	
Treatment	$k-1$	S_3	s_3	$F_3 = \frac{s_3}{s_4}$	$\sigma^2 = \frac{k}{k-1} \sum \gamma_l^2$	
Error	$(k-1)(k-2)$	S_4	s_4	—	σ^2	
Total	k^2-1					

Here P_i -value = $\int_{F_i}^{\infty} f(F)dF$. If $p(F_i) \leq 0.05$, H_0 is rejected.

The rejection of $H_0 = \gamma_l = 0$ leads us to compare the treatments in pairs. The comparison is done by Duncan's multiple range test, where the test statistic is :

$$D_l = d_{\alpha, l, f} \sqrt{\frac{s_d^2}{k}}, \quad l = 2, 3, \dots, k; \quad f = (k - 1)(k - 2).$$

Any particular pair, say, l th and l' th treatment is compared by t -test, where

$$t = \frac{\bar{y}_{..l} - \bar{y}_{..l'}}{\sqrt{\frac{2s_d^2}{k}}}$$

The null hypothesis to be tested by the above t -test is $H_0 : \gamma_l = \gamma_{l'}$, against $H_A : \gamma_l \neq \gamma_{l'}$, $l \neq l' = 1, 2, \dots, k$.

The conclusion will be made in a similar way as it is done in other t -tests.

Example 3.7 : An experiment is conducted to study the productivity of cotton varieties in presence of doses of nitrogen and irrigation levels. Five varieties of cotton are cultivated using 5 doses of nitrogen as urea. Five doses of nitrogen are applied in 5 plots of a row, where 5 rows are used to use 5 levels of irrigation. The design used is LSD. The plot size is 12 m × 8 m. The production of cotton in plots and the plan of treatments in rows and columns are shown below :

Rows	Columns					Total $y_{i..}$	Mean $\bar{y}_{i..}$
	N_1	N_2	N_3	N_4	N_5		
I_1	$C_1 - 10.2$	$C_2 - 11.0$	$C_3 - 9.0$	$C_4 - 12.2$	$C_5 - 8.8$	51.2	10.24
I_2	$C_2 - 11.2$	$C_1 - 11.6$	$C_4 - 12.5$	$C_5 - 9.2$	$C_3 - 10.2$	54.7	10.94
I_3	$C_3 - 9.5$	$C_5 - 10.2$	$C_1 - 12.6$	$C_2 - 14.0$	$C_4 - 12.0$	58.3	11.66
I_4	$C_4 - 13.2$	$C_3 - 12.5$	$C_5 - 10.5$	$C_1 - 12.6$	$C_2 - 15.0$	63.8	12.76
I_5	$C_5 - 9.2$	$C_4 - 14.2$	$C_2 - 11.2$	$C_3 - 11.6$	$C_1 - 13.0$	59.2	11.84
Total $y_{.j}$	53.3	59.5	55.8	59.6	59.0	287.2	—
Mean $\bar{y}_{.j}$	10.66	11.90	11.16	11.92	11.80	—	11.488

Total of treatments, $y_{.l} : 60.0, 62.4, 52.8, 64.1, 47.9$.

Mean of treatment, $\bar{y}_{.l} : 12.00, 12.48, 10.56, 12.82, 9.58$.

- (i) Analyse the data and group the cotton varieties.
- (ii) Which level of irrigation would you recommend for better production?
- (iii) Which dose of nitrogen would you recommend for better production?

Solution : (i) $k = 5$, $G = 287.2$, C.T. = $\frac{G^2}{k^2} = \frac{(287.2)^2}{25} = 3299.3536$.

$$SS \text{ (Total)} = \sum \sum \sum y_{ijl}^2 - \text{C.T.} = 3370.08 - 3299.3536 = 70.7264.$$

$$SS \text{ (Rows)} = \frac{\sum y_{i..}^2}{k} - \text{C.T.} = \frac{16587.5}{5} - 3299.3536 = 18.1464.$$

$$SS \text{ (Columns)} = \frac{\sum y_{.j}^2}{k} - \text{C.T.} = \frac{16527.94}{5} - 3299.3536 = 6.2344.$$

$$SS (\text{Treatments}) = \frac{\sum y_{.l}^2}{k} - \text{C.T.} = \frac{16684.82}{5} - 3299.3536 = 37.6104.$$

$$SS (\text{Error}) = SS (\text{Total}) - SS (\text{Rows}) - SS (\text{Columns}) - SS (\text{Treatments}) \\ = 70.7264 - 18.1464 - 6.2344 - 37.6104 = 8.7352.$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{\text{d.f.}}$	F	F _{0.05}	P-value
Rows (irrigation)	4	18.1464	4.5366	6.23	3.26	0.00
Columns (nitrogen)	4	6.2344	1.5586	2.14	3.26	> 0.05
Treatments	4	37.6104	9.4026	12.92	3.26	0.00
Error	12	8.7352	0.7279	—	—	—
Total	24					

The cotton varieties are significantly different, since $p(F_3) < 0.00$ [$F_1 = 12.92 > F_{0.05}$].

The cotton varieties can be grouped by Duncan's multiple range test, where the test statistic is

$$D_l = d_{.05, l, f} \sqrt{\frac{s_4}{k}}, \quad l = 2, 3, 4, 5; \quad f = 12$$

$$D_2 = 3.08 \sqrt{\frac{0.7279}{5}} = 1.18, \quad D_3 = 3.23 \sqrt{\frac{0.7279}{5}} = 0.81,$$

$$D_4 = 3.33 \sqrt{\frac{0.7279}{5}} = 1.27, \quad D_5 = 3.36 \sqrt{\frac{0.7279}{5}} = 1.28.$$

The means in ascending order are

$$\bar{C}_5 = 9.58, \quad \bar{C}_3 = 10.56, \quad \bar{C}_1 = 12.00, \quad \bar{C}_2 = 12.48, \quad \bar{C}_4 = 12.82.$$

$$\bar{C}_4 - \bar{C}_5 = 12.82 - 9.58 = 3.24 > D_5, \quad \therefore \text{Means are significantly different.}$$

$$\bar{C}_4 - \bar{C}_3 = 12.82 - 10.56 = 2.26 > D_4, \quad \therefore C_3 \text{ and } C_4 \text{ are different.}$$

$$\bar{C}_2 - \bar{C}_5 = 12.48 - 9.58 = 2.90 > D_4, \quad \therefore C_2 \text{ and } C_5 \text{ are different.}$$

$$\bar{C}_1 - \bar{C}_5 = 12.00 - 9.58 = 2.42 > D_3, \quad \therefore C_1 \text{ and } C_5 \text{ are different.}$$

$$\bar{C}_2 - \bar{C}_3 = 12.48 - 10.56 = 1.92 > D_3, \quad \therefore C_2 \text{ and } C_3 \text{ are different.}$$

$$\bar{C}_4 - \bar{C}_1 = 12.82 - 12.00 = 0.82 < D_3, \quad \therefore C_1 \text{ and } C_4 \text{ are similar.}$$

$$\bar{C}_3 - \bar{C}_5 = 10.56 - 9.58 = 0.98 < D_2, \quad \therefore C_3 \text{ and } C_5 \text{ are similar.}$$

$$\bar{C}_1 - \bar{C}_3 = 12.00 - 10.56 = 1.44 > D_2, \quad \therefore C_1 \text{ and } C_3 \text{ are similar.}$$

The underlined means do not differ. $\bar{C}_5, \bar{C}_3, \bar{C}_1, \bar{C}_2, \bar{C}_4$

(ii) Since the levels of irrigation are significantly different by F -test, these levels can be grouped by Duncan's multiple range test, where

$$D_i = d_{.05, i, f} \sqrt{\frac{s_4}{k}}, \quad i = 2, 3, 4, 5; \quad f = 12.$$

The D_i values are equal as calculated above. The means of levels of irrigation are in ascending order as follows :

$$\bar{T}_1 = 10.24, \quad \bar{T}_2 = 10.94 \quad \bar{T}_3 = 11.66 \quad \bar{T}_5 = 11.84 \quad \bar{T}_4 = 12.76$$

$$\bar{T}_4 - \bar{T}_1 = 12.76 - 10.24 = 2.52 > D_5, \quad \therefore \text{levels of irrigation are significantly different.}$$

$$\bar{T}_5 - \bar{T}_1 = 11.84 - 10.24 = 1.60 > D_4, \quad \therefore I_1 \text{ and } I_5 \text{ are different.}$$

$$\bar{T}_4 - \bar{T}_2 = 12.76 - 10.94 = 1.82 > D_4, \quad \therefore I_2 \text{ and } I_4 \text{ are different.}$$

$$\bar{T}_3 - \bar{T}_1 = 11.66 - 10.24 = 1.42 > D_3, \quad \therefore I_1 \text{ and } I_3 \text{ are different.}$$

$$\bar{T}_5 - \bar{T}_2 = 11.84 - 10.94 = 0.90 < D_3, \quad \therefore I_2 \text{ and } I_5 \text{ are similar.}$$

$$\bar{T}_4 - \bar{T}_3 = 12.76 - 11.66 = 1.10 < D_3, \quad \therefore I_3 \text{ and } I_4 \text{ are similar.}$$

$$\bar{T}_2 - \bar{T}_1 = 10.94 - 10.24 = 0.70 < D_2, \quad \therefore I_1 \text{ and } I_2 \text{ are similar.}$$

$$\bar{T}_3 - \bar{T}_2 = 11.66 - 10.94 = 0.72 < D_2, \quad \therefore I_2 \text{ and } I_3 \text{ are similar.}$$

The underlined means do not differ.

$$\underline{\underline{I_1, I_2, I_3, I_5, I_4}}$$

It is observed that the levels of irrigation I_3, I_4 and I_5 are similar but the average production of these three levels are higher. Among these 3 levels I_4 is the best.

(iii) The doses of nitrogen are similar in productivity as is observed by F -test. However, N_2, N_3, N_4 and N_5 are slightly better than N_1 .

3.9 Analysis of Latin Square Design with Missing Observations

Let us consider the analysis of data with one missing observation. Let the observation of l th treatment of j th column in i th row be missing. Let us denote this observation by x . To analyse the data x is to be estimated in such a way that the estimated error sum of squares is minimum.

The total of l th treatment is written as $T_l + x$ and other treatment totals are $y_{.l'}$ ($l' \neq l = 1, 2, \dots, k$). The total of i th row is $R_i + x$ and the total of j th column is $C_j + x$. The totals of other rows and columns are $y_{i'..}$ and $y_{.j'}$. ($i' \neq i = 1, 2, \dots, k; j' \neq j = 1, 2, \dots, k$), respectively. Let $(G + x)$ be the grand total of observations. Now, the estimated error sum of squares with one missing observation is written as

$$\phi = \sum \sum \sum y_{i'j'l'}^2 + x^2 - \sum \frac{y_{i'..}^2}{k} - \frac{(R_i + x)^2}{k} - \sum \frac{y_{.j'}^2}{k} - \frac{(C_j + x)^2}{k} - \sum \frac{y_{.l'}^2}{k} - \frac{(T_l + x)^2}{k} + \frac{(G + x)^2}{k^2}$$

Now $\frac{\partial \phi}{\partial x} = 0$ gives

$$2x - \frac{2(R_i + x)}{k} - \frac{2(C_j + x)}{k} - \frac{2(T_l + x)}{k} + \frac{2(G + x)}{k^2} = 0.$$

$$\therefore x = \frac{k(R_i + C_j + T_l) - 2G}{(k - 1)(k - 2)}$$

The analysis is performed in usual way replacing missing observation by the estimated value of x . The exception in the analysis is that 1 is to be subtracted from total d.f. and hence, from error d.f. for one missing value.

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F
Rows	$k - 1$	S_1	s_1	$F_1 = \frac{s_1}{s_4}$
Columns	$k - 1$	S_2	s_2	$F_2 = \frac{s_2}{s_4}$
Treatments	$k - 1$	S_3	s_3	$F_3 = \frac{s_3}{s_4}$
Error	$(k - 1)(k - 2) - 1$	S_4	s_4	—
Total	$k^2 - 2$			

If the null hypothesis $H_0 : \gamma_l = 0$ is rejected, it is needed to compare the treatments in pairs. The comparison is done as usual by Duncan's multiple range test. But the hypothesis of the type $H_0 : \gamma_l = \gamma_{l'}$, against $H_A : \gamma_l \neq \gamma_{l'}$ ($l \neq l' = 1, 2, \dots, k$), where γ_l is the treatment effect having a missing observation. The test statistic is

$$t = \frac{\bar{y}_{..l} - \bar{y}_{..l'}}{\sqrt{v(\bar{y}_{..l} - \bar{y}_{..l'})}}$$

Here $v(\bar{y}_{..l} - \bar{y}_{..l'})$ is the estimate of $V(\bar{y}_{..l} - \bar{y}_{..l'})$.

Here $V(\bar{y}_{..l} - \bar{y}_{..l'}) = V(\bar{y}_{..l}) + V(\bar{y}_{..l'}) - 2\text{Cov}(\bar{y}_{..l}, \bar{y}_{..l'})$.

We have $V(\bar{y}_{..l'}) = \frac{\sigma^2}{k}$

$$V(\bar{y}_{..l}) = \frac{1}{k^2} V \left[T_l + \frac{k(R_i + C_j + T_l) - 2G}{(k-1)(k-2)} \right]$$

$$V(\bar{y}_{..l}) = \frac{1}{k^2} \left[V(T_l) + \frac{1}{(k-1)^2(k-2)^2} \{k^2V(R_i) + k^2V(C_j) + k^2V(T_l) + 4V(G)\} \right. \\ \left. + 2k^2\text{Cov}(T_l, R_i) + 2k^2\text{Cov}(R_i, C_j) + 2k^2\text{Cov}(C_j, T_l) - 4k\text{Cov}(G, R_i) \right. \\ \left. - 4k\text{Cov}(G, C_j) - 4k\text{Cov}(G, T_l) + \frac{1}{(k-1)(k-2)} \{k\text{Cov}(T_l, R_i) \right. \\ \left. + k\text{Cov}(T_l, C_j) + k\text{Cov}(T_l, T_l) - 2\text{Cov}(T_l, G)\} \right]$$

$$= \frac{1}{k^2} \left[(k-1) + \frac{1}{(k-1)^2(k-2)^2} \{k^2(k-1) + k^2(k-1) + k^2(k-1) + 4(k^2-1) \right. \\ \left. + 2k^2 \times 0 + 2k^2 \times 0 + 2k^2 \times 0 - 4k(k-1) - 4k(k-1) - 4k(k-1) \right. \\ \left. + \frac{2}{(k-1)(k-2)} \{k \cdot 0 + k \cdot 0 + k(k-1) - 2(k-1)\} \right] \sigma^2$$

$$= \frac{\sigma^2}{k^2} \left[(k-1) + \frac{1}{(k-1)^2(k-2)^2} + \{3k^2(k-1) + 4(k^2-1) - 12k(k-1)\} \right. \\ \left. + \frac{2(k-1)(k-2)}{(k-1)(k-2)} \right]$$

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

$$\begin{aligned} \therefore V(\bar{y}_{..l} - \bar{y}_{..l'}) &= [V(\bar{y}_{..l}) + V(\bar{y}_{..l'})], \quad \because \text{Cov}(\bar{y}_{..l}, \bar{y}_{..l'}) = 0 \\ &= \sigma^2 \left[\frac{2}{k} + \frac{1}{(k-1)(k-2)} \right]. \end{aligned}$$

Here σ^2 is estimated by s_4 . Therefore,

$$v(\bar{y}_{..l} - \bar{y}_{..l'}) = s_4 \left[\frac{2}{k} + \frac{1}{(k-1)(k-2)} \right].$$

$$\therefore t = \frac{\bar{y}_{..l} - \bar{y}_{..l'}}{\sqrt{s_4 \left[\frac{2}{k} + \frac{1}{(k-1)(k-2)} \right]}}$$

This t has $(k-1)(k-2) - 1$ d.f. The conclusion will be drawn as usual.

The treatment so estimated is not orthogonal to the effects of row and column due to missing observation. To test the significance of treatment contrast it is necessary to adjust the sum of squares due to treatment. The adjusted treatment sum of squares is found out as follows :

$$SS (\text{Treatment})_{\text{adjusted}} = SS (\text{Total}) - SS (\text{Error})_1 - SS (\text{Row}) - SS (\text{Column})$$

Here $G_1 = \sum_{i' \neq i} \sum_{j' \neq j} \sum_{l' \neq l} y_{i'j'l'} + \hat{x}$, \hat{x} is the estimated value of x

$$\text{C.T.}_1 = \frac{G_1^2}{k^2}, \quad SS (\text{Total})_1 = \sum_i \sum_j \sum_l y_{ijl}^2 - \text{C.T.}_1, \quad SS (\text{Row})_1 = \sum \frac{y_{i.}^2}{k} - \text{C.T.}_1,$$

$$SS (\text{Column})_1 = \sum \frac{y_{.j}^2}{k} - \text{C.T.}_1, \quad SS (\text{Treatment})_1 = \sum \frac{y_{..l}^2}{k} - \text{C.T.}_1$$

$$y_{i.} : y_{i.1}, y_{i.2}, \dots, R_i + \hat{x}, \dots, y_{i.k}$$

$$y_{.j} : y_{.1}, y_{.2}, \dots, C_j + \hat{x}, \dots, y_{.k}$$

$$y_{..l} : y_{..1}, y_{..2}, \dots, T_l + \hat{x}, \dots, y_{..k}$$

$$SS (\text{Error})_1 = SS (\text{Total})_1 - SS (\text{Row})_1 - SS (\text{Column})_1 - SS (\text{Treatment})_1$$

The above analysis is done after replacing the estimated value of $x(\hat{x})$.

Example 3.8 : The following data are related to number of faded signals sent from 4 different stations (namely, A, B, C, D) in 4 days of a week. The data are recorded for 4 weeks following a 4×4 Latin square design. The plan and faded signals of stations recorded in days and weeks are shown below :

Observations of faded signals of stations recorded in days of weeks (y_{ijl})

Days	Weeks			
	W_1	W_2	W_3	W_4
D_1	A-10	B-8	C-7	D-18
D_2	B-8	A-12	D-16	C-9
D_3	C-6	D-16	B-x	A-10
D_4	D-15	C-7	A-11	B-10

- (i) Analyse the data and comment.
 (ii) Is there any difference between station (A) and station (B)?
 (iii) Compare station B with others assuming that the number of faded signals of B are known earlier.
 (iv) Group the stations.

Solution : We have $k = 4$, $G = 163$. The observation of third row and third column is missing. This missing observation is for second treatment. Let R_3 , C_3 and T_2 be the total of third row, third column and second treatment, respectively for original data, where $R_3 = 32$, $C_3 = 34$, $T_2 = 26$. Therefore, estimate of x is

$$\hat{x} = \frac{k(R_3 + C_3 + T_2) - 2G}{(k-1)(k-2)} = \frac{4(32 + 34 + 26) - 2 \times 163}{(4-1)(4-2)} = 7.$$

Also, we have (after replacing the missing value)

$$y_{i..} : 43, 45, 39, 43, G_1 = 170$$

$$y_{.j.} : 39, 43, 41, 47$$

$$y_{.l} : 43, 33, 29, 65$$

$$\bar{y}_{.l} : 10.75, 8.25, 7.25, 16.25$$

Analysis from original data

$$C.T. = \frac{G^2}{k^2 - 1} = \frac{(163)^2}{15} = 1771.27, \quad C.T._1 = \frac{G_1^2}{k^2} = \frac{(170)^2}{16} = 1806.25$$

$$SS \text{ (Total)} = \sum_{i'} \sum_{j'} \sum_{l'} y_{i'j'l'}^2 - C.T. = 1969 - 1771.27 = 197.73$$

$i' \neq j' \neq l' \neq i'$

$$SS \text{ (Row)} = \sum \frac{y_{i..}^2}{k_i} - C.T. = \frac{43^2}{4} + \frac{45^2}{4} + \frac{32^2}{3} + \frac{43^2}{4} - 1771.27 = 0.81$$

$$SS \text{ (Column)} = \sum \frac{y_{.j.}^2}{k_j} - C.T. = \frac{39^2}{4} + \frac{43^2}{4} + \frac{34^2}{3} + \frac{47^2}{4} - 1771.27 = 8.81$$

$$SS \text{ (Treatment)} = \sum \frac{y_{.l}^2}{k_l} - C.T. = \frac{43^2}{4} + \frac{26^2}{3} + \frac{29^2}{4} + \frac{65^2}{4} - 1771.27 = 182.81$$

Analysis after estimating missing observation

$$SS \text{ (Total)}_1 = \sum \sum \sum y_{ijl}^2 - C.T._1 = 2018 - 1806.25 = 211.75$$

$$SS \text{ (Row)}_1 = \sum \frac{y_{i..}}{k} - C.T._1 = \frac{7244}{4} - 1806.25 = 4.75$$

$$SS \text{ (Column)}_1 = \sum \frac{y_{.j.}}{k} - C.T._1 = \frac{7260}{4} - 1806.25 = 8.75$$

$$SS \text{ (Treatment)}_1 = \sum \frac{y_{.l}}{k} - C.T._1 = \frac{8004}{4} - 1806.25 = 194.75$$

$$SS \text{ (Error)}_1 = SS \text{ (Total)}_1 - SS \text{ (Row)}_1 - SS \text{ (Column)}_1 - SS \text{ (Treatment)}_1 \\ = 211.75 - 4.75 - 8.75 - 194.75 = 3.5$$

$$SS \text{ (Treatment)adjusted} = SS \text{ (Total)} - SS \text{ (Error)}_1 - SS \text{ (Row)} - SS \text{ (Column)}$$

$$= 197.73 - 3.5 - 0.81 - 8.81 = 184.61$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	P value
Row	3	4.75	1.583	2.26	5.41	> 0.05
Column	3	8.75	2.917	4.17	5.41	> 0.05
Treatment (adjusted)	3	184.61	61.537	87.91	5.41	< 0.00
Error	5	3.50	0.7	—	—	—
Total	14					

Since $F_3 = 87.91 > F_{0.05}$ [$P - \text{value} < 0.00$], the effects of stations are highly significantly different.

(ii) We need to test the significance of $H_0 : \gamma_1 = \gamma_2$, against $H_A : \gamma_1 \neq \gamma_2$. Since second station has one missing observation, the test statistic is

$$t = \frac{\bar{y}_{..1} - \bar{y}_{..2}}{\sqrt{s_4 \left[\frac{2}{k} + \frac{1}{(k-1)(k-2)} \right]}} = \frac{10.75 - 8.25}{\sqrt{0.7 \left[\frac{2}{4} + \frac{1}{(4-1)(4-2)} \right]}} = 3.66.$$

Since $|t| \geq t_{0.025,5} = 2.571$, H_0 is rejected. Station A and station B are significantly different in respect of faded signals.

(iii) We need to compare station B with others. Here station B can be considered as control treatment. It can be compared with others by Dunnett's test, where the test statistic is

$$D = d_{0.05,k, \text{ error d.f.}} \sqrt{s_4 \left[\frac{2}{k} + \frac{1}{(k-1)(k-2)} \right]}$$

as B has one missing value.

$d_{0.05,k, \text{ error d.f.}} = 5\%$ value from Dunnett's table for k means including control treatment with error d.f. (= 5).

$$\therefore D = 3.29 \sqrt{0.7 \left[\frac{2}{4} + \frac{1}{(4-1)(4-2)} \right]} = 2.247.$$

$$|\bar{y}_{..2} - \bar{y}_{..1}| = 2.5 > D, \text{ A and B are different stations.}$$

$$|\bar{y}_{..2} - \bar{y}_{..3}| = 1.00 < D, \text{ B and C are similar stations.}$$

$$|\bar{y}_{..2} - \bar{y}_{..4}| = 95 > D, \text{ B and D are different stations.}$$

(iv) For grouping we can use Duncan's multiple range test, where the test statistic is

$$D_l = d_{0.05,l,f} \sqrt{\frac{s_4}{k}}; \quad l = 2, 3, 4$$

$d_{0.05,l,f} = 5\%$ tabulated value from Duncan's table for range of l means with $f = \text{error d.f.}$

$$D_2 = 3.65 \sqrt{\frac{0.7}{4}} = 1.53, \quad D_3 = 3.74 \sqrt{\frac{0.7}{4}} = 1.56, \quad D_4 = 3.79 \sqrt{\frac{0.7}{4}} = 1.59$$

The means of stations in ascending order are $\bar{C} = 7.25$, $\bar{B} = 8.25$, $\bar{A} = 10.75$, $\bar{D} = 16.25$
 $\bar{D} - \bar{C} = 9 > D_4$; the means are significantly different.

$\bar{A} - \bar{C} = 3.5 > D_3$; A and C are different.

$\bar{D} - \bar{B} = 8.00 > D_3$; B and D are different.

$\bar{D} - \bar{A} = 5.50 > D_2$; A and D are different.

$\bar{A} - \bar{B} = 2.50 > D_2$; A and D are different.

$\bar{B} - \bar{C} = 1.00 > D_2$; B and C are similar.

The underlined stations are in one group.

C, B, A, D

Analysis with Two Missing Observations : Let the observation of l th treatment in j th column of i th row be missing. Let us denote this observation by x . Again, consider that the observation of l' th treatment of j' th column in i' th row is missing. Let us denote this observation by y . Here $i \neq i' = 1, 2, \dots, k$; $j \neq j' = 1, 2, \dots, k$; $l \neq l' = 1, 2, \dots, k$.

To analyse the data of this $k \times k$ Latin square design the missing observations x and y are to be estimated in such a way that the estimated error sum of squares is minimum. The error sum of squares including the missing values x and y is written as

$$\begin{aligned} \phi = \sum \sum \sum y_{i''j''l''}^2 + x^2 + y^2 - \sum \frac{y_{i''..}^2}{k} - \frac{(R_i + x)^2}{k} - \frac{(R_{i'} + y)^2}{k} \\ - \sum \frac{y_{.j''.}^2}{k} - \frac{(C_j + x)^2}{k} - \frac{(C_{j'} + y)^2}{k} - \sum \frac{y_{.l''}^2}{k} \\ - \frac{(T_l + x)^2}{k} - \frac{(T_{l'} + y)^2}{k} + \frac{2(G + x + y)^2}{k^2}. \end{aligned}$$

Here $i \neq i' \neq i'' = 1, 2, \dots, k$; $j \neq j' \neq j'' = 1, 2, \dots, k$; $l \neq l' \neq l'' = 1, 2, \dots, k$, R_i = total of i th row except x , $R_{i'}$ = total of i' th row except y , C_j = total of j th column except x , $C_{j'}$ = total of j' th column except y , T_l = total of l th treatment except x , $T_{l'}$ = total of l' th treatment except y , G = grand total without x and y .

The value of x and y are to be found out solving the equations $\frac{\partial \phi}{\partial x} = 0$ and $\frac{\partial \phi}{\partial y} = 0$.

$$\text{We have } 2x - \frac{2(R_i + x)}{k} - \frac{2(C_j + x)}{k} - \frac{2(T_l + x)}{k} + \frac{4(G + x + y)}{k^2} = 0$$

$$2y - \frac{2(R_{i'} + y)}{k} - \frac{2(C_{j'} + y)}{k} - \frac{2(T_{l'} + y)}{k} + \frac{4(G + x + y)}{k^2} = 0.$$

Solving these equations, we have

$$x = \frac{k(R_i + C_j + T_l) - 2G}{\frac{k(k-3)(k^2-3k+4)}{(k-1)(k-2)}} - 2 \left\{ \frac{k(R_{i'} + C_{j'} + T_{l'}) - 2G}{k(k-3)(k^2-3k+4)} \right\}$$

$$\text{and } y = \frac{k(R_{i'} + C_{j'} + T_{l'}) - 2G}{\frac{k(k-3)(k^2-3k+4)}{(k-1)(k-2)}} - 2 \left\{ \frac{k(R_i + C_j + T_l) - 2G}{k(k-3)(k^2-3k+4)} \right\}.$$

The missing observations are replaced by the values of x and y calculated according to the above formulae. Then the analysis of the data is performed as usual except that 2 is subtracted from total d.f. and hence, from d.f. of error.

In practice, the values of x and y are also found out by iterative procedure. For this, either value of x or y is calculated taking average of the rows in which they are missing. For example, let x be the average of i th row. Then, we have one missing value y . This y is to be estimated by

$$\hat{y} = \frac{k(R_{i'} + C_{j'} + T_{l'}) - 2G_1}{(k-1)(k-2)}, \quad \text{where } G_1 = G + \bar{R}_i.$$

Replacing \hat{y} in the missing place, x is estimated by

$$\hat{x} = \frac{k(R_i + C_j + T_l) - 2G_2}{(k-1)(k-2)}, \quad \text{where } G_2 = G + \hat{y}.$$

If this \hat{x} is equal to the first value of $x(\bar{R}_i)$, the estimation of missing observations is finished. If not, replacing \hat{x} in missing place y is estimated second time, where

$$\hat{y}_1 = \frac{k(R_{i'} + C_{j'} + T_{l'}) - 2G_3}{(k-1)(k-2)}, \quad \text{where } G_3 = G + \hat{x}.$$

Replacing \hat{y}_1 in the missing place, where x is estimated second time

$$\hat{x}_1 = \frac{k(R_i + C_j + T_l) - 2G_4}{(k-1)(k-2)}, \quad \text{where } G_4 = G + \hat{y}_1.$$

The process of estimation is continued until two consecutive estimates of x are equal or approximately equal.

In this analysis also sum of squares of treatment is adjusted as it is suggested in the previous analysis.

The comparison of two treatments both containing missing observations is done by t -test, where

$$t = \frac{\bar{y}_{..l} - \bar{y}_{..l'}}{\sqrt{s_4 \left(\frac{1}{r_l} + \frac{1}{r_{l'}} \right)}}, \quad \text{where } s_4 = MS (\text{error}).$$

Here r_l and $r_{l'}$ are the effective number of replicates of l th and l' th treatment, respectively. The effective number of replicate is calculated from each row. The component of r_l and $r_{l'}$ is obtained as follows :

Component of $r_l = 1$, if l' th treatment is present in a row and in a column.

$$= \frac{2}{3}, \text{ if } l'\text{th treatment is present either in a row or in a column.}$$

$$= \frac{1}{3}, \text{ if } l'\text{th treatment is missing in a row and in a column.}$$

$$= 0, \text{ if } l\text{th treatment is itself missing in a row.}$$

The value of r_l is the sum of different components calculated from all rows. Similar is the case with $r_{l'}$.

The comparison of two treatments one having missing observation is performed in a similar way as it is proposed before. Pairwise comparison of other treatments are done by usual t -test or by Duncan's multiple range test.

In latin square design if all observations of a row or a column are missing, the design is no longer a Latin square design. The design is called Youden Square design and it is an incomplete block design. The analysis has been discussed by Yates (1936), Yates and Hale (1939) and Das and Giri (1979).

Example 3.9 : An experiment is conducted to study the productivity of a rose variety using 6 doses of nitrogen as urea. The objective is to study the impacts of nitrogen on production of rose. The plot is prepared using 6 doses of potash. During experimentation 6 levels of irrigation are also applied. The number of roses produced per plant after one month of the start of experiment is recorded for analysis. The experiment is conducted through LSD.

Row Levels of Potash	Levels of Irrigation						Total, $y_{i..}$
	I_1	I_2	I_3	I_4	I_5	I_6	
P_1	$F - 21$	$E - 17$	$D - 17$	$C - 20$	$B - 17$	$A - 16$	108
P_2	$E - 18$	$C - 19$	$A - x$	$D - 19$	$F - 22$	$B - 19$	$97 + x$
P_3	$B - 16$	$A - 16$	$F - 22$	$E - 19$	$D - 19$	$C - 21$	113
P_4	$A - 15$	$B - 18$	$E - 20$	$F - y$	$C - 20$	$D - 19$	$92 + y$
P_5	$D - 18$	$F - 24$	$C - 20$	$B - 18$	$A - 20$	$E - 20$	120
P_6	$C - 20$	$D - 18$	$B - 17$	$A - 19$	$E - 19$	$F - 25$	118
Total $y_{.j}$	108	112	$96 + x$	$95 + y$	117	120	$648 + x + y$

Here $A = 30$ kg/ha, $B = 60$ kg/ha, $C = 90$ kg/ha, $D = 120$ kg/ha, $E = 150$ kg/ha, $F = 180$ kg/ha, $P_1 = 0$ kg/ha, $P_2 = 20$ kg/ha, $P_3 = 40$ kg/ha, $P_4 = 60$ kg/ha, $P_5 = 80$ kg/ha, $P_6 = 100$ kg/ha; $I_1 =$ irrigation for 15 minutes in the morning, $I_2 =$ irrigation for 25 minutes in the morning, $I_3 =$ irrigation for 35 minutes in the morning, I_4, I_5 and I_6 are similar to I_1, I_2 and I_3 but in the afternoon.

(i) Analyse the data. (ii) Is there any difference between A and F ? (iii) Is there any difference between E and F ?

Given $y_{.1} : 86 + x, 105, 120, 110, 113, 114 + y$.

Solution : (i) We have $k = 6, R_2 = 97, R_4 = 92, C_3 = 96, C_4 = 95, T_1 = 86, T_6 = 114, G = 648$.

$$x = \frac{k(R_2 + C_3 + T_1) - 2G}{\frac{k(k-3)(k^2-3k+4)}{(k-1)(k-2)}} - 2 \left[\frac{k(R_4 + C_4 + T_6) - 2G}{k(k-3)(k^2-3k+4)} \right]$$

$$= \frac{6(97 + 96 + 86) - 2 \times 648}{\frac{6(6-3)(36-18+4)}{(6-1)(6-2)}} - 2 \left[\frac{6(92 + 95 + 114) - 2 \times 648}{6(6-3)(36-18+4)} \right]$$

$$= 19.09 - 2.57 = 16.52.$$

$$y = \frac{k(R_4 + C_4 + T_6) - 2G}{\frac{k(k-3)(k^2-3k+4)}{(k-1)(k-2)}} - 2 \left[\frac{k(R_2 + C_3 + T_1) - 2G}{(k-1)(k-2)} \right]$$

$$= \frac{6(92 + 95 + 114) - 2 \times 648}{\frac{6(6-3)(36-18+4)}{(6-1)(6-2)}} - 2 \left[\frac{6(97 + 96 + 86) - 2 \times 648}{6(6-3)(36-18+4)} \right]$$

$$= 25.76 - 1.91 = 23.85.$$

Alternative Method to Estimate x and y

The first estimate of $\hat{x} = \bar{R}_2 = \frac{97}{5} = 19.4, G_1 = 648 + 19.4 = 667.4$.

$$\hat{y} = \frac{k(R_4 + C_4 + T_6) - 2G_1}{(k-1)(k-2)} = \frac{6(92 + 95 + 114) - 2 \times 667.4}{(6-1)(6-2)} = 23.56.$$

$$G_2 = G + \hat{y} = 648 + 23.56 = 671.56.$$

$$\hat{x}_1 = \frac{k(R_2 + C_3 + T_1) - 2G_2}{(k-1)(k-2)} = \frac{6(97 + 96 + 86) - 2 \times 671.56}{(6-1)(6-2)} = 16.544.$$

$$G_3 = G + \hat{x}_1 = 648 + 16.544 = 664.544.$$

$$\hat{y}_1 = \frac{k(R_4 + C_4 + T_6) - 2G_3}{(k-1)(k-2)} = \frac{6(92 + 95 + 114) - 2 \times 664.544}{(6-1)(6-2)} = 23.84.$$

$$G_4 = G + \hat{y}_1 = 648 + 23.84 = 671.84.$$

$$\hat{x}_2 = \frac{k(R_2 + C_3 + T_1) - 2G_4}{(k-1)(k-2)} = \frac{6(97 + 96 + 86) - 2 \times 671.84}{(6-1)(6-2)} = 16.52.$$

$$\hat{y}_2 = \frac{6(92 + 95 + 114) - 2 \times 664.52}{(6-1)(6-2)} = 23.85, \text{ where } G_5 = G + \hat{x}_2 = 664.52.$$

$$G_6 = G + \hat{y}_2 = 648 + 23.85 = 671.85.$$

$$\hat{x}_3 = \frac{6(97 + 96 + 86) - 2 \times 671.85}{(6-1)(6-2)} = 16.52.$$

Since $\hat{x}_3 = \hat{x}_2 = 16.52$, it is taken as an estimate of x and hence, estimate of y is 23.85.

Analysis from original data :

$$C.T. = \frac{G^2}{k^2 - 2} = \frac{(648)^2}{34} = 12350.1176.$$

$$SS \text{ (Total)} = \sum_{i \neq i' \neq i''} \sum_{j \neq j' \neq j''} \sum_{l \neq l' \neq l''} y_{ijl}^2 - C.T. = 12508 - 12350.1176 = 157.8824.$$

$$\begin{aligned} SS \text{ (Row)} &= \frac{\sum y_{i.}^2}{k_i} - C.T. \\ &= \frac{(108)^2}{6} + \frac{(97)^2}{5} + \frac{(113)^2}{6} + \frac{(92)^2}{5} + \frac{(120)^2}{6} + \frac{(118)^2}{6} - 12350.1176 \\ &= 17.3157. \end{aligned}$$

$$\begin{aligned} SS \text{ (Column)} &= \frac{\sum y_{.j}^2}{k_j} - C.T. \\ &= \frac{(108)^2}{6} + \frac{(112)^2}{6} + \frac{(96)^2}{5} + \frac{(95)^2}{5} + \frac{(117)^2}{6} + \frac{(120)^2}{6} - 12350.1176 \\ &= 14.2491; \end{aligned}$$

Analysis with Estimated Values :

$$G_1 = G + x + y = 648 + 16.52 + 23.85 = 688.37.$$

$$C.T._1 = \frac{G_1^2}{k^2} = \frac{(688.37)^2}{36} = 13162.5905.$$

$$SS \text{ (Total)} = \sum \sum \sum y_{ijl}^2 - C.T._1 = 13349.7329 - 13162.5905 = 187.1424.$$

$$y_{i.} : 108, 113.52, 113, 115.85, 120, 118; \quad y_{.j} : 108, 112, 112.52, 118.85, 117, 120$$

$$y_{.l} : 102.52, 105, 120, 110, 113, 137.85; \quad \bar{y}_{.l} : 17.09, 17.50, 20.00, 18.33, 18.83, 22.97$$

$$SS (\text{Row})_1 = \frac{\sum y_{i..}^2}{k} - \text{C.T.}_1 = \frac{79065.0129}{6} - 13162.5905 = 14.9119.$$

$$SS (\text{Column})_1 = \frac{\sum y_{.j.}^2}{k} - \text{C.T.}_1 = \frac{79083.0729}{6} - 13162.5905 = 17.9216.$$

$$SS (\text{Treatment})_1 = \frac{\sum y_{.i.}^2}{k} - \text{C.T.}_1 = \frac{79806.9729}{6} - 13162.5905 = 138.5716.$$

$$SS (\text{Error})_1 = SS (\text{Total})_1 - SS (\text{Row})_1 - SS (\text{Column})_1 - SS (\text{Treatment})_1 \\ = 187.1424 - 14.9119 - 17.9216 - 138.5716 = 15.7373.$$

$$SS (\text{Treatment}) \text{ adjusted} = SS (\text{Total}) - SS (\text{Error})_1 - SS (\text{Row}) - SS (\text{Column}) \\ = 157.8827 - 15.7373 - 17.3157 - 14.2491 = 110.5806.$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{\text{d.f.}}$	F	F _{0.05}	P-value
Rows	5	14.9119	2.98238	3.41	2.77	< 0.05
Column	5	17.9216	3.58432	4.10	2.77	< 0.05
Treatment (adjusted)	5	110.5806	22.11612	25.29	2.77	0.00
Error	18	15.7373	0.8743	—	—	—
Total	33					

Since $F_3 = 25.29 > F_{0.05}$ [$P\text{-value} < 0.01$], the treatment effects are highly significant. The doses of nitrogen have significant differential impacts on production of rose.

(ii) We need to test the significance of the hypothesis $H_0 : \gamma_1 = \gamma_6$ against $H_A : \gamma_1 \neq \gamma_6$.

Both the treatments have missing observations. The test statistic is

$$t = \frac{\bar{y}_{.1} - \bar{y}_{.6}}{\sqrt{s_4 \left(\frac{1}{r_1} + \frac{1}{r_6} \right)}}, \text{ where } s_4 = MS (\text{error})_1 = 0.8743.$$

$$r_1 = 1 + 0 + 1 + \frac{1}{3} + 1 + 1 = \frac{13}{3}, \quad r_6 = 1 + \frac{2}{3} + \frac{2}{3} + 0 + 1 + 1 = \frac{13}{3}.$$

$$\therefore t = \frac{17.09 - 22.97}{\sqrt{0.8743 \left(\frac{3}{13} + \frac{3}{13} \right)}} = -9.26.$$

$|t| > t_{0.025, 18} = 2.101$, H_0 is rejected. A and F are significantly different.

(iii) The hypothesis is $H_0 : \gamma_5 = \gamma_6$, against $H_A : \gamma_5 \neq \gamma_6$.

Here E has no missing observation but F has one missing observation. The test statistic is

$$t = \frac{\bar{y}_{.5} - \bar{y}_{.6}}{\sqrt{s_4 \left[\frac{2}{k} + \frac{1}{(k-1)(k-2)} \right]}} = \frac{18.83 - 22.97}{\sqrt{0.8743 \left[\frac{2}{6} + \frac{1}{(6-1)(6-2)} \right]}} = -7.15.$$

$|t| > t_{0.025, 18} = 2.101$, H_0 is rejected. E and F are significantly different.

3.10 Efficiency of Latin Square Design

It is already mentioned that LSD is used to control two-directional external sources of variation, where plots are grouped in rows and columns. Let us now investigate the efficiency of the experiment due to these two-directional groupings. The efficiency of LSD is studied compared to RBD and CRD, where in the former case, plots are grouped perpendicular to one-directional external source of variation and in the latter case no grouping of plots is used.

The analysis of variance table of a $k \times k$ Latin square design is shown below :

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$
Row	$k - 1$	S_1	R
Column	$k - 1$	S_2	C
Treatment	$k - 1$	S_3	T
Error	$(k - 1)(k - 2)$	S_4	E
Total	$k^2 - 1$		

Let us consider that the treatment effect is insignificant and variance due to treatment is, on an average, equal to the error variance. Then the analysis of variance table takes the following shape :

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$
Row	$k - 1$	$(k - 1)R$	R
Column	$k - 1$	$(k - 1)C$	C
Error	$(k - 1)^2$	$(k - 1)^2 E$	E
Total	$k^2 - 1$		

In such a stage if the column classification (or row classification) is not made, the analysis of variance table takes the shape as follows :

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$
Row	$k - 1$	$(k - 1)R$	R
Error	$k(k - 1)$	$(k - 1)C + (k - 1)^2 E$	$\frac{C + (k - 1)E}{k}$
Total	$k^2 - 1$		

This analysis of variance table is equivalent to the analysis of variance table of a randomized block design, where k treatments are used and each one is replicated k times. The treatment effects are assumed insignificant. Therefore, the relative efficiency of LSD compared to RBD is

$$\frac{C + (k - 1)E}{kE}$$

or, $\frac{R + (k - 1)E}{kE}$ [if column classification is not made]

If both row and column classifications are not made, the sum of squares of error is written as $(k - 1)R + (k - 1)C + (k - 1)^2E$ and the mean square due to error is

$$\frac{R + C + (k - 1)E}{k + 1}$$

This error sum of squares is equivalent to the error sum of squares obtained from a completely randomized design, where k treatments are randomly allocated to k^2 plots. Therefore, the efficiency of the LSD compared to CRD is

$$\frac{R + C + (k - 1)E}{(k + 1)E}$$

Cochran (1938, 1940) has shown that the efficiency of LSD compared to CRD is about 222 per cent and it is about 137 per cent compared to RBD. This is observed in analysing the data recorded in Rothamsted Experimental Station. This information indicates that the similar efficient information of a treatment is obtained from LSD, RBD and CRD, if the treatment is replicated 4 to 5 times in LSD, 6 times in RBD and 10 times in CRD. Similar result is also observed by Ma and Harrington (1948). Yates (1935) mentioned that, to obtain similar efficient information of a treatment from RBD, it needs 2 and half times more plot compared to the plots used in a LSD.

3.11 Advantages, Disadvantages and Uses of Latin Square Design

Advantages

- (i) Since Latin square design controls two directional external sources of variation, it controls error more than that is done by randomized block design and completely randomized design. Hence, the estimate of error variance is expected to be smaller.
- (ii) The analysis of data obtained from this design is simple and easier, even it is easier in presence of missing observation. However, the analysis is slightly complicated than that of the randomized block design.
- (iii) The analysis is also done if the observations of a row or a column are lost or missing.
- (iv) Latin square design is an incomplete three-way layout but the plots needed are k^2 instead of k^3 , when the experiment is conducted with k treatments.
- (v) The design can also be planned for field experiment, where the shape of the design is square one. If the variation in the plots is along with line, the plots are taken along that line.

Disadvantages

- (i) Since number of rows, columns and treatments are same, the design is not suitable for large number of treatments. Latin square design is seldom used with more than 10 treatments.
- (ii) If number of treatments is less than 5, the error d.f. becomes smaller and hence, the estimate of error variance is not efficient unless the design is replicated.

- (iii) 2×2 Latin square design is not used unless it is replicated.
- (iv) If two observations of a 3×3 Latin square design are missing, the analysis is not done since the estimate of error variance is not available.
- (v) It is assumed that there is no interaction between any two factors of the experiment.

Uses : The design is used in laboratory experiment, in industrial experiment, in greenhouse experiment, even it is used in field experiment. The design is used in any situation where two-directional external sources of variation are needed to be controlled by design.

Example 3.10 : Find the efficiency of LSD compared to RBD and CRD using the data of Example 3.7.

Solution : Given $k = 5$, $R = 4.5366$, $C = 1.5586$, $T = 9.4026$, $E = 0.7279$. The efficiency of LSD compared to CRD is

$$\frac{C + (k - 1)E}{kE} = \frac{1.5586 + (5 - 1)0.7279}{5 \times 0.7279} = 122.8\%$$

when row classification is not made. If column classification is not made, the efficiency is

$$\frac{R + (k - 1)E}{kE} = \frac{4.5366 + (5 - 1)0.7279}{5 \times 0.7279} = 204.6\%$$

The efficiency of LSD compared to CRD is given by

$$\frac{R + C + (k - 1)E}{(k + 1)E} = \frac{4.5366 + 1.5586 + (5 - 1)(0.7279)}{(5 + 1)0.7279} = 206.2\%$$

3.12 Analysis of Latin Square Design with Several (Equal) Observations Per Cell

Let us consider that a $(k \times k)$ latin square design is conducted with k treatments, where r observations are recorded for l th treatment in j th column of i th row ($i = j = l = 1, 2, \dots, k$). Let y_{ijlm} be the m th observation of l th treatment in j th column of i th row ($m = 1, 2, \dots, r$). The model assumed for the data of such an experiment is

$$y_{ijlm} = \mu + \alpha_i + \beta_j + \gamma_l + \delta_{ijl} + e_{ijlm}, \tag{1}$$

where μ = general mean, α_i = effect of i th row, β_j = effect of j th column, γ_l = effect of l th treatment, δ_{ijl} = interaction of l th treatment in j th column of i th row, e_{ijlm} = random error. Here δ_{ijl} is the effect of r observations corresponding to any column, row and treatment.

Assumption : (i) $e_{ijlm} \sim NID(0, \sigma^2)$.

Further assumption is made regarding observations of row, column, treatment and cell. There are three different possibilities of assumption. These are (ii) the row, column treatment and cell are assumed to be separate populations, (iii) the observations of row, column, treatment and cell are assumed to be random sample observations from respective population, and (iv) except treatment the observations are assumed to be random sample observations from corresponding populations.

Let us consider the analysis first with assumptions (i) and (ii). For this analysis, let us put the restriction

$$\sum \alpha_i = \sum \beta_j = \sum \gamma_l = \sum_i \delta_{ijl} = \sum_j \delta_{ijl} = \sum_l \delta_{ijl} = 0.$$

The estimated error sum of squares is $\phi = \sum \sum \sum [y_{ijlm} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_l - \hat{\delta}_{ijl}]^2$.

The normal equation to obtain $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, $\hat{\gamma}_l$ and $\hat{\delta}_{ijl}$ are :

$$\frac{\partial \phi}{\partial \hat{\mu}} = 0, \quad \frac{\partial \phi}{\partial \hat{\alpha}_i} = 0, \quad \frac{\partial \phi}{\partial \hat{\beta}_j} = 0, \quad \frac{\partial \phi}{\partial \hat{\gamma}_l} = 0, \quad \frac{\partial \phi}{\partial \hat{\delta}_{ijl}} = 0.$$

On simplification, we have

$$\begin{aligned} y_{....} &= rk^2 \hat{\mu} + rk \sum \hat{\alpha}_i + rk \sum \hat{\beta}_j + rk \sum \hat{\gamma}_l + r \sum \sum \sum \hat{\delta}_{ijl} \\ y_{i...} &= rk \hat{\mu} + rk \hat{\alpha}_i + r \sum \hat{\beta}_j + r \sum \hat{\gamma}_l + r \sum_j \sum_l \hat{\delta}_{ijl} \\ y_{.j..} &= rk \hat{\mu} + r \sum \hat{\alpha}_i + rk \hat{\beta}_j + r \sum \hat{\gamma}_l + r \sum_i \sum_l \hat{\delta}_{ijl} \\ y_{..l.} &= rk \hat{\mu} + r \sum \hat{\alpha}_i + r \sum \hat{\beta}_j + rk \hat{\gamma}_l + r \sum_i \sum_j \hat{\delta}_{ijl} \\ y_{ijl.} &= r \hat{\mu} + r \hat{\alpha}_i + r \hat{\beta}_j + r \hat{\gamma}_l + r \hat{\delta}_{ijl}. \end{aligned}$$

There are $(k^2 + 3k + 1)$ normal equations. Among these equations last k^2 are independent and the remaining $(3k + 1)$ are dependent on these last k^2 equations. Hence, to get unique solution of these normal equations, we need to put $(3k + 1)$ restrictions.

The restrictions are

$$\sum \hat{\alpha}_i = \sum \hat{\beta}_j = \sum \hat{\gamma}_l = \sum_i \sum_j \hat{\delta}_{ijl} = \sum_i \sum_l \hat{\delta}_{ijl} = \sum_j \sum_l \hat{\delta}_{ijl} = \sum_i \sum_j \sum_l \hat{\delta}_{ijl} = 0.$$

Under these restrictions the estimates are :

$$\begin{aligned} \hat{\alpha}_i &= \bar{y}_{i...} - \bar{y}_{....}, \quad \hat{\beta}_j = \bar{y}_{.j..} - \bar{y}_{....}, \quad \hat{\gamma}_l = \bar{y}_{..l.} - \bar{y}_{....} \\ \hat{\delta}_{ijl} &= \bar{y}_{ijl.} - \bar{y}_{ij.} - \bar{y}_{i.l.} - \bar{y}_{.jl.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..l.} - 2\bar{y}_{....} \end{aligned}$$

These estimates are independent since covariances of estimates in pairs are zero. For example,

$$\begin{aligned} \text{Cov}(\hat{\alpha}_i, \hat{\gamma}_l) &= \text{Cov}(\bar{y}_{i...} - \bar{y}_{....}, \bar{y}_{..l.} - \bar{y}_{....}) \\ &= \text{Cov}(\bar{y}_{i...}, \bar{y}_{..l.}) - \text{Cov}(\bar{y}_{i...}, \bar{y}_{....}) - \text{Cov}(\bar{y}_{..l.}, \bar{y}_{....}) + v(\bar{y}_{....}) \\ &= \frac{r\sigma^2}{rkrk} - \frac{rk\sigma^2}{rkrk^2} - \frac{rk\sigma^2}{rk^2rk} + \frac{\sigma^2}{rk^2} = 0. \end{aligned}$$

Similarly, all other covariances can be shown as zero. The total sum of squares of the observations can be partitioned as follows :

$$\begin{aligned} \sum_i \sum_j \sum_l \sum_m (y_{ijlm} - \bar{y}_{....})^2 &= \sum_i \sum_j \sum_l \sum_m [(\bar{y}_{i...} - \bar{y}_{....}) + (\bar{y}_{.j..} - \bar{y}_{....}) \\ &\quad + (\bar{y}_{..l.} - \bar{y}_{....}) + (\bar{y}_{ijl.} - \bar{y}_{ij.} - \bar{y}_{i.l.} - \bar{y}_{.jl.} + \bar{y}_{i...} + \bar{y}_{.j..} \\ &\quad + \bar{y}_{..l.} - 2\bar{y}_{....}) + (y_{ijlm} - \bar{y}_{ijl.})]^2. \end{aligned}$$

Squaring the right side and on simplification, we get

$$\sum \sum \sum \sum (y_{ijlm} - \bar{y}_{....})^2 = rk \sum (\bar{y}_{i...} - \bar{y}_{....})^2 + rk \sum (\bar{y}_{.j..} - \bar{y}_{....})^2$$

$$\begin{aligned}
 & +rk \sum (\bar{y}_{..l} - \bar{y}_{....})^2 + r \sum \sum \sum (\bar{y}_{ijl} - \bar{y}_{ij..} - \bar{y}_{i.l} \\
 & \quad - \bar{y}_{.jl} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..l} - 2\bar{y}_{....})^2 \\
 & \quad + \sum \sum \sum \sum (y_{ijlm} - \bar{y}_{ijl})^2 \\
 = & SS \text{ (Row)} + SS \text{ (Column)} + SS \text{ (Treatment)} \\
 & + SS \text{ (Observations)} + SS \text{ (Error)} \\
 = & S_1 + S_2 + S_3 + S_4 + S_5.
 \end{aligned}$$

By assumption all the sum of squares are distributed as chi-square. The d.f. of these sum of squares can be found out as follows :

$$\begin{aligned}
 & Er \sum \sum \sum (\bar{y}_{ijl} - \bar{y}_{ij..} - \bar{y}_{i.l} - \bar{y}_{.jl} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..l} - 2\bar{y}_{....})^2 \\
 & = rE \sum \sum \sum (\delta_{ijl} + \bar{e}_{ijl} - \bar{e}_{i.l} - \bar{e}_{.jl} - \bar{e}_{ij..} + \bar{e}_{j..} + \bar{e}_{..l} + \bar{e}_{i...} - 2\bar{e}_{....})^2 \\
 & = rE \sum \sum \sum \sum \delta_{ijl}^2 + rE \sum \sum \sum (\bar{e}_{ijl} - \bar{e}_{ij..} - \bar{e}_{i.l} - \bar{e}_{.jl} + \bar{e}_{i...} + \bar{e}_{.j..} + \bar{e}_{..l} - 2\bar{e}_{....})^2 \\
 & = r \sum \sum \sum \sum \delta_{ijl}^2 + (k-1)(k-2)\sigma^2 \\
 & = (k-1)(k-2)\sigma^2, \text{ if } \delta_{ijl} = 0.
 \end{aligned}$$

$$\frac{Er \sum \sum \sum (y_{ijl} - \bar{y}_{ij..} - \bar{y}_{i.l} - \bar{y}_{.jl} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..l} - 2\bar{y}_{....})^2}{\sigma^2} = (k-1)(k-2).$$

Thus, $\frac{SS \text{ (observations)}}{\sigma^2}$ is distributed as χ^2 with $(k-1)(k-2)$ d.f. under the restriction $\delta_{ijl} = 0$.

Similarly, the d.f. of other sum of squares are found out. We have

$$\begin{aligned}
 E \sum \sum \sum \sum (y_{ijlm} - \bar{y}_{ijl})^2 & = E \sum \sum \sum \sum (e_{ijlm} - \bar{e}_{ijl})^2 \\
 & = \sum \sum \sum \sum e_{ijlm}^2 + r \sum_i \sum_j \sum_l E(\bar{e}_{ijl}^2) \\
 & \quad - 2r \sum_i \sum_j \sum_l E\bar{e}_{ijl}^2 \\
 & = rk^2\sigma^2 + \frac{rk^2\sigma^2}{r} - \frac{2r\sigma^2k^2}{r} = k^2\sigma^2(r-1).
 \end{aligned}$$

$$\therefore \frac{E \sum \sum \sum \sum (y_{ijlm} - \bar{y}_{ijl})^2}{\sigma^2} = k^2(r-1).$$

Therefore, $\frac{SS \text{ (error)}}{\sigma^2}$ is distributed as χ^2 with $k^2(r-1)$ d.f. under no constraint and hence, to test the significance of $H_0 : \delta_{ijl} = 0$, against $H_A : \delta_{ijl0} \neq 0$, the test statistic is

$$F_4 = \frac{S_4/(k-1)(k-2)}{S_5/k^2(r-1)}.$$

Under H_0 this F_4 has central variance ratio distribution having $(k-1)(k-2)$ and $k^2(r-1)$ d.f. The non-null distribution of F_4 is non-central F with non-centrality parameter

$$\lambda_4 = \frac{r}{2\sigma^2} \sum \sum \sum \delta_{ijl}^2.$$

The null hypothesis is rejected, if $F_4 \geq F_{0.05; (k-1)(k-2), k^2(r-1)}$. The other F -statistics are calculated as usual and the conclusion is drawn similarly.

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$
Row	$k - 1$	S_1	s_1	$F_1 = \frac{s_1}{s_5}$	$\sigma^2 + \frac{rk}{k-1} \sum \alpha_i^2$
Column	$k - 1$	S_2	s_2	$F_2 = \frac{s_2}{s_5}$	$\sigma^2 + \frac{rk}{k-1} \sum \beta_j^2$
Treatment	$k - 1$	S_3	s_3	$F_3 = \frac{s_3}{s_5}$	$\sigma^2 + \frac{rk}{k-1} \sum \gamma_i^2$
Cell	$(k - 1)(k - 2)$	S_4	s_4	$F_4 = \frac{s_4}{s_5}$	$\sigma^2 + \frac{r}{(k-1)(k-2)} \sum \sum \sum \delta_{ijl}^2$
Error	$k^2(r - 1)$	S_5	s_5		σ^2
Total	$rk^2 - 1$				

The comparison of treatment effects in pairs is done as usual by Duncan's multiple range test, where the test statistic is

$$D_l = d_{\alpha, l, f} \sqrt{\frac{s_5}{rk}}, \quad l = 2, 3, \dots, f = k^2(r - 1).$$

Example 3.11 : To study the impact of nitrogen as urea on production of marigold an experiment is conducted using 4 doses of nitrogen [00(A), 30(B), 60(C) and 90(D) kg/ha]. The land is prepared using 30, 60, 90 and 120 kg/ha potash in rows and similar doses of phosphorus in columns. Each doses of nitrogen is applied in 2 plots. The amount of flowers in kg/plot are recorded for analysis.

Doses of Potash	Doses of Phosphorus				Total, $y_{i..}$
	P_1	P_2	P_3	P_4	
K_1	A-2.5, 3.2	B-4.0, 4.2	C-4.5, 5.0	D-5.0, 5.2	33.6
K_2	B-3.5, 4.0	C-5.2, 5.5	D-5.2, 5.5	A-4.0, 3.0	35.9
K_3	C-4.0, 4.0	D-6.0, 6.5	A-4.6, 4.4	B-4.0, 4.8	38.3
K_4	D-4.5, 5.3	A-4.0, 4.2	B-4.0, 4.0	C-4.6, 5.2	35.8
Total $y_{.j.}$	31.0	39.6	37.2	35.8	143.6

$$y_{..l} : 29.9, 32.5, 38.0, 43.2; \quad \bar{y}_{..l} : 3.14, 4.06, 4.75, 5.40.$$

Analyse the data and group the levels of nitrogen.

Solution : We have $r = 2$, $k = 4$, $G = 143.6$, C.T. = $\frac{G^2}{rk^2} = \frac{(143.6)^2}{2 \times 16} = 644.405$.

$$SS (\text{Total}) = \sum \sum \sum \sum y_{ijlm}^2 - \text{C.T.} = 667.24 - 644.405 = 22.835.$$

The observations (y_{ijl}) of treatments in rows and columns are as follows :

Rows	Columns			
	P_1	P_2	P_3	P_4
K_1	5.7	8.2	9.5	10.2
K_2	7.5	10.7	10.7	7.0
K_3	8.0	12.5	9.0	8.8
K_4	9.8	8.2	8.0	9.8

$$SS (\text{Rows}) = \frac{\sum y_{i..}^2}{rk} - \text{C.T.} = \frac{5166.3}{2 \times 4} - 644.405 = 1.3825.$$

$$SS (\text{Columns}) = \frac{\sum y_{.j.}^2}{rk} - \text{C.T.} = \frac{5194.64}{2 \times 4} - 644.405 = 4.925.$$

$$SS (\text{Treatments}) = \frac{\sum y_{.l.}^2}{rk} - \text{C.T.} = \frac{5260.5}{rk} - 644.405 = 13.1575.$$

$$SS (\text{Observations}) = \frac{\sum \sum \sum y_{ijl}^2}{r} - \text{C.T.} - SS (\text{Row}) - SS (\text{Column}) - SS (\text{Treatment})$$

$$= \frac{1330.26}{2} - 644.405 - 1.3825 - 4.925 - 13.1575 = 1.26.$$

$$SS (\text{error}) = SS (\text{Total}) - SS (\text{Row}) - SS (\text{Column}) - SS (\text{Treatment}) - SS (\text{Observation})$$

$$= 22.835 - 1.3825 - 4.925 - 13.1575 - 1.26 = 2.11.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{.05}$	P -value
Row	3	1.3825	0.4608	3.49	3.24	< 0.05
Column	3	4.925	1.6417	12.45	3.24	0.00
Treatment	3	13.1575	4.3858	33.25	3.24	0.00
Observation	6	1.26	0.21	1.59	2.74	> 0.05
Error	16	2.11	0.1319	—	—	—
Total	31					

Since $F_3 = 33.25 > F_{0.05}$ [P -value = 0.00], the levels of nitrogen are significantly different. At this stage the nitrogen levels are grouped by Duncan's multiple range test, where the test statistic is

$$D_l = d_{0.05, l, f} \sqrt{\frac{s_4}{rk}}, \text{ where } f = 16; l = 2, 3, 4.$$

$$D_2 = 3.00 \sqrt{\frac{0.1319}{2 \times 4}} = 0.38, \quad D_3 = 3.15 \sqrt{\frac{0.1319}{2 \times 4}} = 0.40, \quad D_4 = 3.23 \sqrt{\frac{0.1319}{2 \times 4}} = 0.41.$$

The means of nitrogen levels in ascending order are $\bar{A} = 3.74, \bar{B} = 4.06, \bar{C} = 4.75, \bar{D} = 5.40$.

$\bar{D} - \bar{A} = 5.40 - 3.74 = 1.66 > D_4$, \therefore means are significantly different.

$\bar{C} - \bar{A} = 4.75 - 3.74 = 1.01 > D_3$, \therefore A and C are different.

$\bar{D} - \bar{B} = 5.40 - 4.06 = 1.34 > D_3$, \therefore B and D are different.

$\bar{B} - \bar{A} = 4.06 - 3.74 = 0.32 < D_2$, \therefore A and B are similar.

$\bar{C} - \bar{B} = 4.75 - 4.06 = 0.69 > D_2$, \therefore B and C are different.

$\bar{D} - \bar{C} = 5.40 - 4.75 = 0.65 > D_2$, \therefore C and D are different.

The underlined means do not differ significantly. $\underline{\bar{A}}$, $\underline{\bar{B}}$, $\underline{\bar{C}}$, $\underline{\bar{D}}$

3.13 Analysis of p Latin Square Designs

In agricultural experiment or in medical experiment, a treatment or a group of treatments should not be recommended on the basis of its performance observed in one experiment. For example, a particular level of fertilizer may be suitable in one soil condition but not in all soil conditions. To recommend the fertilizer for all agroclimatic conditions, it is necessary to perform the experiment in all agroclimatic conditions and suitability of a fertilizer should be detected by combined analysis of data recorded from all experiments. The experiment may be repeated in different places or in different seasons or in both.

Let us consider that a $k \times k$ Latin square design is repeated over p places, where places are randomly selected. Let y_{hijl} be the result of l th treatment in j th column of i th row in h th place ($h = 1, 2, \dots, p$; $i = j = l = 1, 2, \dots, k$). The model for such data is

$$y_{hijl} = \mu + \alpha_h + \beta_{hi} + \gamma_{hj} + \delta_l + (\alpha\delta)_{hl} + e_{hijl},$$

where μ = general mean, α_h = effect of h -th place, β_{hi} = effect of i -th row in h -th place, γ_{hj} = effect of j -th column in h -th place, δ_l = effect of l -th treatment, $(\alpha\delta)_{hl}$ = interaction of l -th treatment with h -th place and e_{hijl} = random error.

γ_{hj} = effect of j -th column in h -th place, δ_l = effect of l -th treatment, $(\alpha\delta)_{hl}$ = interaction of l -th treatment with h -th place, and e_{hijl} = random error.

Assumption : (i) $\alpha_h \sim NID(0, \sigma_\alpha^2)$, (ii) $\beta_{hi} \sim NID(0, \sigma_\beta^2)$, (iii) $\gamma_{hj} \sim NID(0, \sigma_\gamma^2)$,
(iv) $(\alpha\delta)_{hl} \sim NID(0, \sigma_{\alpha\delta}^2)$, (v) $e_{hijl} \sim NID(0, \sigma^2)$.

Here δ_l is fixed effect and its estimate is $\hat{\delta}_l = \bar{y}_{...l} - \bar{y}_{...}$.

The total sum of squares can be partitioned as follows :

$$\begin{aligned} \sum_p \sum_k \sum_k \sum_k (y_{hijl} - \bar{y}_{...})^2 &= \sum_p \sum_k \sum_k \sum_k [(\bar{y}_{h...} - \bar{y}_{...}) + (\bar{y}_{hi..} - \bar{y}_{h...}) + (\bar{y}_{h.j.} - \bar{y}_{h...}) \\ &\quad + (\bar{y}_{...l} - \bar{y}_{...}) + (\bar{y}_{h..l} - \bar{y}_{h...}) - (\bar{y}_{...l} - \bar{y}_{...}) + (\bar{y}_{hijl} - \bar{y}_{hi..} \\ &\quad - \bar{y}_{h.j.} - \bar{y}_{i..l} + 2\bar{y}_{h...})]^2 \\ &= k^2 \sum (\bar{y}_{h...} - \bar{y}_{...})^2 + k \sum \sum (\bar{y}_{hi..} - \bar{y}_{h...})^2 + k \sum \sum (\bar{y}_{h.j.} - \bar{y}_{h...})^2 + pk \sum (\bar{y}_{...l} - \bar{y}_{...})^2 \\ &\quad + k \sum \sum (\bar{y}_{h..l} - \bar{y}_{h...} - \bar{y}_{...l} + \bar{y}_{...})^2 + \sum \sum \sum \sum (\bar{y}_{hijl} - \bar{y}_{hi..} - \bar{y}_{h.j.} - \bar{y}_{i..l} + 2\bar{y}_{h...})^2 \end{aligned}$$

The main objective of the analysis is to test the significance of the following hypotheses :

- (i) $H_0 : \delta_l = 0$, against $H_A : \delta_l \neq 0$
- (ii) $H_0 : \sigma_{\alpha\delta}^2 = 0$, against $H_A : \sigma_{\alpha\delta}^2 > 0$

- (iii) $H_0 : \sigma_\gamma^2 = 0$, against $H_A : \sigma_\gamma^2 > 0$
- (iv) $H_0 : \sigma_\beta^2 = 0$, against $H_A : \sigma_\beta^2 > 0$
- (v) $H_0 : \sigma_\alpha^2 = 0$, against $H_A : \sigma_\alpha^2 > 0$.

To decide the test statistic for any of the above hypotheses, we need to find $E(SS)$. The expected sum of squares are calculated as follows :

$$\begin{aligned}
 E[SS (\text{error})] &= E \sum_h^p \sum_i^k \sum_j^k \sum_l^k [y_{hijl} - \bar{y}_{hi..} - \bar{y}_{h.j.} - \bar{y}_{h..l} + 2\bar{y}_{h...}]^2 \\
 &= \sum_h \sum_i \sum_j \sum_l E[e_{hijl} - \bar{e}_{hi..} - \bar{e}_{h.j.} - \bar{e}_{h..l} + 2\bar{e}_{h...}]^2 \\
 &= \sum_h \sum_i \sum_j \sum_l [E(e_{hijl})^2 + E(\bar{e}_{hi..}^2) + E(\bar{e}_{h.j.}^2) + E(\bar{e}_{h..l}^2) \\
 &\quad + 4E \sum_h \sum_i \sum_j \sum_l (\bar{e}_{h...}^2)] \\
 &\quad - 2E \sum_h \sum_i \sum_j \sum_l (e_{hijl}, \bar{e}_{hi..}) - 2E \sum_h \sum_i \sum_j \sum_l (e_{hijl}, \bar{e}_{h.j.}) \\
 &\quad - 2E \sum_h \sum_i \sum_j \sum_l (e_{hijl}, \bar{e}_{h..l}) + 4E \sum_h \sum_i \sum_j \sum_l (e_{hijl}, \bar{e}_{h...}) \\
 &\quad + 2E \sum_h \sum_i \sum_j \sum_l (\bar{e}_{hi..}, \bar{e}_{h.j.}) + 2E \sum_h \sum_i \sum_j \sum_l (\bar{e}_{hi..}, \bar{e}_{h..l}) \\
 &\quad - 4E \sum_h \sum_i \sum_j \sum_l (\bar{e}_{hi..}, \bar{e}_{h...}) + 2E \sum_h \sum_i \sum_j \sum_l (\bar{e}_{h.j.}, \bar{e}_{h..l}) \\
 &\quad - 4E \sum_h \sum_i \sum_j \sum_l (\bar{e}_{h.j.}, \bar{e}_{h...}) - 4E \sum_h \sum_i \sum_j \sum_l (\bar{e}_{h..l}, \bar{e}_{h...}) \\
 &= p(k-1)(k-2)\sigma^2.
 \end{aligned}$$

$$\begin{aligned}
 E[SS (\text{Squares})] &= Ek^2 \sum_h (\bar{y}_{h...} - \bar{y}_{....})^2 \\
 &= k^2 E \sum_h [\alpha_h - \bar{\alpha} + \bar{\beta}_h - \bar{\beta}_{..} + \bar{\gamma}_h - \bar{\gamma}_{..} + (\alpha\bar{\delta})_h - (\alpha\bar{\delta})_{..} + \bar{e}_{h...} - \bar{e}_{....}]^2 \\
 &= (p-1)[\sigma^2 + k\sigma_\beta^2 + k\sigma_\gamma^2 + k\sigma_{\alpha\delta}^2 + k^2\sigma_\alpha^2].
 \end{aligned}$$

Other expected sum of squares are calculated similarly.

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	$E(MS)$
Squares	$p-1$	S_1	s_1	$\sigma^2 + k\sigma_\beta^2 + k\sigma_\gamma^2 + k\sigma_{\alpha\delta}^2 + k^2\sigma_\alpha^2$
Row within Squares	$p(k-1)$	S_2	s_2	$\sigma^2 + k\sigma_\beta^2$
Column within Squares	$p(k-1)$	S_3	s_3	$\sigma^2 + k\sigma_\gamma^2$
Treatment	$k-1$	S_4	s_4	$\sigma^2 + k\sigma_{\alpha\delta}^2 + \frac{pk}{k-1} \sum (\delta_l - \bar{\delta})^2$
Squares \times Treatment	$(p-1)(k-1)$	S_5	s_5	$\sigma^2 + k\sigma_{\alpha\delta}^2$
Error	$p(k-1)(k-2)$	S_6	s_6	σ^2
Total	$pk^2 - 1$			

The test statistic for $H_0 : \sigma_{\alpha\delta}^2 = 0$ is $F_5 = s_5/s_6$. If this hypothesis is rejected, the test statistic for $H_0 : \delta_l = 0$ is $F_4 = s_4/s_5$. This F_4 has $(k-1)$ and $(p-1)(k-1)$ d.f. If $H_0 : \sigma_{\alpha\delta}^2 = 0$ is true, then s_4 is to be compared with the pooled value of s_5 and s_6 . That is

$$F_4 = \frac{s_4}{\frac{(p-1)(k-1)s_5 + p(k-1)(k-2)s_6}{(k-1)(pk-p-1)}}$$

This F_4 has $(k-1)$ and $(k-1)(pk-p-1)$ d.f.

The test statistics for $H_0 : \sigma_\gamma^2 = 0$ and $H_0 : \sigma_\beta^2 = 0$ are $F_3 = s_3/s_6$ and $F_2 = s_2/s_6$, respectively. However, the test statistic for $H_0 : \sigma_\alpha^2 = 0$ is not found out directly. Under this latter hypothesis

$$E[(s_1 + 2s_6)] = E[s_2 + s_3 + s_5].$$

Therefore, the test statistic is $F_1 = \frac{s_1 + 2s_6}{s_2 + s_3 + s_5}$.

This F_1 is approximately distributed as variances ratio [Satterthwaite (1946)] with

$$\frac{(s_1 + 2s_6)^2}{\frac{s^2}{p-1} + \frac{(2s_6)^2}{p(k-1)(k-2)}} \quad \text{and} \quad \frac{(s_2 + s_3 + s_5)^2}{\frac{s_2^2}{p(k-1)} + \frac{(s_3^2)^2}{p(k-1)} + \frac{(s_5^2)^2}{(p-1)(k-1)}}$$

Example 3.12 : In a dairy farm an experiment of feeding trial is performed with 4 different types of fodder (F_1, F_2, F_3 and F_4) to identify a best fodder for increased milk production. The fodder is fed to cows of different origin and different lactation periods. The design used is LSD and the experiment is repeated in 3 different sheds. After 4 weeks of start of feeding trial the milk production of one day per experimental cow is recorded for analysis. The milk production data are as follows :

Milk Production (y_{hijl} kg) of Cows

Experiment-1

Origin of cows	Lactation Period			
	L_1	L_2	L_3	L_4
C_1	F_1 -32.5	F_2 -35.1	F_3 -34.6	F_4 -38.6
C_2	F_2 -34.6	F_3 -32.2	F_4 -30.2	F_1 -30.5
C_3	F_3 -30.5	F_4 -35.6	F_1 -28.5	F_2 -25.0
C_4	F_4 -20.4	F_1 -18.2	F_2 -16.4	F_3 -20.2

Total Production of Rows

Experiment	Rows ($y_{hi..}$)			
	1	2	3	4
1	140.8	127.5	119.6	75.2
2	134.8	126.2	118.2	71.9
3	134.0	122.2	115.4	69.0
$y_{.i}$	409.6	375.9	353.2	216.1

Experiment-2

Origin of cows	Lactation Period			
	L_1	L_2	L_3	L_4
C_1	F_1 -30.2	F_2 -33.2	F_3 -35.2	F_4 -36.2
C_2	F_2 -35.6	F_3 -30.0	F_4 -34.1	F_1 -26.5
C_3	F_3 -32.8	F_4 -31.2	F_1 -25.5	F_2 -28.7
C_4	F_4 -17.5	F_1 -15.5	F_2 -18.7	F_3 -20.2

Total Production of Columns

Experiment	Columns ($y_{h.j}$)			
	1	2	3	4
1	118.0	121.1	109.7	114.3
2	116.1	109.9	113.5	111.6
3	118.4	120.4	103.3	98.5
$y_{.j}$	352.5	351.4	326.5	324.4

Experiment-3

Total Production of Treatments

Origin, of cows	Lactation Period			
	L_1	L_2	L_3	L_4
C_1	$F_1-38.2$	$F_2-35.0$	$F_3-30.6$	$F_4-30.2$
C_2	$F_2-33.6$	$F_3-32.0$	$F_4-32.1$	$F_1-24.5$
C_3	$F_3-30.4$	$F_4-36.2$	$F_1-26.2$	$F_2-22.6$
C_4	$F_4-16.2$	$F_1-17.2$	$F_2-14.4$	$F_3-21.2$

Experiment	Treatments ($y_{h..l}$)			
	F_1	F_2	F_3	F_4
1	109.7	111.1	117.5	124.8
2	97.7	116.2	118.2	119.0
3	106.1	105.6	114.2	114.7
$y_{..l}$	313.5	332.9	349.9	358.5

Analyse the data and recommend the best fodder.

Solution : We have $k = 4, p = 3, G = 1354.8, C.T. = \frac{G^2}{pk^2} = 38239.23$.

$$SS \text{ (Total)} = \sum \sum \sum \sum y_{hijl}^2 - C.T. = 40314.91 - 38239.23 = 2075.68.$$

$$SS \text{ (Treatment)} = \frac{\sum y_{..l}^2}{pk} - C.T. = \frac{460056.92}{3 \times 4} - 38239.23 = 98.8467.$$

$$\begin{aligned}
 SS \text{ (Row within squares)} &= \sum_h \left[\sum_i \frac{y_{hi..}^2}{k} - \frac{y_{h...}^2}{k^2} \right], h_{1...} = 463.1, h_{2...} = 451.1, h_{3...} = 440.6. \\
 &= \left(\frac{56040.09}{4} - \frac{(463.1)^2}{16} \right) + \left(\frac{53238.33}{4} - \frac{(451.1)^2}{16} \right) \\
 &\quad + \left(\frac{50967.00}{4} - \frac{(440.6)^2}{16} \right) \\
 &= (14010.0225 - 13403.85) + (13309.5825 - 12718.2006) \\
 &\quad + (12741.75 - 12133.0225) \\
 &= 1806.2819.
 \end{aligned}$$

$$\begin{aligned}
 SS \text{ (Column within squares)} &= \sum_h \left[\sum_j \left(\frac{y_{h.j.}^2}{k} - \frac{y_{h...}^2}{k^2} \right) \right] \\
 &= \left(\frac{53687.79}{4} - \frac{(463.1)^2}{16} \right) + \left(\frac{50894.03}{4} - \frac{(451.1)^2}{16} \right) \\
 &\quad + \left(\frac{48887.86}{4} - \frac{(440.6)^2}{16} \right) \\
 &= (13421.9475 - 13403.85) + (12723.5075 - 12718.2006) \\
 &\quad + (12221.965 - 12133.0225) \\
 &= 112.3469
 \end{aligned}$$

$$SS \text{ (Squares)} = \frac{\sum y_{h...}^2}{k^2} - C.T. = \frac{612081.18}{16} - 38239.23 = 15.8437.$$

$$\begin{aligned}
 SS \text{ (Squares} \times \text{Treatment)} &= \frac{\sum \sum y_{h..l}^2}{k} - C.T. - SS \text{ (Squares)} - SS \text{ (Treatment)} \\
 &= \frac{153544.86}{4} - 38239.23 - 15.8437 - 98.8467 = 32.2946.
 \end{aligned}$$

$$\begin{aligned}
 SS (\text{Error}) &= SS (\text{Total}) - SS (\text{Squares}) - SS (\text{Rows within squares}) \\
 &\quad - SS (\text{Column within squares}) - SS (\text{Treatment}) - SS (\text{Squares} \times \text{treatment}) \\
 &= 2075.68 - 15.8437 - 1806.2819 - 112.3469 - 98.8467 - 32.2946 = 10.0662.
 \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F	F _{0.05}	P-value
Square	2	15.8437	7.92185	0.08	—	—
Row within square	9	1806.2819	200.698	358.88	2.47	0.00
Column within square	9	112.3469	12.483	22.32	2.47	0.00
Treatment	3	98.8467	32.9489	6.12	4.76	< 0.05
Square × treatment	6	32.2946	5.3824	9.62	2.66	> 0.00
Error	18	10.0662	0.55923	—	—	—
Total	47					

Since $F_5 = 1454.70 > F_{0.05}$ ($P\text{-value} < 0.01$), the interaction of square and treatment is highly significant and hence, $F_4 = s_4/s_5 = 6.12$. This $F_4 > F_{0.05}$. Thus, the treatments are significantly different. Here, $F_2 = s_2/s_6$ and $F_3 = s_3/s_6$. The rows and columns are also found significantly different. The milk production varies significantly within the variation in lactation period of cows. Also, it varies with the variation in source of cows.

The test statistic for $H_0 : \sigma_\alpha^2 = 0$ is

$$F_1 = \frac{s_1 + 2s_6}{s_2 + s_3 + s_5} = \frac{7.92185 + 2(0.0037)}{200.698 + 13.5941 + 5.3824} = 0.08.$$

The numerator d.f. of F_1 is

$$\frac{(s_1 + 2s_6)^2}{\frac{s_1^2}{p-1} + \frac{(2s_6)^2}{p(k-1)(k-2)}} = \frac{[7.92185 + 2(0.0037)]^2}{\frac{(7.92185)^2}{3-1} + \frac{[2(0.0037)]^2}{3(4-1)(4-2)}} = 2.$$

The denominator d.f. of F_1 is

$$\frac{(s_2 + s_3 + s_5)^2}{\frac{s_2^2}{p(k-1)} + \frac{s_3^2}{p(k-1)} + \frac{s_5^2}{(p-1)(k-1)}} = \frac{(200.698 + 13.5941 + 5.3824)^2}{\frac{(200.698)^2}{9} + \frac{(13.5941)^2}{9} + \frac{(5.3824)^2}{6}} = 11.$$

Since $F_1 = 0.08 < F_{0.05; 2, 11} = 3.98$, the results observed in different squares are similar.

To identify the best fodder we need to compare all the varieties of fodder. The comparison is done by Duncan's multiple range test, where the test statistic is

$$D_l = d_{0.05, l, f} \sqrt{\frac{s_5}{pk}}, \quad l = 2, 3, 4; \quad f = 6.$$

Here s_5 is used since this s_5 is used as denominator to calculate F_4 . Now

$$D_2 = 3.46 \sqrt{\frac{5.3824}{3 \times 4}} = 2.32, \quad D_3 = 3.58 \sqrt{\frac{5.3824}{3 \times 4}} = 2.40, \quad D_4 = 3.64 \sqrt{\frac{5.3824}{3 \times 4}} = 2.44.$$

The means in ascending order are $\bar{F}_1 = 26.125$, $\bar{F}_2 = 27.742$, $\bar{F}_3 = 29.158$, $\bar{F}_4 = 29.875$.

$$\bar{F}_4 - \bar{F}_1 = 29.875 - 26.125 = 3.75 > D_4, \quad \therefore \text{the means differ significantly.}$$

$$\bar{F}_4 - \bar{F}_2 = 29.875 - 27.742 = 2.133 < D_3, \quad \therefore F_2 \text{ and } F_4 \text{ are similar.}$$

$$\bar{F}_3 - \bar{F}_1 = 29.158 - 26.125 = 3.033 > D_2, \quad \therefore F_1 \text{ and } F_3 \text{ differ significantly.}$$

$$\bar{F}_2 - \bar{F}_1 = 27.742 - 26.125 = 1.617 < D_2, \quad \therefore F_1 \text{ and } F_2 \text{ are similar.}$$

The underlined means do not differ significantly

$$\underline{\bar{F}_1, \bar{F}_2, \bar{F}_3, \bar{F}_4}$$

It is observed that better milk production is recorded from the cows when F_4 is fed. However, F_2, F_3 and F_4 are similar.

3.14 Orthogonal Latin Square Designs

Two Latin square designs of same order with two different sets of treatments are said to be orthogonal if the treatments of one appear once and only once with treatments of another one when one square is superimposed on other. For example, let us consider two squares with one set of treatments A, B and C and another set of treatments α, β and γ as follows.

Square-1	Square-2																		
<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>A</td><td>B</td><td>C</td></tr> <tr><td>B</td><td>C</td><td>A</td></tr> <tr><td>C</td><td>A</td><td>B</td></tr> </table>	A	B	C	B	C	A	C	A	B	<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>α</td><td>β</td><td>γ</td></tr> <tr><td>γ</td><td>α</td><td>β</td></tr> <tr><td>β</td><td>γ</td><td>α</td></tr> </table>	α	β	γ	γ	α	β	β	γ	α
A	B	C																	
B	C	A																	
C	A	B																	
α	β	γ																	
γ	α	β																	
β	γ	α																	

Now, square-1 is superimposed on square-2 and the resultant square takes the following shape :

Square-3		
$A\alpha$	$B\beta$	$C\gamma$
$B\gamma$	$C\alpha$	$A\beta$
$C\beta$	$A\gamma$	$B\alpha$

It is seen that any treatment of square-1 is appeared once and only once with any treatment of square-2. So, square-1 and square-2 are orthogonal.

Orthogonal Latin square designs are available in table prepared by Fisher and Yates (1948). In case of $k \times k$ Latin square design there are maximum $(k - 1)$ orthogonal Latin square designs. If any two Latin square designs of a set of latin square designs are mutually orthogonal, then the set gives orthogonal Latin square designs.

The analysis of this design is done as usual. Here analysis of orthogonal Latin square design of m th order is presented. Let us consider m sets of letters as follows :

$$\begin{array}{cccc}
 A_{11} & A_{12} & \cdots & A_{1k} \\
 A_{21} & A_{22} & \cdots & A_{2k} \\
 \cdots & \cdots & \cdots & \cdots \\
 A_{m1} & A_{m2} & \cdots & A_{mk}
 \end{array}$$

where $m < k$. Let there be k^2 plots for an experiment, where plots are arranged in k rows

and k columns. Now, if the mk letters are arranged in k^2 plots, we shall get orthogonal Latin square design under the following conditions :

(i) In each plot there is a combination of A_1, A_2, \dots, A_m letters. The combination of plots of i -th row and j -th column are denoted by $A_{1(ij)}, A_{2(ij)}, \dots, A_{m(ij)}$.

$$\begin{array}{cccc} \text{(ii)} & A_{t(11)} & A_{t(12)} & \cdots & A_{t(1k)} \\ & A_{t(21)} & A_{t(22)} & \cdots & A_{t(2k)} \\ & \cdots & \cdots & \cdots & \cdots \\ & A_{t(k1)} & A_{t(k2)} & \cdots & A_{t(kk)} \end{array}$$

Here $t = 1, 2, \dots, m$. This above square is a latin square.

(iii) Any combination of A_t letter and $A_{t'}$ letter ($t \neq t'$) will appear once and only once in a plot.

If $m = 2$, then the resultant design is called Graeco latin square design. If k is the power of prime number, then orthogonal latin square of m -th order is available, when $m \leq k - 1$

The model for the analysis of this design is

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{1(ij)} + \gamma_{2(ij)} + \cdots + \gamma_{m(ij)} + e_{ijl},$$

where $y_{ijl} = (ij)$ th value of l th treatment, $\mu =$ general mean, $\alpha_i =$ effect of i th row, $\beta_j =$ effect of j th column, $\gamma_{tl} =$ effect of A_{tl} ($t = 1, 2, \dots, m; i = j = l = 1, 2, \dots, k$).

The normal equations to estimate the parameters of the model are :

$$y_{...} = k^2 \hat{\mu} + k \sum \hat{\alpha}_i + k \sum \hat{\beta}_j + k \sum_t \sum_l \hat{\gamma}_{tl}$$

$$y_{i..} = k \hat{\mu} + k \hat{\alpha}_i + \sum \hat{\beta}_j + \sum \sum \hat{\gamma}_{tl}$$

$$y_{.j.} = k \hat{\mu} + \sum \hat{\alpha}_i + k \hat{\beta}_j + \sum \sum \hat{\gamma}_{tl}$$

$$T_{ul} = k \hat{\mu} + \sum \hat{\alpha}_i + \sum \hat{\beta}_j + k \hat{\gamma}_{ul} + k \sum_{v'} \sum_{v''} \hat{\gamma}_{v'v''}$$

There are $(m + 2)(k - 1) + 1$ independent normal equations among the equations shown above. The solution of these equations provides

$$\hat{\mu} = \bar{y}_{...}, \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \hat{\gamma}_{ul} = \frac{T_{ul}}{k} - \bar{y}_{...}$$

Here T_{ul} is the total result of A_{ul} treatment. The sum of squares due to estimates is

$$SS(\text{estimate}) = \frac{\sum y_{i..}^2}{k} + \frac{\sum y_{.j.}^2}{k} - \frac{y_{...}^2}{k^2} + \sum_{t=1}^m \left[\frac{\sum T_{tl}^2}{k} - \frac{y_{...}^2}{k^2} \right].$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$E(MS)$
Row	$k - 1$	S_1	s_1	$\frac{s_1}{s_e}$	$\sigma^2 + \frac{k}{k-1} \sum \alpha_i^2$
Column	$k - 1$	S_2	s_2	$\frac{s_2}{s_e}$	$\sigma^2 + \frac{k}{k-1} \sum \beta_j^2$
A_1 -letter	$k - 1$	S_3	s_3	$\frac{s_3}{s_e}$	$\sigma^2 + \frac{k}{k-1} \sum \gamma_{1i}^2$
A_2 -letter	$k - 1$	S_4	s_4	\vdots	$\sigma^2 + \frac{k}{k-1} \sum \gamma_{2i}^2$
...	\vdots		
A_m -letter	$k - 1$	S_{m-2}	s_{m-2}	$\frac{s_{m-2}}{s_e}$	$\sigma^2 + \frac{k}{k-1} \sum \gamma_{mi}^2$
Error	$(k - 1)(k - m - 1)$	S_e	s_e	—	σ^2

The null hypothesis is $H_0 : \gamma_{m1} = \gamma_{m2} = \dots = \gamma_{mk} = 0$.

The test statistic for $\gamma_{mk} = 0$ is $F = \frac{s_{m-2}}{s_e}$.

3.15 Graeco Latin Square Design

Let there be three treatments A, B and C . Consider that there is a factor with levels α, β and γ . The variation of this factor along with the variation in rows and columns are needed to be controlled by design so that the treatment effects are estimated avoiding the impacts of three external sources of variation. The plan of such design is as follows :

Graeco Latin Square Design

$A\alpha$	$B\beta$	$C\gamma$
$B\gamma$	$C\alpha$	$A\beta$
$C\beta$	$A\gamma$	$B\alpha$

Here it is observed that each treatment is allocated once and only once in a row and in a column. Also, it is observed that each treatment appears once and only once with each level of third factor, where the levels of third factor are α, β and γ . The arrangement of treatments in rows, columns and with Greek letters is known as Graeco Latin square design.

The design is used to control three external sources of variation. The randomisation of treatment in plots of a Graeco Latin square design is similarly done as it is done in Latin square design. The analysis of data of this design is also similar except that the sum of squares due to Greek letter is calculated and total sum of squares is decomposed into four identified components of variation.

The design is a 4-way design since it has four components viz. row, column, treatment and Greek letter. Since there are 4 factors, the analysis of 3×3 Graeco Latin square design is not done unless the design is repeated several times.

Let y_{ijlm} be the observation of l th treatment of j th column in i th row corresponding to m th level of greek letters. The model for this observation is

$$y_{ijlm} = \mu + \alpha_i + \beta_j + \gamma_l + \delta_m + e_{ijlm},$$

where μ = general mean, α_i = effect of i th row, β_j = effect of j th column, γ_l = effect of l th treatment, δ_m = effect of m th Greek letter and e_{ijlm} = random error.

$i = j = l = m = 1, 2, \dots, k$. The analysis of variance table is shown below :

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	$E(MS)$	F
Row	$k - 1$	S_1	s_1	$\sigma^2 + \frac{k}{k-1} \sum \alpha_i^2$	$\frac{s_1}{s_5}$
Column	$k - 1$	S_2	s_2	$\sigma^2 + \frac{k}{k-1} \sum \beta_j^2$	$\frac{s_2}{s_5}$
Treatment	$k - 1$	S_3	s_3	$\sigma^2 + \frac{k}{k-1} \sum \gamma_l^2$	$\frac{s_3}{s_5}$
Greek letter	$k - 1$	S_4	s_4	$\sigma^2 + \frac{k}{k-1} \sum \delta_m^2$	$\frac{s_4}{s_5}$
Error	$(k - 1)(k - 3)$	S_5	s_4	σ^2	
Total	$k^2 - 1$				

Here $S_1 = \frac{\sum y_{i...}^2}{k} - \text{C.T.}$, $\text{C.T.} = \frac{G^2}{k^2} = \frac{(y_{...})^2}{k^2}$, $S_2 = \frac{1}{k} \sum y_{.j..}^2 - \text{C.T.}$,

$S_3 = \frac{1}{k} \sum y_{..l.}^2 - \text{C.T.}$, $S_4 = \frac{1}{k} \sum y_{...m}^2 - \text{C.T.}$, $S_5 = SS (\text{Total}) - S_1 - S_2 - S_3 - S_4$,

$SS (\text{Total}) = \sum \sum \sum \sum y_{ijlm}^2 - \text{C.T.}$

Chapter 4

Factorial Experiment

4.1 Introduction

The objective of the controlled experiment is to study the behaviour of certain characteristic under some controlled conditions. The characteristic is usually known as treatment, where a treatment has different levels. In some experiment the characteristic of more than one factor is also under investigation simultaneously. For example, let us consider that 5 new varieties of wheat are discovered by a team of agriculture scientists. The team needs to identify the best variety for general agroclimatic condition. To identify the best variety of wheat, an experiment needs to be conducted using different doses of fertilizer and different levels of irrigation. There are three factors, viz., wheat, fertilizer and irrigation. If all the three factors have same levels, an experiment through Latin square design can be performed using doses of fertilizer (say, nitrogen) in rows, irrigation in columns and the varieties of wheat are allocated in rows and columns. If there are only two factors, say, wheat variety and fertilizer, an experiment can be conducted through randomized block design using levels of fertilizer in blocks and varieties of wheat are randomly allocated in all blocks.

In practice, there may be more factors and each factor may have many levels. The basic designs so far we have discussed are not sufficient to study the impacts of more factors simultaneously. Alternatively, all factor levels may be combined and the combination of factor levels may be used as treatments in any experiment. The experiment in which combination of different factor levels are used as treatment is known as *factorial experiment*. These treatments are again allocated to the plots of some basic design. The usual design for factorial experiment is randomized block design, where treatments are allocated to the plots of a block by a random process.

Let there be m factors F_1, F_2, \dots, F_m . Consider that i -th factor ($i = 1, 2, \dots, m$) has s_i levels. The total number of level combinations are $s_1 \times s_2 \times \dots \times s_m$. The experiment in which all these factor level combinations are used as treatment is called factorial experiment. In such experiment the treatment effect is not studied, rather the main effect of each factor and the interaction of two or more factors are studied. For example, let us consider that nitrogen, potash and phosphorus are suitable for any experiment. Consider that each fertilizer has 2 levels. The total number of level combinations is $2 \times 2 \times 2 = 2^3$. If in any experiment 2^3 level combinations are used as treatment, then the experiment is called *factorial experiment*. In general, if there are n factors each having p levels, we have p^n factorial experiment.

Asymmetrical factor experiment : The factorial experiment in which different factors have different levels is called *asymmetrical factorial experiment*. Let us consider that five doses of nitrogen are to be tested in presence of 3 doses of potash and 4 doses of phosphorus. Then total level combinations are $5 \times 3 \times 4 = 60$. Any experiment conducted with such 60-level combinations as treatment is known as *asymmetrical factorial experiment*.

Main effects and interactions : Let there be three factors, viz., nitrogen (N), phosphorus (P) and potash (K) having 2 levels each. Here capital letters are used to indicate a factor and

small letters are used to indicate the levels of the factors. The factor-level combinations can be written as

$$n_0p_0k_0, n_1p_0k_0, n_0p_1k_0, n_1p_1k_0, n_0p_0k_1, n_1p_0k_1, n_0p_1k_1, n_1p_1k_1.$$

We have used 0 as first level and 1 as second level. If it is needed to compare the product of 60 kg/ha of nitrogen with the product of 0 kg/ha, then 0 kg/ha is considered as first level, where nitrogen is absent and 60 kg/ha is considered as second level, where nitrogen is present. Thus, first level and second level can be considered as absent and present of a factor, respectively. We can write the factor levels as

$$000, 100, 010, 110, 001, 101, 011, 111$$

or, (1); n, p, np, k, nk, pk, npk .

In the latter case (1) is used to indicate the absence of all factors.

Let us consider that an experiment is conducted with these 8-level combinations as treatment through a randomized block design. For simplicity, let us denote the result of corresponding treatment by the treatment itself. Hence, the result of the experiments are :

$$(1), n, p, np, k, nk, pk, npk.$$

Now, the effects of nitrogen, phosphorus and potash are evaluated as follows :

Simple effect of nitrogen (N) in absence of phosphorus (P) and potash (K) = $n - (1)$

Simple effect of N in presence of P but in absence of K = $np - p$

Simple effect of N in presence of K but in absence of P = $nk - k$

Simple effect of N in presence of both K and P = $npk - pk$.

The total effect of N, denoted by $[N]$, is the sum of these simple effects. Thus,

$$[N] = (npk - pk) + (nk - k) + (np - p) + (n - (1)) = (n - 1)(p + 1)(k + 1), \text{ symbolically.}$$

The average effect of N, known as main effect of N, is denoted by N is

$$\begin{aligned} N &= \frac{1}{4}[(npk - pk) + (nk - k) + (np - p) + (n - (1))] \\ &= \frac{1}{2^2}(n - 1)(p + 1)(k + 1) = \frac{1}{2^{3-1}}(n - 1)(p + 1)(k + 1) = \frac{[N]}{2^{3-1}}. \end{aligned}$$

Thus, main effect is the average of the simple effects.

Simple effect of P in absence of N and K = $p - (1)$

Simple effect of P in absence of N but in presence of K = $pk - k$

Simple effect of P in absence of K but in presence of N = $pn - n$

Simple effect of P in presence of both N and K = $npk - nk$

Total effect of P, denoted by $[P]$, is the sum of these 4 simple effects. Thus,

$$[P] = (p - (1)) + (pk - k) + (pn - n) + (npk - nk) = (n + 1)(p - 1)(k + 1).$$

Main effect of P is the average of these 4 simple effects. Thus,

$$P = \frac{1}{4}(n + 1)(p - 1)(k + 1) = \frac{[P]}{2^{3-1}}.$$

Simple effect of K in absence of both N and P = $k - (1)$

Simple effect of K in presence of N but in absence of P = $nk - n$

Simple effect of K in absence of N but in presence of P = $pk - p$

Simple effect of K in presence of both N and P = $npk - np$

Total effect of K, denoted by $[K]$, is the sum of these 4 simple effects. We have

$$[K] = (k - (1)) + (nk - n) + (pk - p) + (npk - np) = (n + 1)(p + 1)(k - 1).$$

Main effect of K is the average of these 4 simple effects. We write

$$K = \frac{1}{4}(n + 1)(p + 1)(k - 1) = \frac{[K]}{2^{3-1}}.$$

It is observed that the effect of any factor is measured by subtracting the result of first level from second level [presence minus absence]. Now,

Effect of N in presence of P = $(npk - pk) + np - p$.

Effect of N in absence of P = $(nk - k) + (n - (1))$.

Therefore, the effect of N in presence of P minus the effect of N in absence of P will give the effect of N and P. This effect is called interaction of N and P. The total of this interaction is

$$[NP] = (npk - pk) + (np - p) - \{(nk - k) + (n - (1))\} = (n - 1)(p - 1)(k + 1).$$

This is also linear combination of 4 simple effects. The average of these 4 simple effects (linear combination of simple effects) is known as interaction and is written as

$$NP = \frac{1}{4}(n - 1)(p - 1)(k + 1) = \frac{[NP]}{2^{3-1}}.$$

Similarly, we have

$$[NK] = (npk - pk) + (nk - k) - \{(np - p) + (n - (1))\} = (n - 1)(p + 1)(k - 1).$$

$$NK = \frac{[NK]}{2^{3-1}}, [PK] = (n + 1)(p - 1)(k - 1), PK = \frac{[PK]}{2^{3-1}}.$$

$$[NPK] = (n - 1)(p - 1)(k - 1), NPK = \frac{[NPK]}{2^{3-1}}.$$

If the experiment is replicated r times [experiment conducted in randomized design of r blocks], the main effects and interactions are calculated dividing the total effect by $r2^{3-1}$. Thus, we have

$$N = \frac{[N]}{r2^{3-1}}, P = \frac{[P]}{r2^{3-1}}, K = \frac{[K]}{r2^{3-1}}, NP = \frac{[NP]}{r2^{3-1}}, NK = \frac{[NK]}{r2^{3-1}},$$

$PK = \frac{[PK]}{r2^{3-1}}, NPK = \frac{[NPK]}{r2^{3-1}}$. It is observed that there are $(2^3 - 1)$ effects and interactions of 2^3 -factorial experiment.

It is observed that any total effect is the linear combination (contrast) of factor level combinations, where half the levels are with positive sign and half the levels are with negative sign. The positive sign is observed with a level combination if the factor effect under study

is present in the level combination and negative sign is observed if the factor is absent in the level combination. The signs for different level combinations are found out from a sign table as shown below.

The sign table is prepared writing the level combinations in first row of the table. The effect and interactions are written in the first column of the table. The body of the table is filled up writing +ve sign or -ve sign according to presence or absence of factor(s), respectively in a level combination. The sign is negative if odd number of factors are absent in a level combination.

Sign Table for 2³-Factorial Experiment

Effects or contrasts	Level combinations							
	(1)	<i>n</i>	<i>p</i>	<i>np</i>	<i>k</i>	<i>nk</i>	<i>pk</i>	<i>npk</i>
<i>N</i>	-	+	-	+	-	+	-	+
<i>P</i>	-	-	+	+	-	-	+	+
<i>NP</i>	+	-	-	+	+	-	-	+
<i>K</i>	-	-	-	-	+	+	+	+
<i>NK</i>	+	-	+	-	-	+	-	+
<i>PK</i>	+	+	-	-	-	-	+	+
<i>NPK</i>	-	+	+	-	+	-	-	+

If there are *n* factors each with 2 levels, the experiment is known as 2^{*n*}-factorial experiment. The factors are, say, *A, B, C, D, …, N*. Then

$$[A] = (a - 1)(b + 1)(c + 1) \cdots (n + 1), [B] = (a + 1)(b - 1)(c + 1) \cdots (n + 1)$$

$$[N] = (a + 1)(b + 1) \cdots (n - 1), [AB] = (a - 1)(b - 1)(c + 1) \cdots (n + 1)$$

$$[AC] = (a - 1)(b + 1)(c - 1) \cdots (n + 1), \dots, [AN] = (a - 1)(b + 1)(c + 1) \cdots (n - 1)$$

$$[ABC \cdots N] = (a - 1)(b - 1)(c - 1)(d - 1) \cdots (n - 1).$$

In general, the total effect of any factor or total interaction of any two or more factors are written, symbolically as

$$[X] = (a \pm 1)(b \pm 1)(c \pm 1) \cdots (n \pm 1),$$

where the sign is +ve if a factor is absent in *X* and sign is -ve if a factor is present in *X*. The main effect or interaction is written as

$$X = \frac{1}{r2^{n-1}}(a \pm 1)(b \pm 1)(c \pm 1) \cdots (n \pm 1).$$

There are (2^{*n*} - 1) effects and interactions of the type *X* in a 2^{*n*}-factorial experiment.

Calculation of sum of square : In 2^{*n*}- factorial experiment, there are *n* main effects, ^{*n*}*c*₂ two-factor interactions, ^{*n*}*c*₃ three-factor interactions, … ^{*n*}*c*_{*n*} *n*-factor interactions. In general, there are *nc_r* *r*-factor interactions in 2^{*n*}-factorial experiment. Each effect and interaction are estimated with 1 d.f. The sum of squares of effects and interaction is calculated by

$$SS(X) = \frac{[X]^2}{r2^n}.$$

This sum of squares has 1 d.f. The denominator, except *r*, is calculated as the product sum of squares of coefficient, when [X] is written in the form :

$$[X] = (1, \pm 1)(1, \pm 1)(1, \pm 1) \cdots (1, \pm 1).$$

For example, in case of 2^3 -factorial experiment,

$$[NPK] = (1, -1)(1, -1)(1, -1), [(n - 1)(p - 1)(k - 1)].$$

The product of the sum of squares of coefficient is $[1^2 + (-1)^2][1^2 + (-1)^2][(1)^2 + (-1)^2] = 2^3$. The term r is used if the experiment is replicated r times. If the contrast is written in the form like

$$[NPK] = npk + k + p + n - np - nk - pk - (1),$$

then the denominator is the sum of squares of the coefficients in finding contrast. Thus, for NPK the divisor is

$$[(1)^2 + (1)^2 + (1)^2 + (1)^2 + (-1)^2 + (-1)^2 + (-1)^2 + (-1)^2] = 2^3.$$

The sum of squares due to replication (block) is calculated by

$$SS(\text{replication}) = \frac{\sum R_i^2}{2^n} - \text{C.T.},$$

where $\text{C.T.} = \frac{G^2}{r2^n}$, G = grand total of observations and R_i = total of i -th replication (block); $i = 1, 2, \dots, r$.

$SS(\text{Total}) = \sum \sum \sum y_{ijl}^2 - \text{C.T.}$, where y_{ijl} is the observation of l -th replication of j -th level of a factor A and i -th level of another factor B ; $i = 1, 2$; $j = 1, 2$.

The model for 2^2 -factorial experiment when conducted through randomized block design is

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l + (\alpha\beta)_{ij} + e_{ijl},$$

where y_{ijl} is the result of j -th level of B in presence of i -th level of A in l -th block, μ is the general mean, α_i is the effect of i -th level of A , β_j is the effect of j -th level of B , γ_l is the effect of l -th block, $(\alpha\beta)_{ij}$ is the interaction of i -th level of A with j -th level of B . The analysis of the model is to be performed in a similar way as it is done for other analysis of variance model.

The sign table helps in deciding the contrast. But it is not practically applicable to find out the total effect for calculation of sum of squares. For 2^n -factorial experiment there are $(2^n - 1)$ effects and interactions each having 1 d.f. The total effect and sum of squares of effects and interactions are calculated using Yates' algorithm. The method is discussed below :

Yates' Table to Calculate Effect Total and Sum of Squares

Level combinations	Total yield	Operation			Effects	$SS = \frac{[\]^2}{r2^3}$
		1	2	3 = []		
(1)	y_1]	$y_1 + y_2 = u_1$]	$u_1 + u_2 = V_1$]	$V_1 + V_2$	G	C.T.
a	y_2]	$y_3 + y_4 = u_2$]	$u_3 + u_4 = V_2$]	$V_3 + V_4$	A	$SS(A)$
b	y_3]	$y_5 + y_6 = u_3$]	$u_5 + u_6 = V_3$]	$V_5 + V_6$	B	$SS(B)$
ab	y_4]	$y_7 + y_8 = u_4$]	$u_7 + u_8 = V_4$]	$V_7 + V_8$	AB	$SS(AB)$
c	y_5]	$y_2 - y_1 = u_5$]	$u_2 - u_1 = V_5$]	$V_2 - V_1$	C	$SS(C)$
ac	y_6]	$y_4 - y_3 = u_6$]	$u_4 - u_3 = V_6$]	$V_4 - V_3$	AC	$SS(AC)$
bc	y_7]	$y_6 - y_5 = u_7$]	$u_6 - u_5 = V_7$]	$V_6 - V_5$	BC	$SS(BC)$
abc	y_8]	$y_8 - y_7 = u_8$]	$u_8 - u_7 = V_8$]	$V_8 + V_7$	ABC	$SS(ABC)$

Here $V_1 + V_2 = G = \text{grand total}$, $[A] = V_3 + V_4$, and so on.

Here $V_3 + V_4 = u_5 + u_6 + u_7 + u_8 = y_2 - y_1 + y_4 - y_3 + y_6 - y_5 + y_8 - y_7$.

If y_i is replaced by the notation for its treatment combination, we have

$$[A] = V_3 + V_4 = a - (1) + ab - b + ac - c + abc - bc = (a - 1)(b + 1)(c + 1).$$

Similar is the case for other effects and interactions.

Description of the table

- (i) First column contains the treatment combinations systematically and alphabetically.
- (ii) Second column is to represent the total yield from all replications corresponding to the treatment combinations. The total yields are shown in pairs as it is shown in the table.
- (iii) The operation is done with the total yields in pairs. The first half of the values in first operation is calculated from total yields adding total yields in pairs. Second half of the values is obtained by subtraction of total yields in pairs. The subtraction is done subtracting first value from second value in any pair.
- (iv) Second operation is similarly done as it is done in first operation. This operation is done using the result of first operation.
- (v) Third operation is also done in a similar way as it is done in second operation. But this operation is done using the result of second operation.

All operations are done similarly using the result of preceding operation. The n -th operation gives the effect total in case of 2^n -factorial experiment. In the above table, third operation gives the effect total since it is prepared for 2^3 -factorial experiment.

In the last but one column the effects and interactions are written which corresponds to the level combination. The first element of this column is the grand total of the experimental result. The last column gives the sum of squares of different effects and interactions.

The sum of squares of replication, total and error are calculated as usual. The ratio of mean squares of any effects or interactions to the mean square of error gives the respective F -test for that effects or interactions.

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	$F = \frac{s_i}{s_9}$
Replication or block	$r - 1$	S_1	s_1	
<i>A</i>	1	S_2	s_2	
<i>B</i>	1	S_3	s_3	
<i>C</i>	1	S_4	s_4	
Main effects	3	— S_{10}	— s_{10}	
<i>AB</i>	1	S_5	s_5	
<i>AC</i>	1	S_6	s_6	
<i>BC</i>	1	S_7	s_7	
Two-factor interactions	3	— S_{11}	— s_{11}	
<i>ABC</i>	1	S_8	s_8	
Error	$(r - 1)(2^3 - 1)$	S_9	s_9	
Total	$r2^3 - 1$			

The F -statistic to test the significance of i -th effect [$i = 1, 2, \dots, (2^u - 1)$] is $F_i = s_i/s_9$. The conclusion regarding significance of any effect or interaction is made as usual.

In the above 2^3 -factorial experiment, if any one is interested in testing the significance of two-factor interactions, then

$$F = \frac{s_{11}}{s_9}, \text{ where } s_{11} = \frac{S_{11}}{3} \text{ and } S_{10} = S_5 + S_6 + S_7.$$

Similar process is followed in testing the significance of any higher order interaction as a whole.

In factorial experiment the main effects and lower order interactions are of interest. The higher order interactions or highest order interaction are of less interest. For that reason, the sum of squares due to highest order interaction, even if it is significant, is added to error sum of squares and mean square of any main effect is compared with the pooled mean square error. In 2^3 -factorial experiment ABC is the highest order interaction and if the researcher is least interested in this interaction, the $SS(ABC)$ is to be added to SS (error).

Alternative way to calculate sum of squares : Let there be three factors A, B and C each having two levels. Let us denote these levels by 0 and 1. Then the level combinations are 000, 100, 010, 110, 001, 101, 011, 111. The level combinations are for absence of $ABC, A, B, AB, AC, C, AC, BC$ and ABC respectively. Let $x_i (i = 1, 2, 3)$ be the variable for i -th factor, where the values of any factor are 0 and 1. Then the effects and interactions of different factors at two different levels can be represented by linear equations in pairs as follows :

<p>(i) $1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 = 0$ $ = 1 \mid \text{mod } 2$</p>	<p>(ii) $0 \cdot x_1 + 1 \cdot x_2 + 0 \cdot x_3 = 0$ $ = 1 \mid \text{mod } 2$</p>
<p>(iii) $0 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 = 0$ $ = 1 \mid \text{mod } 2$</p>	<p>(iv) $1 \cdot x_1 + 1 \cdot x_2 + 0 \cdot x_3 = 0$ $ = 1 \mid \text{mod } 2$</p>
<p>(v) $1 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 = 0$ $ = 1 \mid \text{mod } 2$</p>	<p>(vi) $0 \cdot x_1 + 1 \cdot x_2 + 1 \cdot x_3 = 0$ $ = 1 \mid \text{mod } 2$</p>
<p>(vii) $1 \cdot x_1 + 1 \cdot x_2 + 1 \cdot x_3 = 0$ $ = 1 \mid \text{mod } 2$</p>	

The solution of these equations are obtained using values of $x_i (i = 1, 2, 3)$ as 0 and 1.

The solutions are derived according to the principle of finite fields under mod. If p is any prime number or power of a prime then for mod p , we have

$$\begin{aligned}
 0 &= p = 2p = \dots = qp \\
 1 &= p + 1 = 2p + 1 = \dots = qp + 1 \\
 2 &= p + 2 = 2p + 2 = \dots = qp + 2 \\
 &\dots\dots\dots \\
 p - 1 &= 2p - 1 = 3p - 1 = \dots = qp + p - 1
 \end{aligned}$$

The equation

$$1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 = 0 \mid 2$$

$$ = 1 \mid 2$$

indicates that A is present at its two levels but B and C are absent. Therefore, solution of this set of equations gives total of A at first level and second level. Therefore,

$$A_0 = 000 + 010 + 001 + 011$$

$$A_1 = 100 + 110 + 101 + 111.$$

Thus, total effect of A , $[A] = A_1 - A_0$. The total effect is calculated from all r replications. Hence, sum of squares of A is

$$SS(A) = \frac{A_0^2 + A_1^2}{2^{2r}} - \text{C.T.}, \text{ C.T.} = \frac{G^2}{r2^3}.$$

The total effect of B is found out using the solutions of second set of equations, and total result of B at two levels are

$$B_0 = 000 + 100 + 101 + 001$$

$$B_1 = 010 + 110 + 010 + 111, [B] = B_1 - B_0$$

$$SS(B) = \frac{1}{2^{2r}}(B_0^2 + B_1^2) - \text{C.T.}$$

Similarly, other solutions and sum of squares are calculated as follows :

$$C_0 = 000 + 110 + 100 + 010 \quad [C] = C_1 - C_0$$

$$C_1 = 001 + 011 + 101 + 111$$

$$SS(C) = \frac{1}{2^{2r}}(C_0^2 + C_1^2) - \text{C.T.}$$

$$(AB)_0 = 000 + 001 + 110 + 111 \quad [AB] = (AB)_0 - (AB)_1$$

$$(AB)_1 = 100 + 010 + 101 + 011$$

$$SS(AB) = \frac{1}{2^{2r}}[(AB)_0^2 + (AB)_1^2] - \text{C.T.}$$

$$(AC)_0 = 000 + 010 + 101 + 111 \quad [AC] = (AC)_0 - (AC)_1$$

$$(AC)_1 = 100 + 001 + 011 + 110$$

$$SS(AC) = \frac{1}{2^{2r}} [(AC)_0^2 + (AC)_1^2] - \text{C.T.}$$

$$(BC)_0 = 000 + 100 + 011 + 111 \quad [BC] = (BC)_0 - (BC)_1$$

$$(BC)_1 = 001 + 101 + 110 + 010$$

$$SS(BC) = \frac{1}{2^{2r}} [(BC)_0^2 + (BC)_1^2] - \text{C.T.}$$

$$(ABC)_0 = 000 + 110 + 101 + 011 \quad [ABC] = (ABC)_1 - (ABC)_0$$

$$(ABC)_1 = 100 + 010 + 001 + 111$$

$$SS(ABC) = \frac{1}{2^{2r}} [(ABC)_0^2 + (ABC)_1^2] - \text{C.T.}$$

It is observed that main effects are obtained subtracting first level total from second level total, two-factor interaction is obtained subtracting second level total from first level total. Thus, in case of odd order effect/interaction, first level total is subtracted from second level total and even order interaction is calculated subtracting second level total from first level total.

The sum of squares due to block (replication) and error sum of squares due to error are calculated as usual.

Example 4.1 : In an agricultural experiment, the productivity of wheat under nitrogen, phosphorus, potash and irrigation are studied. The levels of nitrogen are 0 kg/ha and 90 kg/ha. The first levels of phosphorus and potash are also 0 kg/ha. The second levels of these two fertilizers are 30 kg/ha. The two levels of irrigation are two times irrigation and 4 times irrigation per week. This 2^4 -factorial experiment is conducted through RBD of 16 plots. There are three blocks for the experiment. The plot size is $40' \times 10'$. The production of wheat per plot is recorded for analysis and given below :

Production of wheat (kg/plot) [y_{ijklm}]

Treatment combinations	(1)	<i>n</i>	<i>p</i>	<i>np</i>	<i>k</i>	<i>nk</i>	<i>pk</i>	<i>npk</i>
Block-1	4.2	4.8	3.5	4.0	3.8	4.2	3.2	3.9
Block-2	3.5	4.2	3.0	4.5	3.8	4.6	3.0	4.6
Block-3	3.0	4.6	3.6	4.6	3.6	4.7	4.2	4.8

Production of wheat (kg/plot)

Treatment combinations	<i>i</i>	<i>ni</i>	<i>pi</i>	<i>npi</i>	<i>ki</i>	<i>nki</i>	<i>pki</i>	<i>npki</i>	Total of block R_i
Block-1	4.0	4.5	3.8	4.6	4.5	4.8	4.0	4.6	66.4
Block-2	4.5	4.6	3.9	4.7	4.6	5.0	4.6	4.2	67.3
Block-3	4.6	4.7	4.2	4.6	4.8	5.0	4.8	4.0	69.8
									203.5

Analyse the data and estimate the effects and interaction.

Solution :

Yates' Table to Calculate Sum of Squares

Treatment combinations	Treatment total	Operations				Effects and interactions	$SS = \frac{[\]^2}{3 \times 2^4}$
		1	2	3	4 = []		
(1)	10.7]	24.3]	47.5]	95.9]	203.5	<i>G</i>	862.7552
<i>n</i>	13.6]	23.2]	48.4]	107.6]	14.1	<i>N</i>	4.1419
<i>p</i>	10.1]	24.7]	52.7]	11.1]	-5.7	<i>P</i>	0.6769
<i>np</i>	13.1]	23.7]	54.9]	3.0]	0.5	<i>NP</i>	0.0052
<i>k</i>	11.2]	26.9]	5.9]	-2.1]	3.1	<i>K</i>	0.2002
<i>nk</i>	13.5]	25.8]	5.2]	-3.6]	-3.1	<i>NK</i>	0.2002
<i>pk</i>	10.4]	28.7]	2.7]	0.7]	-1.3	<i>PK</i>	0.0352
<i>npk</i>	13.3]	26.2]	0.3]	-0.2]	-2.3	<i>NPK</i>	0.1102
<i>i</i>	13.1]	2.9]	-1.1]	0.9]	11.7	<i>I</i>	2.8519
<i>ni</i>	13.8]	3.0]	-1.0]	2.2]	-8.1	<i>NI</i>	1.3669
<i>pi</i>	11.9]	2.3]	-1.1]	-0.7]	-1.5	<i>PI</i>	0.0469
<i>npi</i>	13.9]	2.9]	-2.5]	-2.4]	-0.9	<i>NPI</i>	0.0169
<i>ki</i>	13.9]	0.7]	0.1]	0.1]	1.3	<i>KI</i>	0.0352
<i>nki</i>	14.8]	2.0]	0.6]	-1.4]	-1.7	<i>NKI</i>	0.0602
<i>pki</i>	13.4]	0.9]	1.3]	0.5]	-1.5	<i>PKI</i>	0.0469
<i>npki</i>	12.8]	-0.6]	-1.5]	-2.8]	-3.3	<i>NPKI</i>	0.2269

$$SS \text{ (effects and interactions)} = SS(N) + SS(P) + SS(NP) + \dots + SS(NPKI) = 10.0216.$$

$$SS \text{ (Total)} = \sum \sum \sum \sum \sum y_{ijklkm}^2 - \text{C.T.} = 876.47 - 862.7552 = 13.7148.$$

$$SS \text{ (Blocks)} = \sum \frac{R_i^2}{2^4} - \text{C.T.} = \frac{13810.29}{16} - 862.7552 = 0.3879.$$

$$SS \text{ (error)} = SS \text{ (Total)} - SS \text{ (Blocks)} - SS \text{ (Effects and interactions)} \\ = 13.7148 - 0.3879 - 10.0216 = 3.3053.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$
Blocks	2	0.3879	0.19395	1.70	3.32
<i>N</i>	1	4.1419	4.1419	36.36	4.17
<i>P</i>	1	0.6769	0.6769	5.94	"
<i>K</i>	1	0.2002	0.2002	1.76	"
<i>I</i>	1	2.8519	2.8519	25.04	"
<i>NP</i>	1	0.0052	0.0052	0.04	"
<i>NK</i>	1	0.2002	0.2002	1.76	"
<i>PK</i>	1	0.0352	0.0352	0.31	"
<i>NI</i>	1	1.3669	1.3669	12.00	"
<i>PI</i>	1	0.0469	0.0469	0.41	"
<i>KI</i>	1	0.0352	0.0352	0.31	"
Two-factor interactions	— 6	— 1.6844	— 0.2807	— 2.46	— 2.42
<i>NPI</i>	1	0.0169	0.0169	0.15	4.17
<i>NKI</i>	1	0.0602	0.0632	0.53	"
<i>PKI</i>	1	0.0469	0.0469	0.41	"
<i>NPK</i>	1	0.1102	0.1102	0.97	"
Three-factor interactions	— 4	— 0.2342	— 0.0585	— 0.51	— 2.69
Error	31	3.5322	0.1139		
Total	47				

Usually we are least interested in highest order interaction. Now, let us investigate the significance of the interaction *NPKI*. The test statistic to test the significance of this interaction is

$$F = \frac{MS(NPKI)}{MS(\text{error})} = \frac{0.2269}{0.1102} = 2.06 < F_{0.05;1,30} = 4.17.$$

Therefore, the interaction *NPKI* is insignificant. As it is insignificant, the $SS(NPKI)$ is to be added with error sum of squares. The new error sum of squares in the analysis of variance table is

$$SS(\text{error}) + SS(NPKI) = 3.3053 + 0.2269 = 3.5322.$$

The mean square error to be used for F -statistic is $3.5322/31 = 0.1139$.

It is observed that the effects of nitrogen, phosphorus and irrigation are significant, since the F -statistics related to these effects are greater than $F_{0.05}$. The two-factor interactions as a whole are found significant [$F = 2.46 > F_{0.05}$] but no two-factor interaction is found significant. No three-factor interaction is also found significant.

$$P_0 = 0000 + 1000 + 1010 + 1001 + 0010 + 0011 + 0001 + 1011 \\ = 10.7 + 13.0 + 13.5 + 13.8 + 11.2 + 13.9 + 13.1 + 14.8 = 104.6$$

$$P_1 = 0100 + 0101 + 0110 + 0111 + 1100 + 1101 + 1110 + 1111 \\ = 10.1 + 11.9 + 10.4 + 13.4 + 13.1 + 13.9 + 13.3 + 12.8 = 98.9$$

$$[P] = P_1 - P_0 = 98.9 - 104.6 = -5.7$$

$$SS(P) = \frac{1}{r^{2^4-1}}(P_0^2 + P_1^2) - \text{C.T.} = \frac{(104.6)^2 + (98.9)^2}{3 \times 8} - 862.7552 = 0.6769.$$

Similarly, other sum of squares are calculated. The treatment combinations for different effects at two levels are :

$$K_0 = 0000 + 1000 + 0100 + 1100 + 0001 + 1001 + 0101 + 1101$$

$$K_1 = 0010 + 1010 + 0110 + 1110 + 0011 + 1011 + 0111 + 1111$$

$$[K] = K_1 - K_0$$

$$I_0 = 0000 + 1000 + 0100 + 1100 + 0000 + 1010 + 0110 + 1110$$

$$I_1 = 0001 + 1001 + 0101 + 1101 + 0011 + 1011 + 0111 + 1111$$

$$[I] = I_1 - I_0$$

$$(NP)_0 = 0000 + 0010 + 0001 + 0011 + 1100 + 1110 + 1101 + 1111 \\ = 10.7 + 11.2 + 13.1 + 13.9 + 13.1 + 13.3 + 13.9 + 12.8 = 102.0$$

$$(NP)_1 = 1000 + 1001 + 1010 + 1011 + 0100 + 0101 + 0110 + 0111 \\ = 13.6 + 13.8 + 13.5 + 14.8 + 10.1 + 11.9 + 10.4 + 13.4 = 101.5$$

$$[NP] = (NP)_0 - (NP)_1 = 102 - 101.5 = 0.5$$

$$SS(NP) = \frac{1}{r^{2^4-1}}[(NP)_0^2 + (NP)_1^2] - \text{C.T.} = \frac{(102)^2 + (101.5)^2}{3 \times 8} - 862.7552 = 0.0052$$

$$(NK)_0 = 0000 + 1010 + 0001 + 1110 + 1111 + 0100 + 0101 + 1011$$

$$(NK)_1 = 1000 + 0110 + 0111 + 0010 + 1001 + 0011 + 0010 + 0110$$

$$[NK] = (NK)_0 - (NK)_1$$

$$(NI)_0 = 0000 + 1001 + 1011 + 1101 + 0110 + 1111 + 0100 + 0010$$

$$(NI)_1 = 1000 + 1100 + 1010 + 0001 + 0111 + 0001 + 0101 + 0011$$

$$(PK)_0 = 0000 + 0001 + 1000 + 1001 + 0110 + 1110 + 1111 + 0111$$

$$(PK)_1 = 0100 + 1100 + 0010 + 1010 + 1011 + 1101 + 0101 + 0011$$

$$(PI)_0 = 0000 + 1000 + 0010 + 1010 + 0101 + 1111 + 1101 + 0111$$

$$(PI)_1 = 0100 + 1100 + 0110 + 0001 + 1001 + 0001 + 0011 + 1011$$

$$(KI)_0 = 0000 + 1000 + 0100 + 1100 + 0011 + 1011 + 0111 + 1111$$

$$(KI)_1 = 0010 + 1010 + 0110 + 1110 + 0001 + 1001 + 0101 + 1101$$

$$(NPK)_0 = 0000 + 0001 + 1100 + 1101 + 0111 + 0110 + 1010 + 1011 \\ = 10.7 + 13.1 + 13.1 + 13.9 + 13.4 + 10.4 + 13.5 + 14.8 = 102.9$$

$$(NPK)_1 = 1000 + 0100 + 0010 + 1001 + 0101 + 0011 + 1110 + 1111 \\ = 13.6 + 10.1 + 11.2 + 13.8 + 11.9 + 13.9 + 13.3 + 12.8 = 100.6$$

$$\begin{aligned} \therefore [NPK] &= (NPK)_1 - (NPK)_0 = 100.6 - 102.9 = -2.3 \\ (NPI)_0 &= 0000 + 0010 + 1100 + 0101 + 0111 + 1110 + 1011 + 1001 \\ &= 10.7 + 11.2 + 13.1 + 11.9 + 13.4 + 13.3 + 14.8 + 13.8 = 102.2 \\ (NPI)_1 &= 1000 + 0100 + 0001 + 1010 + 0110 + 0011 + 1101 + 1111 \\ &= 13.6 + 10.1 + 13.1 + 13.5 + 10.4 + 13.9 + 13.9 + 12.8 = 101.3 \\ [NPI] &= (NPI)_1 - (NPI)_0 = 101.3 - 102.2 = -0.9 \\ (NKI)_0 &= 0000 + 0100 + 1010 + 1110 + 1001 + 1101 + 0011 + 0111 \\ &= 10.7 + 10.1 + 13.5 + 13.3 + 13.8 + 13.9 + 13.9 + 13.4 = 102.6 \\ (NKI)_1 &= 1000 + 1100 + 0010 + 0110 + 0001 + 0101 + 1011 + 1111 \\ &= 13.6 + 13.1 + 11.2 + 10.4 + 13.1 + 11.9 + 14.8 + 12.8 = 100.9 \\ [NKI] &= (NKI)_1 - (NKI)_0 = 100.9 - 102.6 = -1.7 \\ (PKI)_0 &= 0000 + 1000 + 0110 + 1110 + 0101 + 1101 + 0011 + 1011 \\ &= 10.7 + 13.6 + 10.4 + 13.3 + 11.9 + 13.9 + 13.9 + 14.8 = 102.5 \\ (PKI)_1 &= 0100 + 1100 + 0010 + 1010 + 0001 + 1001 + 0111 + 1111 \\ &= 10.1 + 13.1 + 11.2 + 13.5 + 13.1 + 13.8 + 13.4 + 12.8 = 101.1 \\ [PKI] &= (PKI)_1 - (PKI)_0 = 101.1 - 102.5 = -1.5 \\ [NPKI]_0 &= 0000 + 1100 + 1010 + 1001 + 0110 + 0101 + 0011 + 1111 \\ &= 10.7 + 13.1 + 13.5 + 13.8 + 10.4 + 11.9 + 13.9 + 12.8 = 100.1 \\ (NPKI)_1 &= 1000 + 0100 + 0010 + 0001 + 1011 + 1101 + 1110 + 0111 \\ &= 13.6 + 10.1 + 11.2 + 13.1 + 14.8 + 13.9 + 13.3 + 13.4 = 103.4 \\ [NPKI] &= (NPKI)_0 - (NPKI)_1 = 100.1 - 103.4 = -3.3. \end{aligned}$$

Variance of effects and interactions : Let us consider that a 2^3 -factorial experiment is conducted through randomized block design having r blocks. The factors are A, B and C , say. Then

$$[A] = r(a-1)(b+1)(c+1) = r(a - (1) + ab - b + ac - c + abc - bc)$$

$$A = \frac{[A]}{r2^{3-1}}$$

Here $[A]$ is a linear combination of $r2^3$ observations and observations are independent. Then

$$V[A] = r2^3\sigma^2.$$

Therefore, $V(A) = \frac{V[A]}{r^2(2^{3-1})^2} = \frac{\sigma^2}{r2^{3-2}}.$

Here σ^2 is estimated by mean square error, i.e., $\hat{\sigma}^2 = MS$ (error).

The variance of A is estimated by $V(A) = \frac{\hat{\sigma}^2}{r2^{3-2}}.$

Since all effects and interactions are the linear combination of $r2^3$ observations, the variance of all effects and interactions is given by

$$V(X) = \frac{\sigma^2}{r2^{3-2}},$$

where X indicates a particular effect or interaction.

In a similar way, if 2^n -factorial experiment is conducted through randomized block design having r blocks, the variance of any effect or interaction is

$$V(X) = \frac{\sigma^2}{r2^{n-2}}.$$

Estimate of this variance is $v(x) = \frac{\sigma^2}{r2^{n-2}}$, $\hat{\sigma}^2 = MS$ (error).

This estimated variance is used to estimate the efficiency of the estimate of effect or interaction, where

$$\text{Efficiency} = \frac{1}{v(x)} = \frac{r2^{n-2}}{\hat{\sigma}^2}.$$

For example, the efficiency of effect and interaction estimated in example 4.1, i.e.,

$$\text{Efficiency} = \frac{r2^{4-2}}{\hat{\sigma}^2} = \frac{3 \times 2^{4-2}}{0.1139} = 105.36.$$

4.2 3^n -Factorial Experiment

Let us first consider 3^2 -factorial experiment for factors A and B each having three levels, say 0, 1, and 2. The level combinations are 00, 01, 02, 10, 11, 12, 20, 21, 22. The level combinations can be arranged as follows :

$B \backslash A$	b_0	b_1	b_2	Total
a_0	a_0b_0	a_0b_1	a_0b_2	A_0
a_1	a_1b_0	a_1b_1	a_1b_2	A_1
a_2	a_2b_0	a_2b_1	a_2b_2	A_2
Total	B_0	B_1	B_2	G

or

$B \backslash A$	b_0	b_1	b_2	Total
a_0	00	01	02	A_0
a_1	10	11	12	A_1
a_2	20	21	22	A_2
Total	B_0	B_1	B_2	G

Any experiment in which these 9-treatment combinations are used as treatment, is known as 3^2 -factorial experiment.

There are 9 treatments in the experiment and using these 9 treatments the effects and interactions with 8 d.f. are to be estimated. Since each factor has 3 levels, the effect of each factor has 2 d.f. and hence, the d.f. of interaction of factors is 4. However, each effect and interaction can be decomposed into effects and interactions of 1 d.f. each. The total effects and mean effects with 1 d.f. are shown below :

We have A_0 = total result of the experiment using first level of A ,

A_1 = total result of the experiment using second level of A ,

A_2 = total result of the experiment using third level of A .

Then $A_1 - A_0$ = linear effect of A due to change to second level of A from first level of A .

$$\begin{aligned} &= a_1b_0 + a_1b_1 + a_1b_2 - a_0b_0 - a_0b_1 - a_0b_2 = (a_1 - a_0)(b_0 + b_1 + b_2) \\ &= (-1, 0, 1)(1, 1, 1), \text{ symbolically.} \end{aligned}$$

Here $(-1, 0, 1)$ and $(1, 1, 1)$ are used to indicate the coefficients of linear combination to estimate contrast of the type $A_1 - A_0$. Again, the change in the experimental result due to the change in the levels of A from second level to third level is

$$\begin{aligned} A_2 - A_1 &= a_2b_0 + a_2b_1 + a_2b_2 - a_1b_0 - a_1b_1 - a_1b_2 \\ &= (a_2 - a_1)(b_0 + b_1 + b_2) = (0, -1, 1)(1, 1, 1), \text{ symbolically.} \end{aligned}$$

Here $A_1 - A_0$ and $A_2 - A_1$ are the change in the experimental result due to the linear change of degree one in the levels of A . The linear total effect of A is the sum of these linear changes. Thus,

$$\begin{aligned} [A'] &= A_2 - A_1 + A_1 - A_0 \\ &= (a_2 - a_0)(b_0 + b_1 + b_2) \\ &= (-1, 0, 1)(1, 1, 1). \end{aligned}$$

Here $[A']$ is used to denote total linear effect of A . The average linear effect of A , if the experiment is replicated r times, is given by

$$A' = \frac{1}{r}(a_2 - a_0)(b_0 + b_1 + b_2).$$

The impact of A due to the changes in the levels of A at two degree is measured by

$$[A''] = (A_2 - A_1) - (A_1 - A_0) = (a_2 - 2a_1 + a_0)(b_0 + b_1 + b_2) = (1, -2, 1)(1, 1, 1), \text{ symbolically.}$$

This $[A'']$ is the total quadratic effect of A . The average quadratic effect of A is

$$A'' = \frac{1}{2r}(1, -2, 1)(1, 1, 1), [A''] = (1, -2, 1)(1, 1, 1).$$

It is observed that linear effect of a factor is a linear combination of observations with coefficients $(-1, 0, 1)$ along with coefficients $(1, 1, 1)$ of other factors. The quadratic effect is a linear combination with coefficients $(1, -2, 1)$ along with coefficients $(1, 1, 1)$ of other factors. Thus, we have

$$\begin{aligned} B' &= \frac{1}{r}(1, 1, 1)(-1, 0, 1), [B'] = (1, 1, 1)(-1, 0, 1) \\ B'' &= \frac{1}{2r}(1, 1, 1)(1, -2, 1), [B''] = (1, 1, 1)(1, -2, 1). \end{aligned}$$

The interaction AB has 4 d.f. and this interaction can be decomposed into 4 components having 1 d.f. each. These are :

$$\begin{aligned} A'B' &= \frac{1}{2r}(-1, 0, 1)(-1, 0, 1), A''B' = \frac{1}{4r}(1, -2, 1)(-1, 0, 1), \\ A'B'' &= \frac{1}{4r}(-1, 0, 1)(1, -2, 1), A''B'' = \frac{1}{8r}(1, -2, 1)(1, -2, 1), \\ [A'B'] &= (-1, 0, 1)(-1, 0, 1), [A''B'] = (1, -2, 1)(-1, 0, 1), \\ [A'B''] &= (-1, 0, 1)(1, -2, 1), [A''B''] = (1, -2, 1)(1, -2, 1). \end{aligned}$$

It is observed that the effect total is the linear combination of observations corresponding to different levels of the factors. For example,

$$[B'] = (a_0 + a_1 + a_2)(b_2 - b_0) = a_0b_2 - a_0b_0 + a_1b_2 - a_1b_0 + a_2b_2 - a_2b_0.$$

The coefficients of linear combination corresponding to factor levels are $(-1, 0, 1)$, $(1, -2, 1)$, $(1, 1, 1)$ for linear effect, quadratic effect and for grand total, respectively. For example, the grand total is

$$\begin{aligned} G &= a_0b_0 + a_0b_1 + a_0b_2 + a_1b_0 + a_1b_1 + a_1b_2 + a_2b_0 + a_2b_1 + a_2b_2 \\ &= (1, 1, 1)(1, 1, 1). \end{aligned}$$

The coefficients in the brackets are for first level, second level and third level, respectively. These coefficients are true for any factor levels in 3^n -factorial experiment.

The average effect (main effect) or interaction effect of any factor or factors is obtained dividing the effect total by appropriate divisor. For 3^2 -factorial experiment the appropriate divisor for A' , A'' , are r and $2r$ respectively, where r is the number of replications of the treatment. The divisor for B' and B'' are similar. The divisor for B is used from the convention that any main effect or interaction is estimated from the difference of the result of two plots. This convention is true for interaction and for quadratic effect. The usual divisor for quadratic effect and interaction is half of the products of sum of absolute coefficients for factors. For example, the divisor for $A'B'$ is half of $(1+1)(1+1) = 2$; the divisor for $A''B''$ is half of $(1+2+1)(1+2+1) = 8$. This convention is true for 3^n -factorial experiment and for all effects and interactions when $n > 2$.

3^3 -Factorial experiment : Let there be 3 factors A, B and C each having three levels, say, 0, 1 and 2. The level combinations are :

000	010	020	100	110	120	200	210	220
001	011	021	101	111	121	201	211	221
002	012	022	102	112	122	202	212	222

Here first notation is for A , second one for B and third notation is for C . If these 27 treatment combinations are used as treatments in any experiment, the experiment is known as 3^3 -factorial experiment.

As there are 27 treatments, we can estimate contrasts of effects and interactions of 26 d.f. $(27 - 1)$. These effects are linear and quadratic. The interactions are :

linear \times linear, linear \times quadratic, quadratic \times linear,
 quadratic \times quadratic, linear \times linear \times linear,
 linear \times linear \times quadratic, linear \times quadratic \times quadratic,
 linear \times quadratic \times linear, quadratic \times linear \times linear,
 quadratic \times linear \times quadratic, quadratic \times quadratic \times linear

and quadratic \times quadratic \times quadratic. The contrasts are :

$$A' = \frac{1}{9r}(-1, 0, 1)(1, 1, 1)(1, 1, 1) = \frac{1}{9r}(a_2 - a_0)(b_0 + b_1 + b_2)(c_0 + c_1 + c_2)$$

$$A'' = \frac{1}{18r}(-1, 2, 1)(1, 1, 1)(1, 1, 1)$$

$$B' = \frac{1}{9r}(1, 1, 1)(-1, 0, 1)(1, 1, 1), \quad B'' = \frac{1}{18r}(1, 1, 1)(1, -2, 1)(1, 1, 1)$$

$$C' = \frac{1}{9r}(1, 1, 1)(1, 1, 1)(-1, 0, 1), \quad C'' = \frac{1}{18r}(1, 1, 1)(1, 1, 1)(1, -2, 1)$$

$$A'B' = \frac{1}{6r}(-1, 0, 1)(-1, 0, 1)(1, 1, 1), \quad A'B'' = \frac{1}{12r}(-1, 0, 1)(1, -2, 1)(1, 1, 1)$$

$$A'C' = \frac{1}{6r}(-1, 0, 1)(1, 1, 1)(-1, 0, 1), \quad A'C'' = \frac{1}{12r}(-1, 0, 1)(1, 1, 1)(1, -2, 1)$$

$$A''B'' = \frac{1}{24r}(1, -2, 1)(-1, -2, 1)(1, 1, 1), \quad A''C'' = \frac{1}{24r}(1, -2, 1)(1, 1, 1)(1, -2, 1)$$

$$A''B' = \frac{1}{12r}(1, -2, 1)(-1, 0, 1)(1, 1, 1), \quad A''C' = \frac{1}{12r}(1, -2, 1)(1, 1, 1)(-1, 0, 1)$$

$$B'C' = \frac{1}{6r}(1, 1, 1)(-1, 0, 1)(-1, 0, 1), \quad B'C'' = \frac{1}{12r}(1, 1, 1)(-1, 0, 1)(1, -2, 1)$$

$$\begin{aligned}
 B''C' &= \frac{1}{12r}(1, 1, 1)(1, -2, 1)(-1, 0, 1), & B''C'' &= \frac{1}{24r}(1, 1, 1)(1, -2, 1)(1, -2, 1) \\
 A'B'C' &= \frac{1}{4r}(-1, 0, 1)(-1, 0, 1)(-1, 0, 1), & A'B'C'' &= \frac{1}{8r}(-1, 0, 1)(-1, 0, 1)(1, -2, 1) \\
 A'B''C' &= \frac{1}{8r}(-1, 0, 1)(1, -2, 1)(-1, 0, 1), & A''B'C' &= \frac{1}{8r}(1, -2, 1)(-1, 0, 1)(-1, 0, 1) \\
 A'B''C'' &= \frac{1}{16r}(-1, 0, 1)(1, -2, 1)(1, -2, 1), & A''B'C'' &= \frac{1}{16r}(1, -2, 1)(-1, 0, 1)(1, -2, 1) \\
 A''B''C' &= \frac{1}{16r}(1, -2, 1)(1, -2, 1)(-1, 0, 1), & A''B''C'' &= \frac{1}{32r}(1, -2, 1)(1, -2, 1)(1, -2, 1).
 \end{aligned}$$

Calculation of sum of squares : Let us describe the procedure for 3²-factorial experiment. There are 8 contrasts. The total of each contrast is a linear combination of the result of different treatment combinations. Thus,

$$[A'] = (-1, 0, 1)(1, 1, 1) = (a_2 - a_0)(b_1 + b_1 + b_2)$$

Here $SS(A') = \frac{[A']^2}{6r}$, where r is the number of replications of a treatment and 6 is the product of sum of squares of the coefficients for different levels of factors. For example, $[(-1)^2 + 0^2 + (1)^2][1^2 + 1^2 + 1^2] = 6$. This rule is true for all contrasts. Thus

$$\begin{aligned}
 SS(A'') &= \frac{[A'']^2}{18r}, & SS(B') &= \frac{[B']^2}{6r}, & SS(B'') &= \frac{[B'']^2}{18r}, \\
 SS(A'B') &= \frac{[A'B']^2}{4r}, & SS(A'B'') &= \frac{[A'B'']^2}{12r}, & SS(A''B') &= \frac{[A''B']^2}{12r}, \\
 SS(A''B'') &= \frac{[A''B'']^2}{36r}.
 \end{aligned}$$

The sum of squares are calculated using Yate's algorithm. This algorithm is described below :

Yates' Table to Calculate Sum of Squares in Case of 3²-Factorial Experiment

Treatment Combinations	Total result of treatment	Operation		Effects and Interacting	Divisor D_i	$SS = \frac{(\quad)^2}{rD_i}$
		1	2 = []			
00	x_1	$x_1 + x_2 + x_3 = y_1$	$y_1 + y_2 + y_3$	G	9	
10	x_2	$x_4 + x_5 + x_6 = y_2$	$y_4 + y_5 + y_6$	A'	6	
20	x_3	$x_7 + x_8 + x_9 = y_3$	$y_7 + y_8 + y_9$	A''	18	
01	x_4	$x_3 - x_1 = y_4$	$y_3 - y_1$	B'	6	
11	x_5	$x_6 - x_4 = y_5$	$y_6 - y_4$	$A'B'$	4	
21	x_6	$x_9 - x_7 = y_6$	$y_9 - y_7$	$A''B'$	12	
02	x_7	$x_1 - 2x_2 + x_3 = y_7$	$y_1 - 2y_2 + y_3$	B''	18	
12	x_8	$x_4 - 2x_5 + x_6 = y_8$	$y_6 - 2y_5 + y_4$	$A'B''$	12	
22	x_9	$x_7 - 2x_8 + x_9 = y_9$	$y_9 - 2y_8 + y_7$	$A''B''$	36	

First column is for treatment combinations in a systematic manner arranged in first, second and third levels of first, second, and third factors and so on. Second column gives the total result of a treatment from all replication. The operations of calculating total effect is done with

the result of total of treatments. The first one-third observations of first operation are obtained by linear combinations of treatment total with coefficients (1, 1, 1), where total treatments are arranged in groups of 8, and linear combination is done for each group. The second one-third observations of first operation is obtained by linear combinations of each group of 3 treatment totals with coefficients (-1, 0, 1) and the last one-third observations of first operation is obtained by linear combination with coefficient (1, -2, 1). The second operation is similarly done as it is done in first operation but the results of second operation is obtained from the results of first operation. The operation is continued up to n -th time. Each time the results are obtained from the results of the preceding operation. The n -th operation gives the total of contrast, where contrasts are systematically written in a column. The system of writing effects and interactions starts with linear and then quadratic and then the product of linear and quadratic contrasts. The first result of n -th operation gives grand total (G). The sum of squares is the square of contrast total divided by appropriate divisor.

Each factor has 3 levels and a factor sum of squares has 2 d.f. The sum of squares of 8 d.f. can be decomposed into 4 components each of 2 d.f. To calculate sum of squares of 2 d.f., the 3 totals of each effect and interaction are found using the solution of equations under mod 3. The 4 components of effects and interactions each with 2 d.f. are A, B, AB and AB_2 . The total of these effects and interactions are obtained using the solutions of the following equations respectively.

$$\begin{array}{l} x_1 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_2 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_1 + x_2 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_1 + 2x_2 = 2 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3 \\ \\ \end{array} \right.$$

Thus, we have

$$\begin{aligned} A_0 &= 00 + 01 + 02, & B_0 &= 00 + 10 + 20, & (AB)_0 &= 00 + 12 + 21, \\ A_1 &= 10 + 11 + 12, & B_1 &= 01 + 11 + 21, & (AB)_1 &= 10 + 01 + 22, \\ A_2 &= 20 + 21 + 22, & B_2 &= 02 + 12 + 22, & (AB)_2 &= 20 + 02 + 11, \\ (AB_2)_0 &= 00 + 11 + 22, & (AB_2)_1 &= 10 + 02 + 21, \\ (AB_2)_2 &= 20 + 12 + 01. \end{aligned}$$

$$\text{Then } SS(A) = \frac{A_0^2 + A_1^2 + A_2^2}{3^{2-1}r} - \text{C.T.}, \text{ where C.T.} = \frac{G^2}{r3^2}$$

$$SS(B) = \frac{1}{r3^{2-1}}(B_0^2 + B_1^2 + B_2^2) - \text{C.T.},$$

$$SS(AB) = \frac{1}{r3^{2-1}}[(AB)_0^2 + (AB)_1^2 + (AB)_2^2] - \text{C.T.}$$

$$SS(AB_2) = \frac{1}{r3^{2-1}}[(AB_2)_0^2 + (AB_2)_1^2 + (AB_2)_2^2] - \text{C.T.}$$

Other sum of squares are calculated as usual.

The effects and interactions having 26 d.f. are estimated from 3^3 -factorial experiment. These effects and interactions can be decomposed into 13 components each of 2 d.f. The totals to calculate sum of squares of these 13 components are found out using the solutions of the following 13 sets of equations.

$$\begin{array}{l} x_1 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_2 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_1 + x_2 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right.$$

$$\begin{array}{l} x_1 + 2x_2 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_1 + x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_1 + 2x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right.$$

$$\begin{array}{l} x_2 + x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_2 + 2x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right.$$

$$\begin{array}{l} x_1 + x_2 + x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_1 + x_2 + 2x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right. \quad \begin{array}{l} x_1 + 2x_2 + x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3, \\ \\ \end{array} \right.$$

$$\begin{array}{l} x_1 + 2x_2 + 2x_3 = 0 \\ = 1 \\ = 2 \end{array} \left| \begin{array}{l} \text{mod } 3. \\ \\ \end{array} \right.$$

These equations are giving solutions for totals of $A, B, C, AB, AB_2, AC, AC_2, BC, BC_2, ABC, ABC_2, AB_2C$ and AB_2C_2 . The totals using solutions are shown below :

$$A_0 = 000 + 010 + 001 + 011 + 012 + 021 + 002 + 020 + 022$$

$$A_1 = 100 + 110 + 101 + 111 + 112 + 121 + 102 + 112 + 122$$

$$A_2 = 200 + 210 + 201 + 211 + 212 + 221 + 202 + 220 + 222$$

$$B_0 = 000 + 100 + 200 + 001 + 101 + 201 + 002 + 102 + 202$$

$$B_1 = 010 + 110 + 210 + 011 + 111 + 211 + 012 + 112 + 212$$

$$B_2 = 020 + 120 + 220 + 021 + 121 + 221 + 022 + 122 + 222$$

$$C_0 = 000 + 100 + 200 + 010 + 110 + 200 + 210 + 120 + 010$$

$$C_1 = 001 + 101 + 201 + 011 + 111 + 201 + 211 + 121 + 011$$

$$C_2 = 002 + 102 + 202 + 012 + 112 + 202 + 212 + 122 + 012$$

$$(AB)_0 = 000 + 120 + 210 + 001 + 121 + 211 + 002 + 122 + 212$$

$$(AB)_1 = 100 + 101 + 102 + 010 + 011 + 012 + 220 + 221 + 222$$

$$(AB)_2 = 200 + 201 + 202 + 020 + 021 + 022 + 110 + 111 + 112$$

$$(AC)_0 = 000 + 010 + 020 + 102 + 112 + 122 + 201 + 211 + 221$$

$$(AC)_1 = 100 + 110 + 120 + 001 + 011 + 021 + 202 + 212 + 222$$

$$(AC)_2 = 200 + 210 + 220 + 002 + 012 + 022 + 101 + 111 + 121$$

$$(AB_2)_0 = 000 + 001 + 002 + 110 + 111 + 112 + 220 + 221 + 222$$

$$(AB_2)_1 = 100 + 101 + 102 + 020 + 021 + 022 + 210 + 211 + 212$$

$$(AB_2)_2 = 200 + 201 + 202 + 010 + 011 + 012 + 120 + 121 + 122$$

$$(AC_2)_0 = 000 + 010 + 020 + 101 + 111 + 121 + 202 + 212 + 222$$

$$(AC_2)_1 = 100 + 110 + 120 + 002 + 012 + 022 + 201 + 211 + 221$$

$$(AC_2)_2 = 200 + 210 + 220 + 001 + 011 + 021 + 102 + 112 + 122$$

$$(BC)_0 = 000 + 100 + 200 + 012 + 112 + 212 + 021 + 121 + 221$$

$$(BC)_1 = 010 + 110 + 210 + 001 + 101 + 201 + 022 + 122 + 222$$

$$\begin{aligned}
 (BC)_2 &= 020 + 120 + 220 + 002 + 102 + 202 + 011 + 111 + 211 \\
 (BC_2)_0 &= 000 + 100 + 200 + 011 + 111 + 211 + 022 + 122 + 222 \\
 (BC_2)_1 &= 010 + 110 + 210 + 002 + 102 + 202 + 021 + 121 + 221 \\
 (BC_2)_2 &= 020 + 120 + 220 + 012 + 112 + 212 + 001 + 101 + 201 \\
 (ABC)_0 &= 000 + 120 + 210 + 012 + 021 + 102 + 201 + 222 + 111 \\
 (ABC)_1 &= 100 + 010 + 001 + 220 + 202 + 022 + 112 + 121 + 211 \\
 (ABC)_2 &= 200 + 020 + 002 + 110 + 101 + 011 + 221 + 212 + 122 \\
 (ABC_2)_0 &= 000 + 011 + 101 + 221 + 112 + 022 + 210 + 120 + 202 \\
 (ABC_2)_1 &= 100 + 010 + 220 + 111 + 002 + 201 + 021 + 212 + 122 \\
 (ABC_2)_2 &= 200 + 020 + 110 + 102 + 012 + 222 + 211 + 121 + 001 \\
 (AB_2C)_0 &= 000 + 011 + 110 + 212 + 121 + 022 + 201 + 220 + 102 \\
 (AB_2C)_1 &= 001 + 012 + 020 + 100 + 111 + 122 + 202 + 210 + 221 \\
 (AB_2C)_2 &= 002 + 010 + 021 + 101 + 112 + 120 + 200 + 211 + 222 \\
 (AB_2C_2)_0 &= 000 + 012 + 021 + 101 + 110 + 122 + 202 + 211 + 220 \\
 (AB_2C_2)_1 &= 002 + 011 + 020 + 100 + 112 + 121 + 201 + 210 + 222 \\
 (AB_2C_2)_2 &= 001 + 010 + 022 + 102 + 111 + 120 + 200 + 212 + 221.
 \end{aligned}$$

The sums of squares are calculated as follows :

$$\begin{aligned}
 SS(A) &= \frac{1}{r3^{3-1}}(A_0^2 + A_1^2 + A_2^2) - \text{C.T.}, \quad SS(B) = \frac{1}{r3^{3-1}}(B_0^2 + B_1^2 + B_2^2) - \text{C.T.} \\
 SS(C) &= \frac{1}{r3^{3-1}}(C_0^2 + C_1^2 + C_2^2) - \text{C.T.}, \quad SS(AB) = \frac{1}{r3^{3-1}}[(AB)_0^2 + (AB)_1^2 + (AB)_2^2] - \text{C.T.} \\
 \dots\dots\dots \\
 SS(ABC) &= \frac{1}{r3^{3-1}}[(ABC)_0^2 + (ABC)_1^2 + (ABC)_2^2] - \text{C.T.}
 \end{aligned}$$

The sums of squares of *A*, *B*, *C*, *AB*, *AC*, *BC* and *ABC* having 2, 2, 2, 4, 4, 4 and 8 d.f., respectively are calculated in a similar way as it is done in three-way classification with *r* observations per cell. Here each of *A*, *B* and *C* has 3 levels. For example, the sum of squares of *AB* of 4 d.f., is calculated from a two-way table as follows :

<i>B</i> <i>A</i>	<i>b</i> ₀	<i>b</i> ₁	<i>b</i> ₂	Total
<i>a</i> ₀	000 + 001 + 002 = <i>y</i> ₁	010 + 011 + 012 = <i>y</i> ₄	020 + 021 + 022 = <i>y</i> ₇	<i>A</i> ₀
<i>a</i> ₁	100 + 101 + 102 = <i>y</i> ₂	110 + 111 + 112 = <i>y</i> ₅	120 + 121 + 122 = <i>y</i> ₈	<i>A</i> ₁
<i>a</i> ₂	200 + 201 + 202 = <i>y</i> ₃	210 + 211 + 212 = <i>y</i> ₆	220 + 221 + 222 = <i>y</i> ₉	<i>A</i> ₂
Total	<i>B</i> ₀	<i>B</i> ₁	<i>B</i> ₂	<i>G</i>

$$SS(AB) = \frac{1}{3r}(y_1^2 + y_2^2 + y_3^2 + \dots + y_9^2) - \text{C.T.} - SS(A) - SS(B).$$

For 3^{*n*}-factorial experiment the effects and interactions having (3^{*n*} - 1) d.f. are estimated. The linear effect of a factor is the linear combination of the levels of that factor with coefficients

(-1, 0, 1) in presence of all other factors. The quadratic effect is the linear combination with coefficients (1, -2, 1) in presence of all other factors. The interaction is also calculated in a similar way as it is done in 3³-factorial experiment.

The total to calculate sum of square having 2 d.f. are calculated using the solution of the equations

$$\begin{matrix} ix_1 + ix_2 + ix_3 + \dots + ix_n = 0 \\ = 1 \\ = 2 \end{matrix} \Bigg| \text{mod } 3, i = 0, 1, 2.$$

Example 4.2 : To study the productivity of balsam apple under nitrogen and phosphorus fertilizers a 3²-factorial experiment is conducted in an agricultural research station. The levels of nitrogen are $n_0 = 40$ kg/ha, $n_1 = 80$ kg/ha and $n_2 = 120$ kg/ha. The levels of phosphorus are $p_0 = 30$ kg/ha, $p_1 = 60$ kg/ha and $p_2 = 90$ kg/ha. Each level combination is applied in 20' × 20' plot. The experiment is replicated in 3 blocks. The productions of balsam apple (kg/plot) are shown below :

Production (kg/plot) of balsam apple in different blocks

Treatment Combinations	Blocks			Total production
	1	2	3	
00	14.6	15.2	15.0	44.8
10	18.2	17.5	18.0	53.7
20	28.0	27.5	26.2	81.7
01	15.2	15.0	16.4	46.6
11	20.6	19.2	20.0	59.8
21	29.4	29.0	30.2	88.6
02	16.4	18.6	17.6	52.6
12	21.6	22.4	24.2	68.2
22	32.2	31.8	33.1	97.1
Total B_i	196.2	196.2	200.7	593.1

Analyse the data and comment on the impacts of fertilizer.

Solution :

Yates' Table to Calculate Sum of Squares in Case of 3²-Factorial Experiment

Treatment combinations	Total of treatment	Operation		Effects and interactions	Divisor D_i	$SS = \frac{[\]^2}{rD_i}$ $r = 3$
		1	2 = []			
00	44.8	180.2	593.1	G	9	13028.43
10	53.7	195.0	123.4	N'	6	845.976
20	81.7	217.9	48.0	N''	18	42.667
01	46.6	36.9	37.7	P'	6	78.961
11	59.8	42.0	7.6	$N'P'$	4	4.813
21	88.6	44.5	-5.8	$N''P'$	12	0.934
02	52.6	19.1	8.1	P''	18	1.215
12	68.2	15.6	-2.6	$N'P''$	12	0.188
22	97.1	13.3	1.2	$N''P''$	36	0.013

$$C.T. = 13028.43, SS(\text{Total}) = \sum \sum \sum y_{ijl}^2 - C.T. = 14015.11 - 13028.43 = 986.68.$$

$$SS(\text{Block}) = \frac{\sum B_i^2}{3^2} - C.T. = \frac{117269.37}{9} - 13028.43 = 1.50.$$

$$SS(\text{error}) = SS(\text{Total}) - SS(\text{Block}) - SS(\text{Effects and interactions}) \\ = 986.68 - 1.50 - 974.767 = 10.413.$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F	P-value
Block	2	1.50	0.75	1.15	> 0.05
N'	1	845.976	845.976	1299.90	0.00
N''	1	42.667	42.667	65.56	0.00
N	— 2	— 888.643	— 444.333	— 682.75	0.00
P'	1	78.961	78.961	121.33	0.00
P''	1	1.215	1.215	1.87	> 0.05
P	— 2	— 80.176	— 40.088	— 61.60	0.00
N'P'	1	4.813	4.813	7.39	0.00
N'P''	1	0.188	0.188	0.29	> 0.05
N''P'	1	0.934	0.934	1.43	> 0.05
N''P''	1	0.013	0.013	0.02	> 0.05
NP	— 4	— 5.948	— 1.487	— 2.28	> 0.05
Error	16	10.413	0.6508	—	
Total	26				

It is observed that both linear and quadratic effects of both the factors are highly significant. The joint effect of nitrogen and phosphorus are not found significant. However, one of the components of interaction of nitrogen and phosphorus, viz., linear \times linear interaction is found highly significant [P -value 0.00 or $F > F_{0.01}$].

The sum of squares of effects and interactions of 2 d.f. are calculated as follows :

$$N_0 = 00 + 01 + 02 = 44.8 + 46.6 + 92.6 = 144.0.$$

$$N_1 = 10 + 11 + 12 = 53.7 + 59.8 + 68.2 = 181.7.$$

$$N_2 = 20 + 21 + 22 = 81.7 + 88.6 + 97.1 = 267.4.$$

$$SS(N) = \frac{1}{3 \times r} (A_0^2 + A_1^2 + A_2^2) - C.T. = \frac{125253.65}{9} - 13028.43 = 888.642.$$

$$P_0 = 00 + 10 + 20 = 44.8 + 53.7 + 81.7 = 180.2.$$

$$P_1 = 01 + 11 + 21 = 46.6 + 59.8 + 88.6 = 195.0.$$

$$P_2 = 02 + 12 + 22 = 52.6 + 68.2 + 97.1 = 217.9.$$

$$SS(P) = \frac{1}{3 \times r} (B_0^2 + B_1^2 + B_2^2) - C.T. = \frac{117977.45}{9} - 13028.43 = 80.176.$$

$$(NP)_0 = 00 + 12 + 21 = 44.8 + 68.2 + 88.6 = 201.6.$$

$$(NP)_1 = 01 + 10 + 22 = 46.6 + 53.7 + 97.1 = 197.4.$$

$$(NP)_2 = 11 + 20 + 02 = 59.8 + 81.7 + 52.6 = 194.1.$$

$$SS(NP) = \frac{1}{3r} [(NP)_0^2 + (NP)_1^2 + (NP)_2^2] - C.T. = \frac{117284.13}{9} - 13028.43 = 3.14.$$

$$(NP_2)_0 = 00 + 11 + 22 = 44.8 + 59.8 + 97.1 = 201.7.$$

$$(NP_2)_1 = 10 + 02 + 21 = 53.7 + 52.6 + 88.6 = 194.9.$$

$$(NP_2)_2 = 20 + 01 + 12 = 81.7 + 46.6 + 68.2 = 196.5.$$

$$SS(NP_2) = \frac{1}{3r} [(NP_2)_0^2 + (NP_2)_1^2 + (NP_2)_2^2] - C.T. = \frac{117281.15}{9} - 13028.43 = 2.809.$$

4.3 4ⁿ-Factorial Experiment

Let there be n factors A, B, C, \dots, N ; each has 4 levels. The levels are usually denoted by 0, 1, 2, 3. The total level combinations are 4^n and if all these level combinations are used as treatments in any experiment, then the experiment is called 4^n -factorial experiment. From this experiment the effects and interactions of $(4^n - 1)$ d.f. are estimated. These effects and interactions can be partitioned into $(4^n - 1)$ components each of one d.f. These are linear, quadratic, cubic effects and the interactions of these effects.

However, since 4 is not a prime number, the interactions cannot be partitioned into components of $(4 - 1) = 3$ d.f. Thus, the total of effects and interactions are not found out using the equations of the type :

$$\left. \begin{aligned} ix_1 + ix_2 + ix_3 + ix_4 &= 0 \\ &= 1 \\ &= 2 \\ &= 3 \end{aligned} \right\} \text{mod } 4, i = 0, 1, 2, 3.$$

Let us explain the analysis for 4^2 -factorial experiment. The factors are assumed to be A and B . The total level combinations are 16 and these are shown in the table below.

$B \backslash A$	b_0	b_1	b_2	b_3	Total of A
a_0	00	01	02	03	A_0
a_1	10	11	12	13	A_1
a_2	20	21	22	23	A_2
a_3	30	31	32	33	A_3
Total of B	B_0	B_1	B_2	B_3	G

From, this 4^2 -factorial experiment one can estimate 15 components of effects and interactions each of 1 d.f. The effects are linear, quadratic and cubic. The two-factor interactions are linear \times linear, linear \times quadratic, linear \times cubic, quadratic \times cubic. The totals of effects and interactions are the linear combinations of treatments. The coefficients for linear effect are $(-3, -1, 1, 3)$ for first, second, third and fourth levels of a factor, respectively. This coefficients are true for any factor. Thus, the total for linear effect of A is obtained by

$$[A_l] = (-3, -1, 1, 3)(1, 1, 1, 1)$$

Similarly, $[B_l] = (1, 1, 1, 1)(-3, -1, 1, 3)$.

The coefficients for quadratic and cubic effects are $(1, -1, -1, 1)$ and $(-1, 3, -3, 1)$ respectively. Therefore, the totals for quadratic effect of A and cubic effect of A are given by

$$\begin{aligned}[A_q] &= (1, -1, -1, 1)(1, 1, 1, 1) \\ [A_c] &= (-1, 3, -3, 1)(1, 1, 1, 1).\end{aligned}$$

Similarly, the total for quadratic and cubic effects of B are :

$$\begin{aligned}[B_q] &= (1, 1, 1, 1)(1, -1, -1, 1) \\ [B_c] &= (1, 1, 1, 1)(-1, 3, -3, 1).\end{aligned}$$

The total for interactions are :

$$\begin{aligned}[A_l B_l] &= (-3, -1, 1, 3)(-3, -1, 1, 3), [A_l B_q] = (-3, -1, 1, 3)(1, 1, 1, 1), \\ [A_l B_c] &= (-3, -1, 1, 3)(-1, 3, -3, 1), [A_q B_l] = (1, -1, -1, 1)(-3, -1, 1, 3), \\ [A_q B_c] &= (1, -1, -1, 1)(-1, 3, -3, 1), [A_q B_q] = (1, -1, -1, 1)(1, -1, -1, 1), \\ [A_c B_l] &= (-1, 3, -3, 1)(-3, -1, 1, 3), [A_c B_q] = (-1, 3, -3, 1)(1, -1, -1, 1), \\ [A_c B_c] &= (-1, 3, -3, 1)(-1, 3, -3, 1).\end{aligned}$$

The sum of squares of the effects and interactions are calculated by

$$SS(Z_i) = \frac{[Z_i]^2}{rD_i},$$

where D_i is the appropriate divisor for i -th effect of interaction. The value of D_i is the product of the sum of squares of coefficients for each factor level. Thus, the D_i value for $A_l (i = l)$ is $[(-3)^2 + (-1)^2 + (1)^2 + (3)^2][1^2 + 1^2 + 1^2 + 1^2] = 80$. D_i for A_q is

$$[1^2 + (-1)^2 + (-1)^2 + 1^2][1^2 + 1^2 + 1^2] = 16.$$

The sum of squares can be calculated using Yate's algorithm. The algorithm is explained below :

- (i) The first column of Yate's table shows the treatment combinations in alphabetical order and in increasing order of levels.
- (ii) The second column is for the total of each treatment from different replications.
- (iii) The treatment totals are grouped. In each group there are 4 totals. The operational columns give the results from linear combinations of each group of totals.

The first one-fourth of the first operation is a linear combination with coefficients $(1, 1, 1, 1)$. The second one-fourth of the first operation is linear combination with coefficients $(-3, -1, 1, 3)$ of each group. The third one-fourth observation of first operation is calculated in a similar way with coefficients $(1, -1, -1, 1)$. The last one-fourth observation of the first operation is calculated in a similar way with coefficients $(-1, 3, -3, 1)$.

- (iv) The subsequent operations are done in a similar way using the results of preceding operation. The n -th operation gives the total for effects and interactions. The first element of the n -th operation gives grand total (G). The other values are for the systematic linear, quadratic, cubic and the corresponding interactions in a systematic fashion. The calculation is shown in the following table.

Yates' Table to Calculate Total for Effects and Interactions

Treatment Combinations	Treatment total	Operation		Effects and interactions	Appropriate divisor D_i
		1	2 = []		
00	x_1	$x_1 + x_2 + x_3 + x_4 = y_1$	$y_1 + y_2 + y_3 + y_4$	G	16
10	x_2	$x_5 + x_6 + x_7 + x_8 = y_2$	$y_5 + y_6 + y_7 + y_8$	A_t	80
20	x_3	$x_9 + x_{10} + x_{11} + x_{12} = y_3$	$y_9 + y_{10} + y_{11} + y_{12}$	A_q	16
30	x_4	$x_{13} + x_{14} + x_{15} + x_{16} = y_4$	$y_{13} + y_{14} + y_{15} + y_{16}$	A_c	80
01	x_5	$3x_4 + x_3 - x_2 - 3x_1 = y_5$	$3y_4 + y_3 - y_2 - 3y_1$	B_t	80
11	x_6	$3x_8 + x_7 - x_6 - 3x_5 = y_6$	$3y_8 + y_7 - y_6 - 3y_5$	$A_t B_t$	400
21	x_7	$3x_{12} + x_{11} - x_{10} - 3x_9 = y_7$	$3y_{12} + y_{11} - y_{10} - 3y_9$	$A_q B_t$	80
31	x_8	$3x_{16} + x_{15} - x_{14} - 3x_{13} = y_8$	$3y_{16} + y_{15} - y_{14} - 3y_{13}$	$A_c B_t$	400
02	x_9	$x_4 - x_3 - x_2 + x_1 = y_9$	$y_4 - y_3 - y_2 + y_1$	B_q	16
12	x_{10}	$x_8 - x_7 - x_6 + x_5 = y_{10}$	$y_8 - y_7 - y_6 + y_5$	$A_t B_q$	80
22	x_{11}	$x_{12} - x_{11} - x_{10} + x_9 = y_{11}$	$y_{12} - y_{11} - y_{10} + y_9$	$A_q B_q$	16
32	x_{12}	$x_{16} - x_{15} - x_{14} + x_{13} = y_{12}$	$y_{16} - y_{15} - y_{14} + y_{13}$	$A_c B_q$	80
03	x_{13}	$x_4 - 3x_3 + 3x_2 - x_1 = y_{13}$	$y_4 - 3y_3 + 3y_2 - y_1$	B_c	80
13	x_{14}	$x_8 - 3x_7 + 3x_6 - x_5 = y_{14}$	$y_8 - 3y_7 + 3y_6 - y_5$	$A_t B_c$	400
23	x_{15}	$x_{12} - 3x_{11} + 3x_{10} - x_9 = y_{15}$	$y_{12} - 3y_{11} + 3y_{10} - y_9$	$A_q B_c$	80
33	x_{16}	$x_{16} - 3x_{15} + 3x_{14} - x_{13} = y_{16}$	$y_{16} - 3y_{15} + 3y_{14} - y_{13}$	$A_c B_c$	400

Example 4.3 : An experiment is conducted in a laboratory to study the mortality level of wood louse using two chemicals, viz., Abate (A) and Benlate (B). Four levels, viz., 0 ppm, 500 ppm, 1000 ppm and 1500 ppm of each chemical are used in the experiment. The experiment is conducted through randomized block design using 2 blocks of 16 plots each. At the beginning of the experiment 30 wood lice are kept at each level combination of chemicals. After 7 days the dead wood lice are counted. The numbers of dead wood lice are shown in the following table.

Treatment combinations		00	01	02	03	10	11	12	13	20	21	22	23	30	31	32	33
Dead wood lice in blocks	1	2	15	16	18	20	22	22	24	23	24	26	23	28	27	28	28
	2	1	17	16	17	19	21	21	23	23	22	25	21	27	26	26	29
Total y_{ij}		3	32	32	35	39	43	43	47	46	46	51	44	55	53	54	57

Analyse the data and comment on the effects of pesticides.

Solution : The observations can be arranged in a 2×2 table showing the observations (y_{ij}) at each level of pesticides as below :

$B \backslash A$	b_0	b_1	b_2	b_3	Total of A_i
a_0	3	32	32	35	102
a_1	39	43	43	47	172
a_2	46	46	51	44	187
a_3	55	53	54	57	219
Total B_i	143	174	180	183	680

$$C.T. = \frac{G^2}{r4^2} = \frac{(680)^2}{2 \times 16} = 14450.$$

$$SS \text{ (Total)} = \sum \sum \sum y_{ijl}^2 - C.T. = 15752 - 14450 = 1302.$$

$$SS(B) = \frac{1}{r4} \sum B_i^2 - C.T. = \frac{116614}{2 \times 4} - 14450 = 126.75$$

$$SS(A) = \frac{1}{r4} \sum A_i^2 - C.T. = \frac{122918}{2 \times 4} - 14460 = 914.75$$

$$SS(AB) = \frac{\sum \sum y_{ij}^2}{r} - C.T. - SS(B) - SS(A) = \frac{31478}{2} - 14450 - 126.75 - 914.75 = 247.50$$

$$SS \text{ (Block)} = \frac{\sum bl_i^2}{4^2} - C.T. = \frac{231272}{16} - 14450 = 4.50,$$

where totals of blocks are $bl_1 = 346$, $bl_2 = 334$.

$$SS \text{ (error)} = SS \text{ (Total)} - SS \text{ (Block)} - SS(B) - SS(A) - SS(AB) = 1302 - 4.5 - 126.75 - 914.75 - 247.56 = 8.5.$$

The sum of squares effects and interactions with 1 d.f. are calculated using Yate's algorithm as follows :

Treatment Combinations	Total of treatment	Operation		Effects and interactions	Divisor D_i	$SS = \frac{[]}{rD_i}$
		1	2 = []			
00	3	143	680	G	16	14450
10	39	174	366	A_l	80	837.225
20	46	180	-38	A_q	16	45.125
30	55	183	72	A_c	80	32.400
01	32	163	126	B_l	80	99.225
11	43	66	-292	$A_l B_l$	400	106.58
21	46	74	80	$A_q B_l$	80	40.00
31	53	63	-14	$A_c B_l$	400	0.245
02	32	-27	-28	B_q	16	24.50
12	43	-4	86	$A_l B_q$	80	46.225
22	51	-8	-14	$A_q B_q$	16	6.125
32	54	1	52	$A_c B_q$	80	16.90
03	35	31	22	B_c	80	3.025
13	47	12	-124	$A_l B_c$	400	19.22
23	44	-2	40	$A_q B_c$	80	10.00
33	57	31	42	$A_c B_c$	400	2.205

It is observed that $SS(A_l) + SS(A_q) + SS(A_c) = S(A)$ which is calculated previously. Similar are the cases for $SS(B)$ and $SS(AB)$. The results are shown in analysis of variance table.

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	P-value
Block	1	4.50	4.50	7.94	< 0.05
A_l	1	837.225	837.225	1477.37	0.00
A_q	1	45.125	45.125	79.63	0.00
A_c	1	32.400	32.400	57.17	0.00
A	— 3	— 914.75	— 304.92	— 538.06	0.00
B_l	1	99.225	99.225	175.09	0.00
B_q	1	24.50	24.50	13.88	0.00
B_c	1	3.025	3.025	5.34	< 0.05
B	— 3	126.75	— 42.25	— 74.55	0.00
$A_l B_l$	1	106.58	106.58	188.07	0.00
$A_l B_q$	1	46.225	46.225	81.57	0.00
$A_l B_c$	1	19.22	19.22	33.92	0.00
$A_q B_l$	1	40.00	40.008	70.58	0.00
$A_q B_q$	1	6.125	6.1258	10.81	< 0.01
$A_q B_c$	1	10.00	10.00	17.65	0.00
$A_c B_l$	1	0.245	0.245	0.43	> 0.05
$A_c B_q$	1	16.90	16.90	29.82	0.00
$A_c B_c$	1	2.205	2.205	3.89	> 0.05
AB	— 9	— 247.50	— 27.5	— 48.53	0.00
Error	15	8.5	0.5667	—	—
Total	31				

The main effects of Abate and Benlate are highly significant. The linear changes, the quadratic changes or the cubic changes in the levels of pesticides are significantly effective to kill wood louse. The joint effect of both the pesticides are also significant.

4.4 p^n -Factorial Experiment

Let there be n factors A, B, C, D, \dots, N each of which has p levels. The levels are, say, $0, 1, 2, \dots, (p - 1)$. The total level combinations are p^n . In an experiment if all these p^n level combinations are used as treatment, then the experiment is called p^n -factorial experiment.

From this experiment $(p^n - 1)$ effects and interactions are estimated. These effects and interactions can be divided into $(p^n - 1)/p - 1$ components. Each of these components has $(p - 1)$ d.f. provided p is a prime number. There are p total for each effect and interaction. These p totals are the totals of p^{n-1} result of treatments. The treatments for any total is identified by the solutions of the equations of the type :

$$\begin{array}{l}
 ix_1 + ix_2 + ix_3 + \dots + ix_n = 0 \\
 = 1 \\
 = 2 \\
 = \vdots \\
 = p - 1
 \end{array}
 \left| \begin{array}{l}
 \dots \\
 \dots \\
 \dots \\
 \dots \\
 \dots
 \end{array} \right.
 \begin{array}{l}
 \dots \\
 \dots \\
 \dots \\
 \dots \\
 \dots
 \end{array}
 \text{mod } p; i = 0, 1, 2, \dots, (p - 1).$$

Here $x_j (j = 1, 2, \dots, n)$ is used to indicate j -th factor. For example, if $n = 2$ and $p = 5$, then there are two factors A and B . Therefore, we have 5 totals of A , where these totals are the sum of results of treatments, the treatments are identified from the solution of the equations :

$$\begin{array}{l} x_1 = 0 \\ \quad = 1 \\ \quad = 2 \\ \quad = 3 \\ \quad = 4 \end{array} \left| \begin{array}{l} \\ \\ \\ \\ \end{array} \right. \text{mod } 5$$

Therefore, we have

$$\begin{aligned} [A_0] &= 00 + 01 + 02 + 03 + 04, & [A_1] &= 10 + 11 + 12 + 13 + 14 \\ [A_2] &= 20 + 21 + 22 + 23 + 24, & [A_3] &= 30 + 31 + 32 + 33 + 34 \\ [A_4] &= 40 + 41 + 42 + 43 + 44. \end{aligned}$$

The sum of squares of A is given by

$$SS(A) = \frac{1}{r5^{2-1}} \sum A_i^2 - \text{C.T.}, \quad \text{C.T.} = \frac{G^2}{r5^2}.$$

In general, if X_i is the total of any effect or interaction for p^n -factorial experiment, the sum of squares of that effect or interaction is

$$SS (\text{Effect or interaction}) = \frac{1}{rp^{n-1}} \sum_{i=1}^{p-1} X_i^2 - \text{C.T.}$$

Here r is the number of replication of the experiment. The sum of squares of replication (block) is calculated as usual.

For $n = 2$ and $p = 5$, the effects and interactions can be divided into 6 components each of 4 d.f. The components are $A, B, AB_1, AB_2, AB_3, AB_4$. The totals of these effects are calculated from the sum of the result of those treatments and identified from the solution of the equations.

$$\begin{aligned} x_1 &= i(i = 0, 1, 2, 3, 4) \text{ mod } 5; & x_2 &= i(i = 0, 1, \dots, 4) \text{ mod } 5 \\ x_1 + x_2 &= i(i = 0, 1, \dots, 4) \text{ mod } 5, & x_1 + 2x_2 &= i(i = 0, 1, \dots, 4) \text{ mod } 5 \\ x_1 + 3x_2 &= i(i = 0, 1, \dots, 4) \text{ mod } 5, & x_1 + 4x_2 &= i(i = 0, 1, \dots, 4) \text{ mod } 5 \text{ respectively.} \end{aligned}$$

The $(5^2 - 1)$ effects and interactions can be partitioned into 24 components each of 1 d.f. These are linear, quadratic, cubic, quartic and the interactions of these effects. Each component is a linear combination of treatments, where the coefficients of different levels of factors are shown in the table below :

Effects	Coefficients for different levels				
	0	1	2	3	4
Linear	-2	-1	0	1	2
Quadratic	2	-1	-2	-1	2
Cubic	-1	2	0	-2	1
Quartic	1	-4	6	-4	1

For example, the total of linear effect of A is

$$A_l = (-2, -1, 0, 1, 2)(1, 1, 1, 1, 1).$$

The sum of squares of these 24 components can be calculated using Yate's algorithm, where the operation will be done to get the linear combination according to the above mentioned coefficients.

4.5 Generalized Interaction

For 2^n -factorial experiment, if the factors are A, B, C, \dots, N , then the interaction of A and B is AB , A and C is AC , A, B and C is ABC and so on. The interaction of A and BC is ABC . The interaction of A and AC or C and AC can also be determined. This is done under mod 2. Thus, the interaction of A and AC is $A \times AC = A^2C = C$ under mod 2. Here, under mod 2, $A^2 = A^0 = 1$. Similarly,

$$A \times ABC = A^2BC = BC \mid \text{mod } 2$$

$$AB \times BC = AB^2C = AC \mid \text{mod } 2$$

$$A \times BC = ABC.$$

Similar types of interactions can be identified under module 3 in case of 3^n -factorial experiment. Thus,

$$A \times AB = A_2B = A_4B_2 = AB_2 \mid \text{mod } 3$$

$$AB \times AB_2 = A_2B_3 = A_4B_6 = A \mid \text{mod } 3$$

$$AC \times AC_2 = A_2C_3 = A_4C_6 = A \mid \text{mod } 3$$

$$ABC \times BC = AB_2C_2$$

$$ABC_2 \times AC_2 = A_2BC_4 = A_4B_2C_8 = AB_2C_2 \mid \text{mod } 3$$

Here C, BC, AC and ABC for 2^n -factorial experiment and AB_2, A, AB_2C_2 , for 3^n -factorial experiment are known as generalized interaction.

The generalized interactions are very much important in confounded factorial experiment and in fractional replication of factorial experiment.

4.6 Confounded Factorial Experiment

In factorial experiment the treatments are the level combinations of factors and usually the experiment is conducted through randomized block design. If factors or levels or both are large, the level combinations become large which needs blocks of larger number of plots. Due to the use of blocks of large number of plots, there is a chance of heterogeneity in the plots within a block. The heterogeneity among the plots within a block may distort the objective of blocking. This problem can be avoided if blocks of smaller number of plots are used in the experiment. In that case all treatments are not allocated to the plots of a block, rather a portion of treatments are allocated in plots of a block so that the block contrast represents a higher order interaction of factors. Thus, in any replication more than one block is used in the experiment. Usually, for p^n -factorial experiment blocks of multiple of p are used per replication, and treatments are allocated in plots within a block so that block contrasts represent one or more higher order interactions. This method of allocation of treatments to the plots within blocks per replication is known as confounding and the experiment is called confounded factorial experiment.

In this technique the blocks are incomplete and the incomplete blocks are smaller in number of plots. The blocks with smaller number of plots are expected to be more homogeneous. As a result block homogeneity increases and the efficiency of the experiment increases by reducing the experimental error.

Since higher order interaction and block contrast are entangled due to the use of confounded factorial experiment, the interaction which is less important to the researcher is usually confounded. The information on confounded interaction is lost but unconfounded effects and interactions are estimated with more efficiency. However, the confounded technique is used in such a way that no main effect is confounded with blocks.

It is already mentioned that, in confounded factorial experiment the number of blocks per replication are more than one. Usually, for 2^n -factorial experiment the number of blocks per replication is 2^{n-k} ($k < n$) and for 3^n -factorial experiment the number of blocks is 3^{n-k} ($k < n$). For p^n -factorial experiment the number of blocks is p^{n-k} and there are p^k plots per block. Due to the use of p^{n-k} blocks $(p^{n-k} - 1)/(p - 1)$ interactions or components of interactions are confounded with blocks.

It has been discussed that the total of treatment contrast is calculated from the treatments results, where treatments are decided solving the equations of the types :

$$\begin{array}{l|l} ix_1 + ix_2 + \dots + ix_n = 0 \\ = 1 \\ = 2 \\ \vdots \\ = p - 1 \end{array} \quad \text{mod } p; \quad i = 0, 1, 2, \dots, (p - 1).$$

The treatment contrast is the linear combination of the total of the contrast calculated from different solutions of the above types of equations. Thus, for 2^3 -factorial experiment ABC is the highest order interaction and the equations to get the total of ABC at two levels are :

$$\begin{array}{l|l} x_1 + x_2 + x_3 = 0 \\ = 1 \end{array} \quad \text{mod } 2$$

so that we have

$$\begin{aligned} (ABC)_0 &= 000 + 011 + 110 + 101 \\ (ABC)_1 &= 100 + 010 + 001 + 111. \end{aligned}$$

Therefore, the total of ABC contrast is

$$\begin{aligned} [ABC] &= (ABC)_1 - (ABC)_0 = (111 + 100 + 010 + 001) - (000 + 011 + 110 + 101) \\ &= (abc + a + b + c) - ((1) + bc + ab + ac) \end{aligned}$$

Now, if the treatments to get $(ABC)_1$ are allocated in Block-1 and the treatments to get $(ABC)_0$ are allocated in Block-2 within a replication, we get

$$\begin{aligned} [ABC] &= \text{total result of Block-1} - \text{total result of Block-2} \\ &= B_1 - B_2, \text{ where } B_i = \text{total of Block-}i, \quad i = 1, 2. \end{aligned}$$

That is

Total of blocks						
Replication	Block-1	abc	a	b	c	B_1
	Block-2	(1)	ab	ac	bc	B_2

It is observed that the block contrast $[B_1 - B_2]$ is equal to the treatment contrast ABC . Therefore, the ABC interaction is confounded with blocks of a replication and the information of this interaction with 1 d.f. is lost.

Again, for 3^2 -factorial experiment the totals of treatment contrast AB_2 are calculated from the sum of results of treatments which are obtained from the solution of the equations :

$$\begin{array}{l} x_1 + 2x_2 = 0 \\ = 1 \\ = 2 \end{array} \Bigg| \text{ mod } 3.$$

Thus, we have $(AB_2)_0 = 00 + 11 + 22$, $(AB_2)_1 = 10 + 21 + 02$, $(AB_2)_2 = 20 + 01 + 12$.
Now, the treatments are allocated in blocks of 3 plots as under :

					Total of blocks
Replication	Block-1	00	11	22	B_1
	Block-2	10	21	02	B_2
	Block-3	20	01	12	B_3

Here block contrast may be of the type $(B_3 - B_1)$ or $(B_1 - 2B_2 + B_3)$. Both the contrasts are equivalent to the contrast of AB_2 . Therefore, AB_2 interaction is confounded with blocks and information of this interaction with 2 d.f. is lost.

So far we have discussed the confounding of one component of contrast of treatment in one replication. In practice, more than one component of treatment contrasts can be confounded with blocks in a replication. For example, let us consider the allocation of treatments of 2^4 -factorial experiment in blocks of 4 plots each, where

$$\begin{aligned} (ABCD)_0 &= 0000 + 1100 + 1010 + 1001 + 0011 + 0101 + 0110 + 1111 \\ &= (1) + ab + ac + ad + cd + bd + bc + abcd \\ (ABCD)_1 &= 1000 + 0100 + 0010 + 0001 + 1101 + 1110 + 0111 + 1011 \\ &= a + b + c + d + abd + abc + bcd + acd. \end{aligned}$$

These totals of $(ABCD)_0$ and $(ABCD)_1$ are calculated using the result of treatments, where treatments are identified from the solution of the equations :

$$\begin{array}{l} x_1 + x_2 + x_3 + x_4 = 0 \\ = 1 \end{array} \Bigg| \text{ mod } 2.$$

Again, for AB interaction, we have

$$\begin{array}{l} x_1 + x_2 = 0 \\ = 1 \end{array} \Bigg| 2.$$

$$\begin{aligned} (AB)_0 &= 0000 + 1100 + 0011 + 0010 + 0001 + 1111 + 1101 + 1110 \\ &= (1) + ab + cd + c + d + abcd + abd + abc. \\ (AB)_1 &= 1000 + 0100 + 1011 + 0111 + 1010 + 0110 + 1001 + 0101 \\ &= a + b + acd + bcd + ac + bc + ad + bd. \end{aligned}$$

It is observed that half of treatments for $(AB)_0$ are used for $(ABCD)_0$ and another half are used for $(ABCD)_1$. Similar is the case with $(AB)_1$. Now, one can allocate 2^{4-2} treatments in blocks within a replication as follows :

Replication	Blocks	Treatments in plots				Total
1	1	(1)	ab	abcd	cd	B_1
	2	ac	ad	bd	bc	B_2
	3	a	b	acd	bcd	B_3
	4	abc	abd	c	d	B_4

It is observed that

$$[ABCD] = B_1 + B_2 - B_3 - B_4, [AB] = B_1 + B_4 - B_2 - B_3, \\ [CD] = B_1 + B_3 - B_2 - B_4.$$

All three interactions are block contrasts. These are confounded with blocks. Three interactions each with 1 d.f. are confounded with blocks. Here CD is automatically confounded with blocks, where CD is the generalized interaction of $ABCD$ and AB [$AB \times ABCD = CD \mid \text{mod } 2$].

In 2^n -factorial experiment half of treatment contrasts for an interaction at first level is also used for treatment contrast of another interaction at first level. In other words, half of treatments with negative sign for a treatment are with positive signs for another treatment contrast and half of treatments with positive sign for a treatment contrast are with positive signs for another treatment contrast. Thus, if positive treatments of both contrasts and negative treatments of both contrasts are allocated randomly in two separate blocks and the remaining negative and positive treatments are allocated separately in another two blocks, then two treatment contrasts (interactions) and their generalized interaction is confounded with blocks. In such a situation the block size is reduced to one-fourth and block homogeneity is expected more and the efficiency of the experiment is expected to be increased.

Two or more interactions can be confounded in p^n -factorial experiment also reducing the block size to p^{n-k} . For example, let us consider the simultaneous confounding of AB and AC each of 2 d.f. in a replication in case of 3^3 -factorial experiment. We have

$$(AB)_0 = 000 + 120 + 210 + 001 + 121 + 211 + 002 + 122 + 212 \\ (AB)_1 = 100 + 101 + 102 + 220 + 221 + 222 + 010 + 011 + 012 \\ (AB)_2 = 200 + 201 + 202 + 020 + 021 + 022 + 110 + 111 + 112 \\ (AC)_0 = 000 + 010 + 020 + 102 + 112 + 122 + 201 + 211 + 221 \\ (AC)_1 = 100 + 110 + 120 + 001 + 011 + 021 + 202 + 212 + 222 \\ (AC)_2 = 200 + 210 + 220 + 002 + 012 + 022 + 101 + 111 + 121.$$

The treatments are allocated as follows :

Replication	Blocks	Treatments in blocks			Block total
1	1	000	211	122	B_1
	2	120	001	212	B_2
	3	210	121	002	B_3
	4	100	222	011	B_4
	5	110	021	202	B_5
	6	101	220	012	B_6
	7	102	221	010	B_7
	8	200	022	111	B_8
	9	201	020	112	B_9

It is observed that

$$(AB)_0 = B_1 + B_2 + B_3, (AB)_1 = B_4 + B_6 + B_7, (AB)_2 = B_8 + B_9 + B_5, \\ (AC)_0 = B_1 + B_7 + B_9, (AC)_1 = B_4 + B_5 + B_2, (AC)_2 = B_3 + B_6 + B_8.$$

Since the treatments required to calculate $(AB)_0$ or $(AB)_1$ or $(AB)_2$ are allocated in 3 different blocks, the interaction AB is confounded with blocks. Similar is the case with AC . The generalized interaction of AB and AC is $AB \cdot AC = A_2BC = A_4B_2C_2 = AB_2C_2 \pmod 3$. Here, we have

$$(AB_2C_2)_0 = B_1 + B_6 + B_5, (AB_2C_2)_1 = B_3 + B_4 + B_9, (AB_2C_2)_2 = B_2 + B_7 + B_8.$$

Therefore, AB_2C_2 interaction is also confounded with blocks. In total, 3 interactions each of 2 d.f. are confounded with blocks in the above mentioned replication.

In using this confounded technique the block size is reduced further. The plots in a block are $\frac{1}{3^2} \times 3^3 = 3$. Again, if we use blocks of 9 plots allocating treatments of B_1, B_5 and B_6 in a block, treatments of B_3, B_4 and B_9 in another block and the treatments of B_2, B_7 and B_8 in another block, then 2 d.f. of AB_2C_2 will be confounded with block.

We have discussed the technique of confounding by allocating treatments to blocks of smaller number of plots in one replication. In practice, there are many replications.

One or many interactions can be confounded in all replications or in some replications. Accordingly, the technique of confounding is of two types, viz., total confounding and partial confounding.

Total confounding : If an interaction is confounded in all replications of an experiment, the confounded technique is known as total confounding. For example, let us consider 2^4 -factorial experiment conducted through randomized block design replicated three times using two blocks per replication. The treatments in all three replications are as follows :

Arrangement of treatments of 2^4 -factorial experiments in plots of blocks within replications

Replication-1		Replication-2		Replication-3	
Block-1	Block-2	Block-1	Block-2	Block-1	Block-2
(1)	a	ab	abc	abcd	abd
ab	b	ac	acd	ab	bcd
ac	c	(1)	bcd	ac	acd
bc	d	bc	a	(1)	a
ad	abc	ad	b	ad	b
bd	abd	bd	c	bd	c
cd	acd	cd	d	cd	d
abcd	bcd	abcd	abd	bc	abc
Total B_{11}	B_{12}	B_{21}	B_{22}	B_{31}	B_{32}

Here $[ABCD] = B_{11} - B_{12}$ in replication-1.

$[ABCD] = B_{21} - B_{22}$ in replication-2

and $[ABCD] = B_{31} - B_{32}$ in replication-3.

In all replications $ABCD$ interaction is confounded with blocks. $ABCD$ interaction is totally confounded. The information of $ABCD$ is totally lost.

Partial confounding : If any interaction is confounded with blocks in some replications but not in all, the interaction is partially confounded. The technique of allocating treatments in blocks of a replication so that block contrasts represent an interaction in some replications

but not in all is called *partial confounding*. For example, let us consider the allocation of treatments in blocks of replications as follows :

Replications	Blocks	Treatments in plots								Total of blocks
1	1	(1)	ab	cd	abcd	abc	abd	d	c	B_{11}
	2		a	b	ac	bc	ad	bd	acd	bcd
2	1	(1)	cd	ab	abcd	acd	bcd	a	b	B_{21}
	2		c	d	abc	abd	ac	bc	bd	ad

Here $[AB] = B_{11} - B_{12}$ and $[CD] = B_{21} - B_{22}$.

Thus, interaction AB is confounded with blocks in replication-1, and interaction CD is confounded with blocks in replication-2. Both the interactions are partially confounded. Information of AB can be estimated from replication-2 and information of CD can be estimated from replication-1. Here $[AB]$ from replication = $B_{11} - B_{12}$.

$[AB]_1 = [AB]$ from unconfounded replication = $[AB] - [AB]$ from replication-1

$$\therefore SS(AB) \text{ from unconfounded blocks} = \frac{[AB]_1^2}{(r-1)2^3}$$

The partial confounded interaction sum of squares is calculated in a similar way as above. The sum of squares from all replications are calculated as usual by Yate's method.

The effect total of totally confounded interaction is also confounded as usual by Yate's method. But its sum of squares is not calculated. The sum of squares of totally confounded interaction is included with error sum of squares.

For p^n -factorial experiment no interaction is totally confounded with blocks. Only $(p-1)$ d.f. of an interaction is confounded with blocks of one replication. The $(p-1)$ d.f. of an interaction is confounded with blocks within one replication but the remaining $(p-2)$ components each of $(p-1)$ d.f. are not confounded in that replication and hence, the sum of squares of these remaining components are calculated from the replication. For example, let us consider the confounding of AB interaction of 3^2 -factorial experiment, where arrangements of treatments in plots are as follows :

Replications	Blocks	Treatments in plots			Total of blocks
1	1	00	12	21	B_{11}
	2	10	01	22	B_{12}
	3	20	02	11	B_{13}
2	1	00	11	22	B_{21}
	2	10	02	21	B_{22}
	3	01	20	12	B_{23}

Here $(AB)_0 = B_{11}$, $(AB)_1 = B_{12}$ and $(AB)_2 = B_{13}$; $(AB_2)_0 = B_{21}$, $(AB_2)_1 = B_{22}$ and $(AB_2)_2 = B_{23}$. Therefore, AB with 2 d.f. is confounded with blocks within replication-1. The interaction AB_2 with 2 d.f. is confounded with blocks within replication-2. The interaction AB can be estimated from replication-2 and AB_2 can be estimated from replication-1 from such experiment of two replications. The interaction AB has 4 d.f., but the information of all 4 d.f. are not lost from an experiment.

The sum of squares of effects and interactions are calculated as usual. The sum of squares of totally confounded interactions in case of 2^n -factorial experiment are not calculated. The sum of squares of partially confounded interactions are calculated from those replications in which these are not confounded. The procedure is shown above. Other steps of analysis remain same as these are used in analysing unconfounded data of unconfounded factorial experiment. The d.f. of totally confounded interaction is added with the d.f. of error, but d.f. of partially confounded interaction is shown with a separate sign in the analysis of variance table. Let us now discuss some examples of totally and partially confounded factorial experiment.

Example 4.4 : An experiment is conducted to study the productivity of maize using two doses of nitrogen as urea, two doses of phosphorus, two doses of potash and two levels of cow dung. The doses of nitrogen (N) are 60 kg/ha and 90 kg/ha, doses of phosphorus (P) are 40 kg/ha and 80 kg/ha, doses of potash (K) are 30 kg/ha and 60 kg/ha and the levels of cow dung are 2 m.ton/ha and 3 m.ton/ha. The level combinations are used as treatments and treatments are applied in blocks of 4 plots per replication. The experiment is conducted through randomized block design and treatments are replicated 3 times. The productions of maize in 1 m² plot are recorded for analysis.

Production of Maize (in kg/plot, y_{ijklkm} km)

Replication	Blocks	Production against treatment per plot				Total of blocks B_{ij}
1	1	(1) - 2.5	$np - 2.8$	$npkd - 3.2$	$kd - 2.2$	10.7
	2	$nk - 3.0$	$nd - 2.7$	$pd - 2.0$	$pk - 2.2$	9.9
	3	$n - 2.7$	$p - 2.6$	$nk d - 3.4$	$pk d - 2.7$	11.4
	4	$npk - 3.0$	$npd - 2.8$	$k - 2.6$	$d - 2.5$	10.9
2	1	(1) - 2.6	$np - 3.0$	$npkd - 3.6$	$kd - 2.6$	11.8
	2	$nk - 3.4$	$nd - 2.6$	$pd - 2.2$	$pk - 2.6$	10.8
	3	$n - 2.6$	$p - 2.2$	$nk d - 3.0$	$pk d - 2.8$	10.6
	4	$npk - 3.6$	$npd - 3.4$	$k - 2.8$	$d - 2.0$	11.8
3	1	(1) - 2.4	$np - 3.0$	$npkd - 3.0$	$kd - 2.6$	11.0
	2	$nk - 3.6$	$nd - 2.9$	$pd - 2.4$	$pk - 2.0$	10.9
	3	$n - 3.0$	$p - 2.6$	$nk d - 2.8$	$pk d - 3.0$	11.4
	4	$npk - 3.4$	$npd - 3.6$	$k - 2.6$	$d - 2.4$	12.0

Analyse the data and comment on the use of different fertilizers.

Solution : This is a 2^4 -factorial experiment, where 3 interactions are confounded with blocks. Number of replications $r = 3$. It is observed that,

$$[NP] = B_{11} + B_{14} - B_{12} - B_{13}, [KD] = B_{11} + B_{13} - B_{12} - B_{14}$$

and $[NPKD] = B_{11} + B_{12} - B_{13} - B_{14}$.

Therefore, NP, KD and $NPKD$ interactions are confounded with blocks. The same treatment arrangements are noted in all three replications. Hence, all three interactions are totally confounded with blocks.

SS (Blocks within replications)

$$= \sum_{i=1}^3 \left[\frac{\sum B_{ij}^2}{4} - \frac{(\sum B_{ij})^2}{4^2} \right].$$

Here $\sum B_{1j} = 42.9$, $\sum B_{2j} = 45.0$, $\sum B_{3j} = 45.3$

$$\begin{aligned} &= \left[\frac{1}{4} \{ (10.7)^2 + (9.9)^2 + (11.4)^2 + (10.9)^2 \} - \frac{(42.9)^2}{16} \right] \\ &+ \left[\frac{1}{4} \{ (11.8)^2 + (10.8)^2 + (10.6)^2 + (11.8)^2 \} - \frac{(45.0)^2}{16} \right] \\ &+ \left[\frac{1}{4} \{ (11.0)^2 + (10.9)^2 + (11.4)^2 + (12.0)^2 \} - \frac{(45.3)^2}{16} \right] \\ &= (115.3175 - 115.0256) + (126.87 - 126.562) + (128.4425 - 128.2556) = 0.7863. \end{aligned}$$

Yates' Table to Calculate Sum of Squares

Treatment combinations	Total of treatment	Operations				Effects and interactions	$SS = \frac{[]^2}{r \cdot 2^4}$ $r = 3$
		1	2	3	4 = []		
(1)	7.5]	15.8]	32.0]	66.8]	133.2	G	369.63 = C.T.
(n)	8.3]	16.2]	34.8]	66.4]	15.0	N	4.6875
p	7.4]	18.0]	31.5]	7.4]	2.2	P	0.1008
np	8.8]	16.8]	34.9]	7.6]	3.2	NP	—
k	8.0]	15.1]	2.2]	-0.8]	6.2	K	0.8008
nk	10.0]	16.4]	5.2]	-3.0]	1.6	NK	0.0533
pk	6.8]	16.6]	4.5]	1.8]	-1.2	PK	0.03
npk	10.0]	18.3]	3.1]	1.4]	-1.8	NPK	0.0675
d	6.9]	0.8]	0.4]	2.8]	-0.4	D	0.0033
nd	8.2]	1.4]	-1.2]	3.4]	0.2	ND	0.0008
pd	6.6]	2.0]	1.3]	3.0]	3.8	PD	0.3008
npd	9.8]	3.2]	1.7]	-1.4]	-0.4	NPD	0.0033
kd	7.4]	1.3]	0.6]	-1.6]	0.6	KD	—
$nk d$	9.2]	3.2]	1.2]	0.4]	-4.4	NKD	0.4033
$pk d$	8.5]	1.8]	1.9]	0.6]	2.0	PKD	0.0833
$npk d$	9.8]	1.3]	-0.5]	-2.4]	-3.0	$NPKD$	—

$$SS \text{ (unconfounded effects and interactions)} = SS(N) + SS(P) + \dots$$

$$+ SS(NKD) + SS(PKD) = 6.5347.$$

$$SS \text{ (Total)} = \sum \sum \sum \sum \sum y_{ijklm}^2 - \text{C.T.} = 378.54 - 369.63 = 8.91.$$

$$SS \text{ (Error)} = SS \text{ (Total)} - SS \text{ (Blocks within replications)}$$

$$- SS \text{ (Unconfounded effects and interactions)}$$

$$= 8.91 - 0.7863 - 6.5347 = 1.589.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$	P-value
Blocks within-replication	9	0.7863	0.0874	1.43	2.30	> 0.05
N	1	4.6875	4.6875	76.72	4.22	0.00
P	1	0.1008	0.1008	1.65	"	> 0.05
K	1	0.8008	0.8008	13.11	"	0.00
D	1	0.0033	0.0033	0.05	"	> 0.05
Main effects	— 4	5.5924	1.3981	22.88	2.74	0.00
NK	1	0.0533	0.0533	0.87	4.22	> 0.05
PK	1	0.03	0.03	0.49	"	"
ND	1	0.0008	0.0008	0.01	"	"
PD	1	0.3008	0.3008	4.92	"	< 0.05
Two-factor interaction	— 4	0.3849	0.0962	1.57	2.74	> 0.05
NPK	1	0.0675	0.0675	1.10	4.22	"
NPD	1	0.0033	0.0033	0.05	"	"
NKD	1	0.4033	0.4033	6.60	"	< 0.05
PKD	1	0.0833	0.0833	1.36	"	> 0.05
Three-factor interactions	— 4	0.5574	0.1393	2.28	2.74	> 0.05
Error	26	1.589	0.0611	—		
Total	47					

The *F*-statistics for which *P*-values are greater than 0.05 indicate insignificant effects or interactions. The main effects are found highly significant, but only the effects of nitrogen and potash are highly significant. The overall two-factor interactions are found insignificant but the interaction of phosphorus and cow-dung are significant. The interaction of nitrogen, potash and cow-dung is found significant though overall three-factor interactions are found insignificant.

The main effects are :

$$N = \frac{[N]}{r2^{4-1}} = \frac{15.0}{3 \times 8} = 0.625, P = \frac{[P]}{r2^{4-1}} = \frac{2.2}{3 \times 8} = 0.092,$$

$$K = \frac{[K]}{r2^{4-1}} = \frac{6.2}{3 \times 8} = 0.258, D = \frac{[D]}{r2^{4-1}} = \frac{-0.4}{24} = -0.017.$$

The estimated variance of any one of the effects or interactions is :

$$V(X) = \frac{MS(\text{error})}{r2^{n-2}} = \frac{0.0611}{3 \times 2^{4-2}} = 0.005,$$

where *X* is effect or interaction.

Example 4.5 : To study the mortality capacity of Abate (*A*) and Benlate (*B*) an experiment is conducted using these pesticides at two levels on wood lice. The levels of concentration of these pesticides used in the experiment are 500 ppm and 1000 ppm. Thirty wood lice are kept under these pesticides at room temperature and at temperature 30 °C. This 2³-factorial experiment is conducted through randomized block design of 4 plots each. The experiment is

repeated three times. After 7 days of the start of the experiment the number of dead wood lice are counted. The data are given below.

Number of dead wood lice (y_{ijk}) under different treatment

Replications	Blocks	Treatment and dead wood lice				Total of blocks B_{ij}
1	1	(1) - 18	at - 20	abt - 28	b - 25	91 = B_{11}
	2	a - 22	t - 20	bt - 24	ab - 27	93 = B_{12}
2	1	(1) - 17	bt - 22	abt - 27	a - 24	90 = B_{21}
	2	b - 24	t - 21	at - 22	ab - 26	93 = B_{22}
3	1	a - 24	b - 25	t - 21	abt - 28	98 = B_{31}
	2	ab - 24	at - 23	bt - 24	(1) - 15	86 = B_{32}

Analyse the data and interpret the result of using pesticides.

Solution : This is a 2^3 -factorial experiment where three interactions, viz., AT , BT and ABT are partially confounded with blocks. The number of replications are $r = 3$. Here

$$[AT] = B_{11} - B_{12}, [BT] = B_{21} - B_{22}, [ABT] = B_{31} - B_{32}.$$

Thus, AT is confounded partially in replication-1, BT is confounded partially in replication-2 and ABT is partially confounded in replication-3. The sum of squares of AT , BT and ABT are to be calculated, respectively from replication-2 and replication-3; replication-1 and replication-3; replication-1 and replication-2. Here

$$\sum B_{1j} = 184, \sum B_{2j} = 183, \sum B_{3j} = 184.$$

Yates' Table to calculate sum of squares

Treatment combinations	Total of treatment	Operation			Effects and interactions	$SS = \frac{\square^2}{r^2}$ $r = 3$
		1	2	3 = []		
(1)	50]	120]	271]	551	G	12650.042
a	70]	151]	280]	39	A	63.375
b	74]	127]	23]	57	B	135.375
ab	77]	153]	16]	-7	AB	2.042
t	62]	20]	31]	9	T	3.375
at	65]	3]	26]	-7	AT	—
bt	70]	3]	-17]	-5	BT	—
abt	83]	13]	10]	27	ABT	—

$$\begin{aligned}
 SS \text{ (blocks within replications)} &= \sum_i \left[\frac{\sum_i B_{ij}^2}{4} - \frac{(\sum B_{ij})^2}{2^3} \right] \\
 &= SS \text{ (blocks within replication-1)} + SS \text{ (blocks within replication-2)} \\
 &\quad + SS \text{ (blocks within replication-3)} \\
 &= \left[\frac{91^2 + 93^2}{4} - \frac{(184)^2}{8} \right] + \left[\frac{90^2 + 93^2}{4} - \frac{(183)^2}{8} \right] + \left[\frac{98^2 + 86^2}{4} - \frac{(184)^2}{8} \right] \\
 &= (4232.5 - 4232.0) + (4187.25 - 4186.125) + (4250 - 4232) = 0.5 + 1.125 + 18.0 = 18.625.
 \end{aligned}$$

Total effect of *AT* from replication-2 and replication-3 is

$$[AT]_1 = [AT] - [AT] \text{ from replication-1} = -7 - (B_{11} - B_{12}) = -7 - (91 - 93) = -5.$$

$$SS(AT) = \frac{[AT]_1^2}{(r-1)2^3} = \frac{(-5)^2}{2 \times 8} = 1.5625.$$

Total effect of *BT* from replication-1 and replication-3 is

$$[BT]_1 = [BT] - [BT] \text{ from replication-2} = -5 - (90 - 93) = -5 + 3 = -2.$$

$$SS(BT) = \frac{[BT]_1^2}{(r-1)2^3} = \frac{(-2)^2}{2 \times 8} = 0.25.$$

Total effect of *ABT* from replication-1 and replication-2 is

$$[ABT]_1 = [ABT] - [ABT] \text{ from replication-3} = 27 - (B_{31} - B_{32}) = 27 - (98 - 86) = 15.$$

$$SS(ABT) = \frac{[ABT]_1^2}{(r-1)2^3} = \frac{(15)^2}{2 \times 8} = 14.0625.$$

$$SS \text{ (Effect and interactions)} = SS(A) + SS(B) + \dots + SS(ABT) = 220.042.$$

$$SS \text{ (Total)} = \sum \sum \sum \sum y_{ijkl}^2 - C.T. = 12909 - 12650.042 = 258.958.$$

$$SS \text{ (Error)} = SS \text{ (Total)} - SS \text{ (Blocks within replications)} - SS \text{ (Effects and interactions)} \\ = 258.958 - 18.625 - 220.042 = 20.291.$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F	F _{0.05}	P-value
Blocks within replications	3	18.625	6.208	3.98	3.41	< 0.05
<i>A</i>	1	63.375	63.375	40.60	4.67	0.00
<i>B</i>	1	135.375	135.375	86.73	"	0.00
<i>T</i>	1	3.375	3.375	2.16	"	> 0.05
Main effects	— 3	— 202.125	67.375	— 43.17	3.41	0.00
<i>AB</i>	1	2.042	2.042	1.31	4.67	> 0.05
<i>AT</i>	1'	1.5625	1.5625	1.00	"	"
<i>BT</i>	1'	0.25	0.25	0.16	"	"
<i>ABT</i>	1'	14.0625	14.0625	9.01	"	0.00
Error	13	20.291	1.5608	—		
Total	23					

It is observed that the main effects are highly significant. However, the effect of temperature is not significant. The partially confounded three-factor interaction is also highly significant. The results indicate that increasing the levels of concentration of abate and benlate more wood lice can be killed.

The estimates of confounded effect are :

$$AT = \frac{[AT]_1}{(r-1)2^{3-1}} = \frac{-5}{2 \times 4} = -0.625, \quad BT = \frac{[BT]_1}{(r-1)2^{3-1}} = \frac{-2}{2 \times 4} = -0.25,$$

$$ABT = \frac{[ABT]_1}{(r-1)2^{3-1}} = \frac{15}{2 \times 4} = 1.875$$

The variance of any of the unconfounded effects or interactions is

$$V(X) = \frac{MS(\text{Error})}{r2^{3-2}}, \text{ where } X \text{ is effect or interaction}$$

$$= \frac{1.5608}{3 \times 2} = 0.2601$$

and variance of confounded interaction is

$$V(X_1) = \frac{MS(\text{Error})}{(r-1)2^{3-1}}, \text{ where } X_1 \text{ is confounded effect or interaction}$$

$$= \frac{1.5608}{2 \times 2} = 0.3902.$$

Therefore, unconfounded interactions or effects are estimated more efficiently than the confounded interactions.

Example 4.6 : An experiment is conducted to study the productivity of marigold using three doses of nitrogen as urea and 3 doses of phosphorus. The doses of nitrogen are 30 kg/acre, 60 kg/acre and 90 kg/acre; the doses of phosphorus are 20 kg/acre, 40 kg/acre and 60 kg/acre. The plants are cultivated in the plots at distances of 6", 9" and 12". The experiment is a 3³-factorial experiment conducted through randomized block design of 3 plots each. The number of flowers per plant and the arrangement of treatments in blocks are shown below :

Replications	Blocks	Number of flowers (y_{ijk}) per plant under different treatments			Total of b_{ij}
1	1	000-12	012-21	021-18	51 = B_{11}
	2	100-16	121-24	112-19	59 = B_{12}
	3	200-20	212-23	221-20	63 = B_{13}
	4	222-28	210-18	201-16	62 = B_{14}
	5	120-20	102-17	111-18	55 = B_{15}
	6	220-18	202-18	211-22	58 = B_{16}
	7	022-26	010-15	001-14	55 = B_{17}
	8	110-18	101-20	122-22	60 = B_{18}
	9	002-14	011-16	020-15	45 = B_{19}
2	1	012-19	021-19	000-13	51 = B_{21}
	2	112-20	121-23	100-17	60 = B_{22}
	3	212-22	200-18	221-19	59 = B_{23}
	4	222-27	210-19	201-17	63 = B_{24}
	5	120-19	111-18	102-18	55 = B_{25}
	6	220-19	202-20	211-21	60 = B_{26}
	7	022-25	001-18	010-19	62 = B_{27}
	8	101-22	110-21	122-24	67 = B_{28}
	9	002-18	011-19	020-14	51 = B_{29}

Analyse the data and comment on the use of fertilizer.

Solution : It is observed that

$$(ABC)_0 = B_{11} + B_{14} + B_{15}, (ABC)_1 = B_{12} + B_{16} + B_{17}, (ABC)_2 = B_{13} + B_{18} + B_{19},$$

$$(BC)_0 = B_{11} + B_{13} + B_{12}, (BC)_1 = B_{14} + B_{17} + B_{18}, (BC)_2 = B_{15} + B_{16} + B_{19}.$$

Therefore, ABC and BC interactions each with 2 d.f are confounded with blocks. the generalized interaction of ABC and BC is AB_2C_2 and it is also confounded with blocks. Since the arrangement of treatments in plots of blocks in second replication is similar to that in replication-1, the three interactions ABC, BC and AB_2C_2 each of 2 d.f. are also confounded with blocks. The three interactions each of 2 d.f. are totally confounded with blocks. The sum of squares of these 3 interactions will not be calculated. Sum of squares of other effects and interactions each of 2 d.f. are calculated as usual.

The treatment totals are given below :

Treatments	000 001 002	010 011 012	020 021 022	100 101 102	
Total of treatments	25 32 32	34 35 40	29 37 51	33 42 35	
Treatments	110 111 112	120 121 122	200 201 202	210 211 212	220 221 222
Total of treatments	39 36 39	39 47 46	38 33 38	37 43 45	37 39 55

$$G = 1036, C.T = \frac{G^2}{r3^3} = \frac{(1036)^2}{2 \times 27} = 19875.8518.$$

$$SS \text{ (Total)} = \sum \sum \sum \sum y_{ijkl}^2 - C.T. = 2082 - 19875.8518 = 606.1482.$$

$$R_1 = \sum B_{ij} = 508, R_2 = \sum B_{2j} = 528,$$

$$SS \text{ (Replications)} = \frac{1}{27} \sum R_i^2 - C.T. = \frac{536848}{27} - 19875.8518 = 7.4075$$

$$SS \text{ (Blocks with replications)} = \sum_l \left[\frac{\sum B_{ij}^2}{3} - \frac{(\sum B_{ij})^2}{27} \right]$$

$$= \left[\frac{28934}{3} - \frac{(508)^2}{27} \right] + \left[\frac{31210}{3} - \frac{(528)^2}{27} \right]$$

$$= (9644.667 - 9557.926) + (10403.333 - 10325.333) = 164.741.$$

$$A_0 = 000 + 001 + 002 + 010 + 011 + 012 + 020 + 021 + 022 = 315$$

$$A_1 = 100 + 101 + 102 + 110 + 111 + 112 + 120 + 121 + 122 = 356$$

$$A_2 = 200 + 201 + 202 + 210 + 211 + 212 + 220 + 221 + 222 = 365$$

$$B_0 = 000 + 001 + 002 + 100 + 101 + 102 + 200 + 201 + 202 = 308$$

$$B_1 = 010 + 011 + 012 + 110 + 111 + 112 + 210 + 211 + 212 = 348$$

$$B_2 = 020 + 021 + 022 + 120 + 121 + 122 + 220 + 221 + 222 = 380$$

$$C_0 = 000 + 010 + 020 + 100 + 110 + 120 + 200 + 210 + 220 = 311$$

$$C_1 = 001 + 011 + 021 + 101 + 111 + 121 + 201 + 211 + 221 = 344$$

$$C_2 = 002 + 012 + 022 + 102 + 112 + 122 + 202 + 212 + 222 = 381$$

$$(AB)_0 = 000 + 001 + 002 + 120 + 121 + 122 + 210 + 211 + 212 = 346$$

$$(AB)_1 = 100 + 101 + 102 + 010 + 011 + 012 + 220 + 221 + 222 = 350$$

$$(AB)_2 = 200 + 201 + 202 + 020 + 021 + 022 + 110 + 111 + 112 = 340$$

$$(AB_2)_0 = 000 + 001 + 002 + 110 + 111 + 112 + 220 + 221 + 222 = 334$$

$$(AB_2)_1 = 100 + 101 + 102 + 210 + 211 + 212 + 020 + 021 + 022 = 352$$

$$(AB_2)_2 = 200 + 201 + 202 + 120 + 121 + 122 + 010 + 011 + 012 = 350$$

$$(AC)_0 = 000 + 010 + 020 + 102 + 112 + 122 + 201 + 211 + 221 = 323$$

$$(AC)_1 = 100 + 110 + 120 + 001 + 011 + 021 + 202 + 212 + 222 = 353$$

$$(AC)_2 = 200 + 210 + 220 + 002 + 012 + 022 + 101 + 111 + 121 = 360$$

$$(AC_2)_0 = 000 + 010 + 020 + 202 + 212 + 222 + 101 + 111 + 121 = 351$$

$$(AC_2)_1 = 100 + 110 + 120 + 201 + 211 + 221 + 002 + 012 + 022 = 349$$

$$(AC_2)_2 = 200 + 210 + 220 + 001 + 011 + 021 + 102 + 112 + 122 = 336$$

$$(BC_2)_0 = 000 + 100 + 200 + 011 + 111 + 211 + 022 + 122 + 222 = 362$$

$$(BC_2)_1 = 010 + 110 + 210 + 002 + 102 + 202 + 021 + 121 + 221 = 338$$

$$(BC_2)_2 = 020 + 120 + 220 + 012 + 112 + 212 + 001 + 101 + 201 = 336$$

$$(ABC_2)_0 = 000 + 011 + 101 + 221 + 112 + 022 + 210 + 120 + 202 = 345$$

$$(ABC_2)_1 = 100 + 010 + 220 + 111 + 002 + 201 + 021 + 212 + 122 = 333$$

$$(ABC_2)_2 = 200 + 020 + 110 + 102 + 012 + 222 + 211 + 121 + 001 = 358$$

$$(AB_2C)_0 = 000 + 011 + 110 + 212 + 121 + 022 + 201 + 220 + 102 = 347$$

$$(AB_2C)_1 = 001 + 012 + 020 + 100 + 111 + 122 + 202 + 210 + 221 = 330$$

$$(AB_2C)_2 = 002 + 010 + 021 + 101 + 112 + 120 + 200 + 211 + 222 = 359$$

$$SS(A) = \frac{1}{9r}(A_0^2 + A_1^2 + A_2^2) - \text{C.T.} = \frac{359186}{9 \times 2} - 19875.8518 = 78.9260$$

$$SS(B) = \frac{1}{9r}(B_0^2 + B_1^2 + B_2^2) - \text{C.T.} = \frac{360368}{9 \times 2} - 19875.8518 = 144.5926$$

$$SS(C) = \frac{1}{9r}(C_0^2 + C_1^2 + C_2^2) - \text{C.T.} = \frac{360218}{9 \times 2} - 19875.8518 = 136.2593$$

$$SS(AB) = \frac{1}{9r}[(AB)_0^2 + (AB)_1^2 + (AB)_2^2] - \text{C.T.} = \frac{357816}{9 \times 2} - 19875.8518 = 2.8149$$

$$SS(AB_2) = \frac{1}{9r}[(AB_2)_0^2 + (AB_2)_1^2 + (AB_2)_2^2] - \text{C.T.} = \frac{357960}{9 \times 2} - 19875.8518 = 10.8149$$

$$SS(AC) = \frac{1}{9r}[(AC)_0^2 + (AC)_1^2 + (AC)_2^2] - \text{C.T.} = \frac{358538}{9 \times 2} - 19875.8518 = 42.9260$$

$$SS(AC_2) = \frac{1}{9r}[(AC_2)_0^2 + (AC_2)_1^2 + (AC_2)_2^2] - \text{C.T.} = \frac{357898}{9 \times 2} - 19875.8518 = 7.3704$$

$$SS(BC_2) = \frac{1}{9r}[(BC_2)_0^2 + (BC_2)_1^2 + (BC_2)_2^2] - \text{C.T.} = \frac{358184}{9 \times 2} - 19875.8518 = 23.2593$$

$$SS(ABC_2) = \frac{1}{9r}[(ABC_2)_0^2 + (ABC_2)_1^2 + (ABC_2)_2^2] - \text{C.T.} = \frac{358078}{9 \times 2} - 19875.8518 = 17.3704$$

$$SS(AB_2C) = \frac{1}{9r}[(AB_2C)_0^2 + (AB_2C)_1^2 + (AB_2C)_2^2] - \text{C.T.} = \frac{358190}{9 \times 2} - 19875.8518 = 23.5926.$$

$$SS(\text{Effects and interactions}) = SS(A) + SS(B) + SS(C) + \dots + SS(AB_2C) = 487.9264$$

$$SS(\text{Error}) = SS(\text{Total}) - SS(\text{Replications}) - SS(\text{Effects and interaction})$$

$$= 606.1482 - 7.4074 - 487.9264 = 110.8144.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{.05}$	P -value
Replications	1	7.4074	7.4074	2.14	4.15	> 0.05
A	2	78.9260	39.463	11.40	3.30	0.00
B	2	144.5926	72.2963	20.88	"	0.00
C	2	136.2593	68.1296	19.67	"	0.00
AB	2	2.8149	1.4074	0.41	"	> 0.05
AB ₂	2	10.8149	5.4074	1.56	"	> 0.05
AB	— 4	— 13.6298	— 3.4074	0.98	2.67	> 0.05
AC ₂	2	7.3704	3.6852	1.06	3.30	"
AC	2	42.926	21.463	6.20	"	< 0.05
AC	— 4	— 50.2964	— 12.5741	3.63	2.67	< 0.05
BC ₂	2	23.2593	11.6296	3.36	3.30	< 0.05
ABC ₂	2	17.3704	8.6852	2.51	"	> 0.05
AB ₂ C	2	23.5926	11.7963	3.41	"	< 0.05
Error	32	110.8144	3.46295	—	—	—
Total	53					

Here *A* is used for nitrogen, *B* is used for phosphorus and *C* is used for spacing.

It is observed that the changes in the levels of nitrogen, phosphorus and spacing significantly increase the production of flower. The joint impact of nitrogen and spacing is also found significant.

4.7 Fractional Replication of Factorial Experiment

In factorial experiment if the factors or levels of factors or both are large, the treatment combinations become large. The use of a large number of treatments creates problem, specially if the experiment is conducted through randomized block design. The problem is obviated using smaller number of plots per block, where there are multiple blocks per replication. The technique of allocation of treatment to the plots of a block of smaller size than the number of treatment is known as *confounding*. Sometimes the confounding technique is also not sufficient to reduce the block size. Further reduction of block size is needed for certain types of larger factorial experiment.

One of the technique of selecting a set of treatments from all treatments is known as use of fractional replication of treatments so that the number of plots per block is reduced to $1/p^k (k < n)$ for p^n -factorial experiment. Finney (1945) is the first man to introduce such a technique to select a fraction of treatments from all treatments. The selected treatments are used for the experiment and experiment is performed in a similar way as it is done for other factorial experiment.

The fraction of treatments is profitably used in any experiment if the experimenter is not interested in any higher order interaction or interactions. It is noted that any effect or interaction is expressed as a linear combination of results of treatments, where sum of the coefficients of the linear combination is zero.

For example, the interaction *ABC* in 2^3 -factorial experiment is

$$ABC = \frac{1}{2^{3-1}}(a-1)(b-1)(c-1) = \frac{1}{2^2} [abc + a + b + c - (1) - ab - ac - bc].$$

An experiment can be conducted using all 8 treatments allocating them in two separate blocks so that block contrast represents ABC interaction. The technique is known as *confounding*. Thus, we have

						Total of block
Replication	Block-1	abc	a	b	c	B_1
	Block-2	(1)	ab	ac	bc	B_2

$[ABC] = B_1 - B_2$, the block size is reduced to half. Here ABC interaction is not estimated. Here $abc, a, b,$ and c treatments are $\frac{1}{2} 2^3$ - factorial experiment. Another half of treatments are : $(1), ab, ac, bc$. If in away experiment either half of treatments are used, the experiment is known as *fractional factorial experiment*.

The use of abc, a, b and c as treatments in an experiment or the use of $(1), ab, ac$ and bc as treatments do not give information on ABC interaction. Either half of the treatments are not suitable to estimate ABC contrast.

Defining contrast : The interaction which can not be estimated in any factorial experiment due to the use of fractional replication of treatments is known as defining contrast. Thus, in the above mentioned case of $\frac{1}{2} 2^3$ -factorial experiment if abc, a, b and c are used as treatment, the interaction ABC cannot be estimated. The interaction is known as *defining contrast*.

The interactions which is not important for the research can be used as defining contrast. If $\frac{1}{p^k} p^n$ -factorial experiment is conducted, k interactions can be used as defining contrasts and the block sizes are reduced further. The generalized interactions of k defining contrasts are also defining contrasts and those interactions cannot be estimated. For example, if $\frac{1}{2^2} 2^3$ -factorial experiment is used, then we need only two treatments. These two treatments can be selected from abc, a, b and c . We know,

$$[AB] = abc + c - a - b.$$

Then, either abc and c or a and b can be used for the experiment and in that case AB interaction cannot be estimated. Here AB is also defining contrast. The generalized interaction of ABC and AB is $ABC \cdot AB = C$. Therefore, ABC, AB and C are defining contrasts. We write :

$$I = ABC = AB = C, \text{ where } I \text{ is used to indicate defining contrast.}$$

Principal block : In $\frac{1}{2} 2^3$ -factorial experiment if ABC is used as defining contrast, the treatments are either abc, a, b and c or $(1), ab, bc$ and ac . The either group of treatments constitutes principal block. Thus, the treatments which are used in any fractional replication of factorial experiment constitutes the principal block. If ABC and AB are used as defining contrast, then abc and c or a and b constitute the principal blocks. The effects and interactions are estimated using the treatments of principal block.

Aliases : For $\frac{1}{2} 2^3$ -factorial experiment if ABC is the defining contrast, then abc, a, b and c treatments constitute the principal block. From these four treatments, we have

$$\begin{aligned}
 [A] &= abc + a - b - c, [B] = abc + b - a - c = [AC] \\
 [C] &= abc + c - a - b, [AB] = abc + c - a - b = [C] \\
 [AC] &= abc + b - a - c, [BC] = abc + a - b - c = [A].
 \end{aligned}$$

Thus, $[A] = [BC], [B] = [AC], [C] = [AB]$.

Here $BC = A$, we call, BC is the alias of A ,

$AC = B$, we call, AC is the alias of B and $AB = C$, we call, A is the alias of C .

The aliases are the generalized interaction of any effect or interaction with defining contrast. We have $I = ABC$. Its generalized interactions are :

$$A \cdot ABC = BC, B \cdot ABC = AC, C \cdot ABC = AB.$$

Again, if (1), ab, ac and bc treatments constitute the principal blocks, then

$$\begin{aligned} [A] &= ab + ac - bc - (1), & [AB] &= ab + (1) - ac - bc = -[C] \\ [B] &= ab + bc - ac - (1), & [AC] &= ac + (1) - ab - bc = -[B] \\ [C] &= ac + bc - ab - (1), & [BC] &= bc + (1) - ab - ac = -[A] \end{aligned}$$

If 2^3 -factorial experiment is replicated r -times, effects and interactions of 7 d.f. can be estimated. These are A, B, C, AB, AC, BC and ABC . But if $\frac{1}{2}2^3$ -factorial experiment is replicated r times, we have 4 treatments to estimate effects and interactions of $(4 - 1) = 3$ d.f. As ABC is the defining contrast, its effect cannot be estimated. The remaining six effects and interactions are A, B, C, AB, AC and BC . But it is observed that $A = BC, B = AC$ and $C = AB$. Therefore, three effects or interactions can be estimated from this experiment.

The sum of squares of effects and interactions are calculated as usual. The effect totals are calculated from the results of treatments, where treatments are identified from the solution of equations of the type

$$\begin{aligned} ix_1 + ix_2 + ix_3 &= 0 \\ &= 1 \end{aligned} \Bigg| \text{ mod } 2; i = 0, 1.$$

The solutions are obtained from the treatment 111, 100, 010, 001. Thus, the solution providing total of A_0 and A_1 are obtained from the equation :

$$\begin{aligned} x_1 &= 0 \\ &= 1 \end{aligned} \Bigg| \text{ mod } 2.$$

We have $(A_0)_0 = 010 + 001, \quad SS(A) = \frac{1}{r^{2^3-1}}[A_0^2 + A_1^2] - \text{C.T.},$

$$(A)_1 = 111 + 100.$$

Similarly, other sum of squares are calculated.

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F
Block	$r - 1$			
A	1			
B	1			
C	1			
Error	$3(r - 1)$			
Total	$r2^2 - 1$			

$\frac{1}{4}2^4$ -Factorial experiment : Let us consider that there are 4 factors A, B, C and D and each of them has 2 levels. Total number of treatment combinations are 16. But we need to conduct the experiment with $\frac{1}{4}2^4 = 4$ treatments. To select 4 treatments we can use two

interactions as defining contrasts. Let these interactions be AB and CD . The treatments to be used to calculate total of $(AB)_0$ and $(AB)_1$ are selected from the solution of the equations :

$$\begin{array}{l} x_1 + x_2 = 0 \\ = 1 \end{array} \Bigg| \text{ mod } 2.$$

Thus, we have

$$(AB)_0 = 0000 + 0010 + 0001 + 0011 + 1100 + 1101 + 1110 + 1111$$

$$(AB)_1 = 1000 + 0100 + 1010 + 0110 + 1001 + 0101 + 0111 + 1011.$$

Either group of treatments for $(AB)_0$ and $(AB)_1$ can be considered for $\frac{1}{2}2^4$ -factorial experiment. Let us consider that the treatments used to get $(AB)_0$ are under consideration. From these 8 treatments we need to select four treatments which will be used to get $(CD)_0$ and $(CD)_1$. The treatments are selected from the solutions of the equations :

$$\begin{array}{l} x_2 + x_3 = 0 \\ = 1 \end{array} \Bigg| \text{ mod } 2.$$

Thus, we have

$$(CD)_0 = 0000 + 1100 + 0011 + 1111$$

$$(CD)_1 = 0010 + 0001 + 1101 + 1110.$$

Either group of 4 treatments for $(CD)_0$ or for $(CD)_1$ can be used as required 4 treatments for the experiment. Let the selected groups be 0000, 1100, 0011 and 1111.

Since AB and CD are used as defining contrasts, their generalized interaction $AB \cdot CD = ABCD$ is also defining contrast. Therefore,

$$I = AB = CD = ABCD.$$

The alias group of effects and interactions are :

$$A = B = ACD = BCD$$

$$C = ABC = D = ABD$$

$$BC = AC = BD = AD.$$

Thus, we can estimate effects and interactions of 3 d.f. Let these effects and interaction be A , C and BC . Thus, we have

$$(A)_0 = 0000 + 0011, (A)_1 = 1100 + 1111, (C)_0 = 0000 + 1100,$$

$$(C)_1 = 0011 + 1111, (BC)_0 = 0000 + 1111, (BC)_1 = 1100 + 0011.$$

$SS(A) = \frac{1}{r^2}(A_0^2 + A_1^2) - \text{C.T.}$, where r is the number of replications of treatments. The other sum of squares are calculated as usual and in a similar way.

ANOVA Table

Sources of variation	d.f.
Replications	$r - 1$
A	1
C	1
BC	1
Error	$3(r - 1)$
Total	$r^2 - 1$

$\frac{1}{3}3^2$ -Factorial experiment : Let the three factors be A, B and C and each has three levels. Total level combinations are 9. We need to select three treatments for our experiment. To select 3 treatments let us consider AB_2 as defining contrast. Then, we have

$$(AB_2)_0 = 00 + 11 + 22, (AB_2)_1 = 10 + 21 + 02, (AB_2)_2 = 20 + 12 + 01.$$

There are 3 groups of treatments used to get $(AB_2)_0, (AB_2)_1$ and $(AB_2)_2$. Any of the group of treatments can be used for the required experiment. Let the selected treatments be 00, 11 and 22.

Here $I = AB_2$. The alias group of effects and interactions are

$$A = AB = B.$$

Since we have only 3 treatments to be used for the experiment, we can estimate effects or interactions of 2 d.f. We can estimate the effect of A , where A has 2 d.f. Also, we have $(A)_0 = 00, (A)_1 = 11, (A)_2 = 22$. Then

$$SS(A) = \frac{1}{r}(A_0^2 + A_1^2 + A_2^2) - C.T.$$

ANOVA Table

Sources of variation	d.f.
Replication	$r - 1$
A	1
Error	$2r - 1$
Total	$3r - 1$

Confounding in fractional replication : The fractional replication of factorial experiment helps us in reducing the number of treatments so that the treatments are allocated in homogeneous plots within a block. Further reduction of block size can be achieved if the selected group of treatments are allocated in different blocks within a replication so that the block contrast represent a higher order interaction. The higher order interaction is confounded with blocks. The technique of allocating the selected treatments in different blocks within a replication when block contrast represent a higher order interaction is known as confounding in fractional replication. The technique is useful if further reduction in block size is needed for any experiment.

The confounded interaction and its generalized interaction with defining contrast is also confounded with blocks. Therefore, the information on defining contrast and confounded contrast are lost in such an experiment. Hence, confounding is done in such a way that no main effect is confounded with blocks.

$\frac{1}{3}3^3$ -Factorial experiment in 3 blocks : Let there be 3 factors A, B and C each having 3 levels. Total number of level combinations are 27 and we need to select 9 treatment combinations for our experiment. To select 9 treatments, we need to use an interaction as defining contrast. Let the defining contrast be AB_2C_2 .

Then, we have

$$(AB_2C_2)_0 = 000 + 012 + 021 + 101 + 110 + 122 + 202 + 211 + 220$$

$$(AB_2C_2)_1 = 002 + 011 + 020 + 100 + 112 + 121 + 201 + 210 + 222$$

$$(AB_2C_2)_2 = 001 + 010 + 022 + 102 + 111 + 120 + 200 + 212 + 221.$$

There are 3 groups of treatments to calculate $(AB_2C_2)_0$, $(AB_2C_2)_1$ and $(AB_2C_2)_2$. In each group there are 9 treatments. Any of the group can be selected for the experiment. Let the selected group of treatment be 000, 012, 021, 101, 110, 122, 202, 211 and 220. These 9 treatments are to be allocated in 3 blocks of 3 plots each. This is possible if a higher order interaction is confounded with blocks. Let this interaction be BC_2 . Then, using the above mentioned 9 treatments, we have

$$(BC_2)_0 = 000 + 122 + 211, (BC_2)_1 = 021 + 110 + 202, (BC_2)_2 = 012 + 101 + 220.$$

Therefore, the block contents per replication will be

$$\text{Block-1 } \begin{array}{|c|c|c|} \hline 000 & 122 & 211 \\ \hline \end{array}, \text{ Block-2 } \begin{array}{|c|c|c|} \hline 021 & 110 & 202 \\ \hline \end{array}, \text{ Block-3 } \begin{array}{|c|c|c|} \hline 012 & 101 & 220 \\ \hline \end{array}$$

The generalized interaction of BC_2 and AB_2C_2 is AC . Hence, AC interaction is also confounded with blocks. The alias group of effects and interactions are :

$$A = ABC = BC$$

$$B = AC_2 = ABC_2$$

$$C = AB_2 = AB_2C$$

$$AB = AC = BC_2.$$

Therefore, we can estimate effects and interactions having 6 d.f. These effects are A , B and C . The interactions AB , AC and BC_2 are confounded with blocks. We have

$$(A)_0 = 000 + 021 + 012, (A)_1 = 122 + 110 + 101, (A)_2 = 211 + 202 + 220,$$

$$SS(A) = \frac{1}{3r}(A_0^2 + A_1^2 + A_2^2) - \text{C.T.}$$

$SS(B)$ and $SS(C)$ are calculated in a similar way.

ANOVA Table

Sources of variation	d.f.
Blocks within replication	$2r$
A	2
B	2
C	2
Error	$7(r - 1)$
Total	$9r - 1$

Example 4.7 : An experiment is conducted to study the change in milk production of cows after giving 3 new foods, say A , B and C , to cows of 3 lactation period. The cows are considered as factor D . Each factor A , B and C has 3 levels. The new foods are given to cows in 3 different sheds. In each shed there are 9 cows and each shed is considered as a block. The experiment is conducted according to $\frac{1}{3^2}3^4$ -factorial plan. The milk production (in kg per cow) is recorded in a day during the experiment. The plan of treatment in blocks and milk production per cow are shown below :

Milk production (in kg/cow) according to food and lactation period

Levels of treatment <i>ABCD</i>	Milk production in sheds (blocks)			Total of treatments
	1	2	3	
0000	18.5	19.0	19.0	56.5
1100	23.4	24.2	23.8	71.4
1010	25.0	26.2	25.8	77.0
2110	30.0	31.5	31.0	92.5
0120	28.5	27.5	27.0	83.0
0210	27.3	28.0	27.0	82.3
2020	32.4	33.0	32.6	98.0
2200	35.0	34.2	34.3	103.5
1220	26.0	26.5	25.8	78.3
Total of blocks B_i	246.1	250.1	246.3	742.5

Analyse the data and comment on the suitability of new foods.

Solution : We have $r = 3$, number of treatments are 9.

The defining contrasts are $I = AB_2C_2D = AB_2C_2D_2 = AB_2C_2 = D$.

The alias group of effects and interactions are :

$$A = AD = AD_2 = BC = ABC = BCD = BCD_2 = ABCD = ABCD_2$$

$$B = AC_2 = AC_2D = AC_2D_2 = BD = BD_2 = ABC_2 = ABC_2D = ABC_2D_2$$

$$C = AB_2 = CD = CD_2 = AB_2D = AB_2C = AB_2CD = AB_2CD_2 = AB_2D_2$$

$$AB = AC = BC_2 = ACD = ACD_2 = ABD = ABD_2 = BC_2D = BC_2D_2.$$

Therefore, we can estimate effects and interactions having 8 d.f. These effects and interactions are A, B, C and AB .

Here $C.T. = \frac{G^2}{r9} = \frac{(742.5)^2}{3 \times 9} = 20418.75.$

$$SS \text{ (Total)} = \sum \sum \sum \sum \sum y_{ijklm}^2 - C.T. = 20966.55 - 20418.75 = 547.8.$$

$$SS \text{ (Blocks)} = \frac{1}{9} \sum B_i^2 - C.T. = \frac{183778.91}{9} - 20418.75 = 1.129.$$

$$(A)_0 = 0000 + 0120 + 0210 = 221.8, (A)_1 = 1100 + 1010 + 1220 = 226.7.$$

$$(A)_2 = 2020 + 2200 + 2110 = 294.0.$$

$$SS(A) = \frac{1}{3r} (A_0^2 + A_1^2 + A_2^2) - C.T. = \frac{187024.13}{3 \times 3} - 20418.75 = 361.709.$$

$$(B)_0 = 0000 + 1010 + 2020 = 231.5, (B)_1 = 1100 + 2110 + 0120 = 246.9.$$

$$(B)_2 = 0210 + 2200 + 1220 = 264.1.$$

$$SS(B) = \frac{1}{3r} (B_0^2 + B_1^2 + B_2^2) - C.T. = \frac{184300.67}{9} - 20418.75 = 59.102.$$

$$(C)_0 = 0000 + 1100 + 2200 = 231.4, (C)_1 = 1010 + 2110 + 0210 = 251.8.$$

$$(C)_2 = 0120 + 2020 + 1220 = 258.3.$$

$$SS(C) = \frac{1}{3r} (C_0^2 + C_1^2 + C_2^2) - C.T. = \frac{184185.69}{9} - 20418.75 = 46.327.$$

$$(AB)_0 = 0000 + 2110 + 1220 = 227.3, (AB)_1 = 1010 + 0120 + 2200 = 263.5.$$

$$(AB)_2 = 1100 + 0210 + 2020 = 251.7.$$

$$SS(AB) = \frac{1}{3r}(AB_0^2 + AB_1^2 + AB_2^2) - \text{C.T.} = \frac{184450.43}{9} - 20418.75 = 75.742.$$

$$SS(\text{Error}) = SS(\text{Total}) - SS(\text{Blocks}) - SS(A) - SS(B) - SS(C) - SS(AB) \\ = 547.8 - 1.129 - 361.709 - 59.102 - 46.327 - 75.742 = 3.791.$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{\text{d.f.}}$	F	F _{0.05}	P-value
Blocks	2	1.129	0.5645	2.38	3.63	> 0.05
A	2	361.709	180.8545	763.42	"	0.00
B	2	59.102	29.551	124.74	"	0.00
C	2	46.327	23.1635	97.78	"	0.00
AB	2	75.742	37.871	159.86	"	0.00
Error	16	3.791	0.2359	—	—	—
Total	26					

It is observed that all the three foods are highly significantly effective in increasing milk production. The joint effect of first two foods (AB) is also highly significant.

$$\text{Here } \bar{A}_0 = 24.64, \bar{A}_1 = 25.19, \bar{A}_2 = 32.67; \bar{B}_0 = 25.72, \bar{B}_1 = 27.43$$

$$\bar{B}_2 = 29.34; \bar{C}_0 = 25.71, \bar{C}_1 = 27.98, \bar{C}_2 = 28.81.$$

It is observed that with the change in the levels of food the milk production increases. This is true for all the 3 foods. However, the significance in the difference of the means can be verified by Duncan's multiple range test, where the test statistic is

$$D_k = d_{0.05, k, f} \sqrt{\frac{MS(\text{Error})}{9}}; k = 2, 3; f = 16.$$

$$\text{Here } D_2 = d_{0.05, 2, 16} \sqrt{\frac{MS(\text{Error})}{9}} = 3.00 \sqrt{\frac{0.2369}{9}} = 0.486$$

$$D_3 = d_{0.05, 3, 16} \sqrt{\frac{MS(\text{Error})}{9}} = 3.15 \sqrt{\frac{0.2369}{9}} = 0.511.$$

All the means are significantly different from each other. This is true for all the three foods.

4.8 Advantages and Disadvantages of Factorial Experiment

Advantages :

- (i) The main effects of many factors can be studied simultaneously from one experiment.
- (ii) The interaction of many factors can be estimated from one single experiment.
- (iii) The main effects are estimated from the results of each plot and hence, efficiency of the experiment increases in factorial experiment.
- (iv) The main effects and interaction of several factors can be studied using minimum experimental materials.

Disadvantages :

- (i) If factors or levels or both are more, the treatment combinations become large. Therefore, in conducting the experiment through randomized block design the homogeneity of plots within a block may be lost.
- (ii) The interaction of factors in case of higher levels is not easily interpreted.

4.9 Advantages and Disadvantages of Confounding**Advantages :**

- (i) Since all treatments are not allocated in a block, the block size is reduced in respect of number of plots and hence, homogeneity of blocks is achieved and the efficiency of the experiment increases.
- (ii) Large number of treatments can be allocated within the blocks in a replication.

Disadvantages :

- (i) The information of confounded interaction is lost.
- (ii) Some interaction is estimated from smaller number of replications, specially when an interaction is partially confounded.
- (iii) No interaction can be confounded totally in case of factorial experiments with higher levels of factors.
- (iv) The analysis is slightly complicated and complication arises if there is any interaction of treatment with incomplete blocks.

4.10 Asymmetrical Factorial Experiment

The factorial experiments, so far we have discussed, are symmetrical in number of levels. In practice, different factors may have different levels. If different factors have different levels, then the factorial experiment is known as asymmetrical factorial experiment. For example, let us consider that A is a factor having p levels and B is another factor having q levels. The total number of level combinations are pq . If all pq levels are considered as treatments in any factorial experiment, the experiment is known as asymmetrical factorial experiment.

The asymmetrical factorial experiment may be conducted through randomized block design or any other designs. The important designs used for asymmetrical factorial experiment are (i) Randomized block design, (ii) Split-plot design and (iii) Split-split-plot design.

$p \times q$ -Factorial experiment : Let there be two factors A and B , where A has p levels and B has q levels. The total number of level combinations are pq . Let these pq treatments are allocated in pq plots of a block, where there are r blocks to replicate each treatment r times. Let y_{ijl} be the observation of i -th level of B corresponding to j -th level of A in i -th block; $i = 1, 2, \dots, r$; $j = 1, 2, \dots, p$; $l = 1, 2, \dots, q$. The model for y_{ijl} observation is

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l + (\beta\gamma)_{jl} + e_{ijl}, \quad (61)$$

where μ = general mean, α_i = effect of i -th block, β_j = effect of j -th level of A , γ_l = effect of l -th level of B , $(\beta\gamma)_{jl}$ = interaction of j -th level of A with l -th level of B and e_{ijl} = random error.

The estimates of parameters and sum of squares of effects and interactions are easily obtained in a similar way as these are obtained in two-way classification with several observations per cell. Here

$$SS(A) = \frac{1}{qr} \sum y_{.j.}^2 - \text{C.T.}, \quad SS(B) = \frac{1}{pr} \sum y_{.l.}^2 - \text{C.T.}$$

$$SS(AB) = \frac{1}{r} \sum \sum y_{.jl}^2 - \text{C.T.} - SS(A) - SS(B)$$

where $\text{C.T.} = \frac{G^2}{pqr}$ and $G = \text{grand total} = \sum \sum \sum y_{ijl}$; $SS(\text{Block}) = \frac{1}{pq} \sum y_{i..}^2 - \text{C.T.}$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f}$	$F = s_i/s_5, i = 1, 2, 3, 4$
Blocks	$r - 1$	S_1	s_1	
A	$p - 1$	S_2	s_2	
B	$q - 1$	S_3	s_3	
AB	$(p - 1)(q - 1)$	S_4	s_4	
Error	$(pq - 1)(r - 1)$	S_5	s_5	
Total	$pqr - 1$			

The effect A has $(p - 1)$ d.f. and the effect B has $(q - 1)$ d.f. This effects can be expressed as linear, quadratic, cubic, quartic, and so on components each of 1 d.f. Each contrast of 1 d.f. is a linear combination of results of treatments. Depending on the number of levels, the coefficients of different levels combinations are different. The coefficients of each combination to calculate effect total and the divisor to calculate sum of squares are shown in the following table for $p = 2, q = 6$ [2 × 6 factorial experiment].

Table below shows coefficients and divisor to calculate effect total and sum of squares of effects and interactions of 2 × 6-factorial experiment.

Effects and interactions	Levels of factors												Divisor D_i
	a_0						a_1						
	b_0	b_1	b_2	b_3	b_4	b_5	b_0	b_1	b_2	b_3	b_4	b_5	
A	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	12
B ₁	-5	-3	-1	1	3	5	-5	-3	-1	1	3	5	140
AB ₁	5	3	1	-1	-3	5	-5	-3	-1	1	3	5	140
B ₂	5	-1	-4	-4	-1	5	5	1	-4	-4	-1	5	168
AB ₂	-5	1	4	4	1	-5	5	-1	-4	-4	-1	5	168
B ₃	-5	7	4	-4	-7	5	-5	7	4	-4	-7	5	328
AB ₃	5	-7	-4	4	7	-5	-5	7	4	-4	-7	5	328
B ₄	1	-3	2	2	-3	1	1	-3	-2	2	-3	1	56
AB ₄	-1	3	-2	-2	3	-1	1	-3	-2	2	-3	1	56
B ₅	-1	5	-10	10	-5	1	-1	5	-10	10	-5	1	504
AB ₅	1	-5	10	-10	5	-1	-1	5	-10	10	-5	1	504

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	$F = s_i/s_9, i = 1, 2, \dots, 8$
Blocks	$r - 1$	S_1	s_1	
A	$p - 1$	S_2	s_2	
B	$q - 1$	S_3	s_3	
C	$m - 1$	S_4	s_4	
AB	$(p - 1)(q - 1)$	S_5	s_5	
AC	$(p - 1)(m - 1)$	S_6	s_6	
BC	$(q - 1)(m - 1)$	S_7	s_7	
ABC	$(p - 1)(q - 1)(m - 1)$	S_8	s_8	
Error	$(r - 1)(pqm - 1)$	S_9	s_9	
Total	$pqrm - 1$			

Example 4.8 : An experiment is conducted in an agricultural research station to study the productivity of high yielding variety of rice using 4 levels of nitrogen, 3 levels of phosphorus and 2 levels of potash. The levels of nitrogen are 30 kg/acre, 60 kg/acre, 90 kg/acre and 120 kg/acre. The levels of phosphorus are 30 kg/acre, 60 kg/acre and 90 kg/acre. The levels of potash are 20 kg/acre and 40 kg/acre. The experiment is conducted through randomized block design having 3 blocks. The production of rice (kg/plot) of each treatment in different replications are shown below :

Treatment combinations			Blocks			Total	Treatment combinations			Blocks			Total
n	p	k	1	2	3	$y_{.jlk}$	n	p	k	1	2	3	$y_{.jlk}$
0	0	0	5.2	5.6	5.5	16.3	2	0	0	9.2	9.0	9.5	27.7
0	0	1	5.3	5.6	5.4	16.3	2	0	1	9.7	9.2	9.6	28.5
0	1	0	6.2	6.0	6.0	18.2	2	1	0	10.8	11.0	10.9	32.7
0	1	1	6.5	6.5	6.2	19.2	2	1	1	11.8	11.6	11.0	34.4
0	2	0	7.0	6.8	6.7	20.5	2	2	0	12.6	11.6	11.4	35.6
0	2	1	7.2	7.0	6.6	20.8	2	2	1	12.5	11.8	11.8	36.1
0	0	0	8.2	8.0	8.2	24.4	3	0	0	12.0	12.0	12.1	36.1
1	0	1	8.8	8.8	8.6	26.2	3	0	1	12.4	12.2	12.0	36.6
1	1	0	10.2	10.5	10.4	31.1	3	1	0	12.8	12.6	12.5	37.9
1	1	1	10.5	10.4	10.6	31.5	3	1	1	12.9	12.7	12.6	38.2
1	2	0	10.8	10.4	10.7	31.9	3	2	0	12.8	12.8	12.8	38.4
1	2	1	10.2	10.7	10.6	31.5	3	2	1	12.8	12.7	12.8	38.3
Total $y_{i..}$						287.9							420.5

Analyse the data and comment on the suitability of different fertilizers.

Solution : Here $r = 3, p = 4, q = 3, m = 2, G = 708.4 = 287.9 + 420.5$

$$C.T. = \frac{G^2}{pqrm} = \frac{(708.4)^2}{4 \times 3 \times 2 \times 3} = 6969.869$$

$$SS(\text{Total}) = \sum \sum \sum \sum y_{ijkl}^2 - C.T. = 7408.74 - 6969.869 = 438.871$$

$$SS(\text{Block}) = \frac{1}{pqm} \sum y_{i..}^2 - C.T. = \frac{167285.06}{4 \times 3 \times 2} - 6969.869 = 0.342$$

The production of nitrogen and phosphorus ($y_{.jl}$)

$P \backslash N$	0	1	2	Total $y_{j..}$
0	32.6	37.4	41.3	111.3
1	50.6	62.6	63.4	176.6
2	56.2	67.1	71.7	195.0
3	72.7	76.1	76.7	225.5
Total $y_{.l}$	212.1	243.2	253.1	708.4

The production of nitrogen and potash ($y_{.jk}$)

$k \backslash N$	0	1
0	55.0	56.3
1	87.4	89.2
2	96.0	99.0
3	112.4	113.1
Total $y_{...k}$	350.8	357.6

Production of phosphorus and potash ($y_{.lk}$)

$P \backslash K$	0	1	2
0	104.5	119.9	126.4
1	107.6	123.3	126.7

$$SS(N) = \frac{1}{qrm} \sum y_{j..}^2 - \text{C.T.} = \frac{132450.5}{3 \times 3 \times 2} - 6969.869 = 388.492$$

$$SS(P) = \frac{1}{prm} \sum y_{.l}^2 - \text{C.T.} = \frac{168192.26}{4 \times 3 \times 2} - 6969.869 = 38.142$$

$$SS(K) = \frac{1}{pqr} \sum y_{.k}^2 - \text{C.T.} = \frac{250938.4}{4 \times 3 \times 3} - 6969.869 = 0.642$$

$$\begin{aligned} SS(NP) &= \frac{1}{rm} \sum \sum y_{.jl}^2 - \text{C.T.} - SS(N) - SS(P) \\ &= \frac{44427.02}{3 \times 2} - 6969.869 - 388.492 - 38.142 = 8.000 \end{aligned}$$

$$\begin{aligned} SS(NK) &= \frac{1}{rq} \sum \sum y_{.jk}^2 - \text{C.T.} - SS(N) - SS(K) \\ &= \frac{66232.46}{3 \times 3} - 6969.869 - 388.492 - 0.642 = 0.159 \end{aligned}$$

$$\begin{aligned} SS(PK) &= \frac{1}{rp} \sum \sum y_{.lk}^2 - \text{C.T.} - SS(P) - SS(K) \\ &= \frac{84106.76}{12} - 6969.869 - 38.142 - 0.642 = 0.244 \end{aligned}$$

$$\begin{aligned} SS(NPK) &= \frac{1}{r} \sum \sum \sum y_{.jlk}^2 - \text{C.T.} - SS(N) - SS(P) - SS(k) - SS(NP) \\ &\quad - SS(NK) - SS(PK) \\ &= \frac{22222.01}{3} - 6969.869 - 388.492 - 38.142 - 0.642 - 8.000 - 0.159 - 0.244 \\ &= 1.507 \end{aligned}$$

$$\begin{aligned}
 SS(\text{error}) &= SS(\text{Total}) - SS(\text{Blocks}) - SS(N) - SS(P) - SS(K) - SS(NP) - SS(NK) \\
 &\quad - SS(PK) - SS(NPK) \\
 &= 438.871 - 0.342 - 388.492 - 38.142 - 0.642 - 8.000 - 0.159 - 0.244 - 1.507 \\
 &= 1.343
 \end{aligned}$$

ANOVA Table

Sources of variation	d.f	SS	MS = $\frac{SS}{d.f}$	F	F ₀₅	P-value
Blocks	2	0.342	0.171	5.86	3.21	0.00
N	3	388.492	129.497	4434.84	2.82	0.00
P	2	38.142	19.071	653.12	3.21	0.00
K	1	0.642	0.642	21.99	4.06	0.00
NP	6	8.000	1.333	45.66	2.31	0.00
NK	3	0.159	0.053	1.82	2.82	> 0.05
PK	2	0.244	0.122	4.18	3.21	< 0.05
NPK	6	1.507	0.251	8.60	2.31	0.00
Error	46	1.343	0.0292	—	—	—
Total	71					

It is observed that except the interaction *NK* all other effects and interactions are significant. The production of rice significantly increases with the increase in the levels of fertilizers. The mean productions are

$$\begin{aligned}
 \bar{N}_0 &= 6.18, \bar{N}_1 = 0.81, \bar{N}_2 = 10.83, \bar{N}_3 = 12.53; \bar{P}_0 = 8.84, \bar{P}_1 = 10.13, \bar{P}_2 = 10.54 \\
 \bar{K}_0 &= 9.74, \bar{K}_1 = 9.93
 \end{aligned}$$

The first two groups of means can be compared pairwise using Duncan's multiple range test, where the test statistic to compare means of nitrogen is :

$$\begin{aligned}
 D_k &= d_{.05,k,f} \sqrt{\frac{MS(\text{error})}{qrm}}; \quad k = 2, 3, 4; \quad f = 46, \\
 D_2 &\approx 2.77 \sqrt{\frac{0.0292}{18}} = 0.111, \quad D_3 \approx 2.92 \sqrt{\frac{0.0292}{18}} = 0.118 \\
 D_4 &\approx 3.02 \sqrt{\frac{0.0292}{18}} = 0.122
 \end{aligned}$$

Thus, each mean of nitrogen is significantly different from other. The production significantly increases with the increase in levels of nitrogen.

The test statistic to compare the means of phosphorus is :

$$\begin{aligned}
 D_k &= d_{.05,k,f} \sqrt{\frac{MS(\text{error})}{prm}}; \quad k = 2, 3; \quad f = 46. \\
 D_2 &\approx 2.77 \sqrt{\frac{0.0292}{24}} = 0.097, \quad D_3 \approx 2.92 \sqrt{\frac{0.0292}{24}} = 0.102
 \end{aligned}$$

Thus, with the increase in the levels of phosphorus the production increases significantly and this is true for every level of phosphorus.

4.11 Split-Plot Design

To conduct the asymmetrical factorial experiment some factor may need larger experimental materials compared to the need of materials of other factor. For example, in agricultural experiment if irrigation is a factor having different levels, it needs large experimental area to apply irrigation of different levels. In such a case if another factor is fertilizer having different levels, a large area used for one level of irrigation may be splitted into several smaller plots where in the smaller plots different doses of a fertilizer can be applied. This means that the larger area, usually known as whole-plot, can be splitted into several sub-plots. In each replication if there are several whole plots and each whole-plot is divided into several sub-plots or split-plots, then the levels of a factor which needs larger materials can be randomly allocated to whole-plots and the levels of another factor which needs smaller materials can be randomly allocated to the split-plots within a whole-plot. The two-steps randomisation process can be replicated several times. Then the resultant design is known as split-plot design.

Let there be two factors A and B , where A has p levels and B has q levels. The levels of A are such that they need larger experimental materials to be used in the experiment and the levels of B are such that they need smaller materials to be used in the experiment. In such a case, the entire experimental materials per replication are divided into p whole-plots and each whole-plot is divided into q split-plots. The p levels of A are randomly allocated to whole-plots and q levels of B are randomly allocated to q split-plots of a whole-plot. The process of randomisation is replicated r times. Then the resultant design is known as split-plot design.

There are two different steps of randomisation in split-plot design and hence, it may be considered as the combination of more than one randomized block design. If the observations of split-plots are added, we have the data of randomized block design, where p treatments are allocated randomly in plots of r blocks. Therefore, the observations of whole-plot treatment (A) replicated r times can be analysed in a similar way as it is done for the data of randomized block design. This analysis is usually known as *whole-plot analysis* or *whole-plot comparison* and during this analysis we get one error sum of squares which is usually known as *whole-plot error* or *error-1*. Again, adding the observations of a treatment combination for r replicates, we get the two-way classified data for factor A and factor B . These two-way data can also be analysed similar to the analysis of data of randomized block design having r observations per cell. This second step of analysis also gives one error sum of squares which is usually known as *sub-plot error* or *error-2*.

The levels of A and B may be same. In that case the experiment is symmetrical factorial experiment. If levels of A and B are different, the experiment is asymmetrical factorial experiment. Therefore, the split-plot experiment is a factorial experiment. The arrangement of treatments in whole-plots and split-plots of a blocks can be shown as follows :

		Whole-plots ($W_i, i = 1, 2, \dots, p$)																		
		a_1				a_2				...				a_p						
Block or replication		b_1	b_2	...	b_q	b_1	b_2	...	b_q			...						b_2	...	b_q
	Total of whole-plots	A_1				A_2				...				A_p						

Here $A_1 - A_2$ is a contrast of A . It is also a contrast of two whole-plots, viz. $W_1 - W_2$. Hence, contrast of A (whole-plot factor) is entangled with contrast of whole-plots. That is, whole-plot

factor is confounded with whole plots and hence, split-plot experiment may be considered as confounded factorial experiment. Again, the p -levels of A may be the combinations of several factors and q levels of B may also be the combinations of another group of factors. If the level combinations in both cases are more, the treatments in whole-plots and sub-plots may be allocated in a similar way as it is done in confounded factorial experiment.

The whole-plot treatments may be allocated in whole-plots following the technique of randomized block design or latin square design. Accordingly, the name of the design is split-plot design using randomized block design technique or split-plot design using Latin square design technique.

Advantages and Disadvantages of Split-plot Design

Advantages :

- (i) Since the whole-plot factor needs larger experimental materials, these can be sub-divided into several parts, and levels of second factor can be allocated to the sub-plots. The effect of second factor is also estimated simultaneously with first factor.
- (ii) There are pq plots per block. But due to the use of sub-plots in a whole-plot, the design is more efficient than the randomized block design having pq plots. The efficiency is observed in estimating sub-plot factor effect and interaction of sub-plot and whole-plot factors.
- (iii) The efficiency of split-plot design increases compared to the randomized block design if whole-plot treatment is allocated in whole-plots using Latin square design technique or incomplete Latin square design technique.

Disadvantages :

- (i) Since whole-plot factor is confounded with whole-plots, the effect of whole-plot factor is estimated with less efficiency.
- (ii) The data of split-plot experiment become non-orthogonal since the observations of sub-plots within a whole-plot are correlated.
- (iii) The analysis of data of split-plot design becomes complicated if there are missing observations.

Analysis of split-plot design : Let y_{ijl} be the observation of i -th level of B in j -th whole-plot of i -th replication. The model for y_{ijl} ($i = 1, 2, \dots, r; j = 1, 2, \dots, p; l = 1, 2, \dots, q$) observation is

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l + (\beta\gamma)_{jl} + e_{ijl},$$

where μ = general mean, α_i = effect of i -th block (replication), β_j = effect of j -th level of A , γ_l = effect of l -th level of B , $(\beta\gamma)_{jl}$ = interaction of j -th level of A with l -th level of B , e_{ijl} = random error.

The assumption for analysis of the data is

- (i) $E(e_{ijl}) = 0$, (ii) $E[e_{ijl}, e_{i'j'l'}] = 0$, if $i \neq i', j \neq j', l \neq l'$.
 $= \rho\sigma^2$, if $i = i', j = j', l \neq l'$
 $= \sigma^2$, if $i = i', j = j', l = l'$.

To analyse the data the following restrictions can be imposed :

$$\sum \alpha_i = \sum \beta_j = \sum \gamma_l = \sum_j (\beta\gamma)_{jl} = \sum_l (\beta\gamma)_{jl} = 0.$$

Since the observations of a whole-plot in any replication are non-orthogonal, an orthogonal transformation is needed to the vector of q observations of j -th whole-plot in i -th block. This is done as follows :

$$\begin{bmatrix} U_{ij} \\ Z_{ij1} \\ Z_{ij2} \\ \vdots \\ Z_{ijq-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{q}} & \frac{1}{\sqrt{q}} & \cdots & \frac{1}{\sqrt{q}} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ a_{31} & a_{32} & \cdots & a_{3q} \\ \dots & \dots & \dots & \dots \\ a_{q-1,1} & a_{q-1,2} & \cdots & a_{q-1,q} \end{bmatrix} \begin{bmatrix} y_{ij1} \\ y_{ij2} \\ \vdots \\ y_{ijq} \end{bmatrix},$$

where $\sum_l a_{l'l} = 0$, $\sum_l a_{l'l}^2 = 1$, $\sum_l a_{l'l} a_{l''l} = 0$, $l' \neq l'' = 1, 2, \dots, q-1$

$$\sum_{l'=1}^{q-1} a_{l'l}^2 = 1 - \frac{1}{q}, \quad \sum_{l'=1}^{q-1} \sum_{l''=1}^{q-1} a_{l'l} a_{l''l} = -\frac{1}{q}, \quad l \neq k = 1, 2, \dots, q.$$

We have
$$U_{ij} = \frac{1}{\sqrt{q}} \sum_l y_{ijl} = \frac{1}{\sqrt{q}} \sum_l (\mu + \alpha_i + \beta_j + \gamma_l + (\beta\gamma)_{jl} + e_{ijl})$$

$$= \sqrt{q}(\mu + \alpha_i + \beta_j) + \delta_{ij}, \quad \text{where } \delta_{ij} = \frac{1}{\sqrt{q}} e_{ij}.$$

Thus, the estimates of μ , α_i and β_j are to be obtained from the observations U_{ij} ($i = 1, 2, \dots, r$; $j = 1, 2, \dots, p$).

Here
$$V(\delta_{ij}) = V(U_{ij}) = \frac{1}{q} V(y_{ij1} + y_{ij2} + \cdots + y_{ijq})$$

$$= \frac{1}{q} \left[\sum_l V(y_{ijl}) + \sum_{l \neq l'} \sum_{l''}^q \text{Cov}(y_{ijl}, y_{ijl''}) \right]$$

$$= \sigma^2 [1 + (q-1)\rho].$$

This is the variance of first kind of error.

Again,
$$Z_{ijl'} = a_{l'1}y_{ij1} + a_{l'2}y_{ij2} + \cdots + a_{l'q}y_{ijq}, \quad l' \neq l = 1, 2, \dots, q$$

$$= \sum_{l=1}^q a_{l'l} [\gamma_l + (\beta\gamma)_{jl}] + \epsilon_{ijl'} \quad \text{where } \epsilon_{ijl'} = \sum_{l=1}^q a_{l'l} e_{ijl}.$$

The estimates of γ_l and $(\beta\gamma)_{jl}$ are to be estimated from $Z_{ijl'}$ observations.

Here
$$V(\epsilon_{ijl'}) = V[a_{l'1}y_{ij1} + a_{l'2}y_{ij2} + \cdots + a_{l'q}y_{ijq}]$$

$$= \sum_l a_{l'l}^2 V(y_{ijl}) + \sum_{l \neq k} \sum_{l''} a_{l'l} a_{l''l} \text{Cov}(y_{ijl}, y_{ijl''}) = \sigma^2 (1 - \rho).$$

This is the variance of the second kind of error.

Let $w_1 = 1/\sigma^2 [1 + (q-1)\rho]$, $w_2 = 1/\sigma^2 (1 - \rho)$. Since two-error variances are not similar, weighted least squares method is to be applied to estimate the parameters. The weighted estimated error sum of squares is

$$\phi = w_1 \sum \sum \sum [U_{ij} - \sqrt{q}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)]^2 + w_2 \sum \sum \sum [Z_{ijl'} - \sum_l a_{l'l}(\hat{\gamma}_l + (\hat{\beta}\gamma)_{jl})]^2.$$

The normal equations are

$$\frac{\partial \phi}{\partial \hat{\mu}} = 0, \quad \frac{\partial \phi}{\partial \hat{\alpha}_i} = 0, \quad \frac{\partial \phi}{\partial \hat{\beta}_j} = 0, \quad \frac{\partial \phi}{\partial \hat{\gamma}_l} = 0 \quad \text{and} \quad \frac{\partial \phi}{\partial (\hat{\beta\gamma})_{jl}} = 0.$$

Solving these equations, we get

$$\hat{\mu} = \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad \hat{\gamma}_l = \bar{y}_{.l.} - \bar{y}_{...}$$

$$(\hat{\beta\gamma})_{jl} = \bar{y}_{.jl} - \bar{y}_{.j.} - \bar{y}_{.l.} + \bar{y}_{...}$$

It is observed that the estimates of parameters do not depend on the error variances. But the estimates are obtained from U_{ij} observations ($i = 1, 2, \dots, r; j = 1, 2, \dots, p$), where U_{ij} observations are obtained adding the observations of sub-plots of a whole-plot. Therefore, the sum of squares due to $\hat{\mu}$, and $\hat{\alpha}_i$ are to be found out from y_{ij} observations. Hence, we have

$$q \sum \sum (\bar{y}_{ij.} - \bar{y}_{...})^2 = pq \sum (\bar{y}_{i..} - \bar{y}_{...})^2 + qr \sum (\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$+ q \sum \sum (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$= SS(\hat{\alpha}_i) + SS(\hat{\beta}_j) + SS(\text{Error-1}) = S_1 + S_2 + S_3.$$

The second step of analysis is to be performed using the observations $Z_{ijl'}$ ($i = 1, 2, \dots, r; j = 1, 2, \dots, p; l' = 1, 2, \dots, q-1$) However, on simplification, we have

$$S_4 = SS(\hat{\gamma}_l) = pr \sum (\bar{y}_{.l.} - \bar{y}_{...})^2,$$

$$S_5 = SS(\beta\hat{\gamma}_{jl}) = r \sum \sum (\bar{y}_{.jl} - \bar{y}_{.j.} - \bar{y}_{.l.} + \bar{y}_{...})^2$$

$$S_6 = SS(\text{Error-2}) = \sum \sum \sum (y_{ijl} - \bar{y}_{.jl} - \bar{y}_{ij.} + \bar{y}_{.j.})^2$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	$E(MS)$	F
Replications	$r - 1$	S_1	s_1	$\sigma_1^2 + \frac{pq}{r-1} \sum \alpha_i^2$	$F_1 = s_1/s_3$
A	$p - 1$	S_2	s_2	$\sigma_1^2 + \frac{qr}{p-1} \sum \beta_j^2$	$F_2 = s_2/s_3$
Error-1	$(r - 1)(p - 1)$	S_3	s_3	σ_1^2	—
B	$q - 1$	S_4	s_4	$\sigma_2^2 + \frac{pr}{q-1} \sum \gamma_l^2$	$F_3 = s_4/s_6$
AB	$(p - 1)(q - 1)$	S_5	s_5	$\sigma_2^2 + \frac{r}{(p-1)(q-1)} \sum \sum (\beta\gamma)_{jl}^2$	$F_4 = s_5/s_6$
Error-2	$p(r - 1)(q - 1)$	S_6	s_6	σ_2^2	—
Total	$pqr - 1$				

Here $\sigma_1^2 = \sigma^2[1 + (q - 1)\rho]$, $\sigma_2^2 = \sigma^2(1 - \rho)$; $E(s_3) = \sigma_1^2$, $E(s_6) = \sigma_2^2$. The estimate of ρ is $\hat{\rho} = \frac{s_3 - s_6}{s_3 + (q - 1)s_6}$. It is known that ρ is positive. Therefore, $s_3 > s_6$.

The main objective of this analysis is to test the significance of the hypotheses

- (i) $H_0 : \beta_j = 0$, against $H_A : \beta_j \neq 0$;
- (ii) $H_0 : \gamma_l = 0$, against $H_A : \gamma_l \neq 0$ and
- (iii) $H_0 : (\beta\gamma)_{jl} = 0$, against $H_A : (\beta\gamma)_{jl} \neq 0$.

The test statistic for H_0 (i) is F_2 , where the null-distribution of F_2 is variance ratio F and the non-null distribution of F_2 is non-central F with non-centrality parameter

$$\lambda_2 = \frac{qr}{2\sigma_1^2} \sum \beta_j^2.$$

The d.f. of F_2 are $(p-1)$ and $(p-1)(r-1)$. Therefore, $F_2 \geq F_{0.05;(p-1),(p-1)(r-1)}$ leads us to reject the null hypothesis (i).

The test statistic for H_0 (ii) is F_3 . It has $(q-1)$ and $p(r-1)(q-1)$ d.f. Under H_0 (ii) F_3 is distributed as variance ratio. Therefore, if $F_3 \geq F_{0.05;(q-1),p(r-1)(q-1)}$, H_0 is rejected. The non-null distribution of F_3 is non-central F with non-centrality parameter

$$\lambda_3 = \frac{pr}{2\sigma_2^2} \sum \gamma_l^2.$$

The test statistic for H_0 (iii) is F_4 . It has $(p-1)(q-1)$ and $p(r-1)(q-1)$ d.f. The null distribution of F_4 is variance ratio distribution. Hence, if $F_4 \geq F_{0.05;(p-1)(q-1),p(r-1)(q-1)}$, H_0 is rejected. The non-null distribution of F_4 is non-central with non-centrality parameter

$$\lambda_4 = \frac{r}{2\sigma_2^2} \sum_j \sum_l (\beta\gamma)_{jl}^2.$$

If the hypothesis $H_0 : \beta_j = 0$ is rejected, we need to compare the pairs of all whole-plot treatments ($A_j, j = 1, 2, \dots, p$). The hypothesis for this is

$$H_0 : \beta_j = \beta_{j'}, \text{ against } H_A : \beta_j \neq \beta_{j'}, \text{ for all } j \neq j' = 1, 2, \dots, p.$$

If the comparison is planned for a particular pair, the test statistic is

$$t_1 = \frac{\bar{y}_{.j} - \bar{y}_{.j'}}{\sqrt{\frac{2s_3}{qr}}},$$

where t_1 is distributed as Student's t with $(p-1)(r-1)$ d.f. The calculated value of $t_1 \geq t_{0.025,(p-1)(r-1)}$ leads us to reject the null hypothesis.

The comparison of all pairs is done by Duncan's multiple range test, where the test statistic is

$$D_k = d_{0.05,k,f} \sqrt{\frac{s_3}{qr}}, \quad k = 2, 3, \dots, p; \quad f = (p-1)(r-1).$$

At this stage of analysis one may need to compare any two levels of whole-plot treatment in which a particular level of sub-plot treatment is used. The hypothesis for this is

$$H_0 : \beta_j^l = \beta_{j'}^l, \text{ against } H_A : \beta_j^l \neq \beta_{j'}^l, \quad (j \neq j' = 1, 2, \dots, p).$$

Here β_j^l = effect of j -th level of A in presence of l -th level of B . The test statistic for this hypothesis is

$$F_5 = \frac{r(\bar{y}_{.jl} - \bar{y}_{.j'l})^2}{2S^2}, \text{ where } S^2 = \sum_i \sum_j (\bar{y}_{ijl} - \bar{y}_{.jl} - \bar{y}_{.il} + \bar{y}_{..l})^2 / (r-1)(p-1).$$

This S^2/σ^2 is distributed as central χ^2 with $(r-1)(p-1)$ d.f. and under $H_0(\bar{y}_{.jl} - \bar{y}_{.j'l})^2 / \frac{2\sigma^2}{r}$ is distributed as χ^2 with 1 d.f. Therefore, F_5 is distributed as variance ratio with 1 and $(r-1)(p-1)$ d.f. The non-null distribution of F is non-central with non-centrality parameter :

$$\lambda_5 = \frac{1}{2\sigma_1^2} [\beta_j - \beta_{j'} + (\beta\gamma)_{jl} - (\beta\gamma)_{j'l}]^2.$$

The test statistic for the hypothesis $H_0 : \beta_j^l = \beta_{j'}^{l'}, \quad j \neq j' = 1, 2, \dots, p; \quad l \neq l' = 1, 2, \dots, q$ is

$$F_6 = \frac{r(\bar{y}_{ijl} - \bar{y}_{.j'l'})^2}{2S^2}.$$

The distribution of F_5 and F_6 is same except the non-centrality parameter.

The Duncan's multiple range test to compare the means of $\bar{y}_{.jl}$ and $\bar{y}_{.j'l}$ or the means of $\bar{y}_{.jl}$ and $\bar{y}_{.j'l'}$ ($j \neq j' = 1, 2, \dots, p; l \neq l' = 1, 2, \dots, q$) is given by

$$D_k = d_{0.05, k, f_1} \sqrt{\frac{S^2}{r}}; \quad k = 2, 3, \dots, p; \quad f_1 = (r - 1)(p - 1).$$

The rejection of H_0 (ii) leads us to test the significance of

$$H_0 : \gamma_l = \gamma_{l'} \text{ against } H_A : \gamma_l \neq \gamma_{l'}; \quad l \neq l' = 1, 2, \dots, q.$$

The test statistic for this hypothesis is

$$t_2 = \frac{\bar{y}_{.l} - \bar{y}_{.l'}}{\sqrt{\frac{2s_6}{pr}}}.$$

This t_2 follows Student's t distribution with $p(r - 1)(q - 1)$ d.f. For multiple comparison of the means $\bar{y}_{.l}$ ($l = 1, 2, \dots, q$) the test statistic is

$$D_k = d_{0.05, k, f_2} \sqrt{\frac{s_6}{pr}}; \quad k = 2, 3, \dots, q; \quad f_2 = p(r - 1)(q - 1).$$

The test statistic to compare the means of sub-plot treatment at any particular level of whole-plot treatment is

$$t = \frac{\bar{y}_{.jl} - \bar{y}_{.j'l'}}{\sqrt{\frac{2s_6}{r}}}.$$

Here also Duncan's multiple range test can be performed to compare all pairs of means at a particular level of whole-plot treatment.

Efficiency of split-plot design : The analysis of variance table for split-plot experiment can be rewritten as follows :

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$
Replications	$r - 1$	S_1	s_1
A	$p - 1$	S_2	s_2
Error-1	$(p - 1)(r - 1)$	S_3	s_3
B	$q - 1$	S_4	s_4
AB	$(p - 1)(q - 1)$	S_5	s_5
Error-2	$p(r - 1)(q - 1)$	S_6	s_6
Total	$pqr - 1$		

Consider that the effects of A, B and interaction of AB are insignificant and the effect of A is, on an average, equals the error variance-1. The impacts of B and AB equal the error variance-2. Then the analysis of variance table takes the following shape :

ANOVA Table

Sources of variation	d.f.	SS
Replications	$r - 1$	$(r - 1)s_1$
Error-1	$r(p - 1)$	$r(p - 1)s_3$
Error-2	$pr(q - 1)$	$pr(q - 1)s_6$
Total	$pqr - 1$	

The total error sum of squares is $r(p-1)s_3 + pr(q-1)s_6$. It has $r(pq-1)$ d.f. The above analysis of variance table looks like the analysis of variance table of randomized block experiment with pq treatments, where treatment effects are insignificant. Therefore, the efficiency of split-plot design compared to randomized block design is

$$\frac{(p-1)s_3 + p(q-1)s_6}{(pr-1)s_6}$$

Example 4.9 : In an agricultural research station an experiment is conducted to study the productivity of a maize variety using 4 doses of nitrogen fertilizer under three different levels of irrigation. The levels of irrigation are daily one time irrigation (I_1) two times irrigation (I_2) and alternate day two times irrigation (I_3). The nitrogen fertilizer as urea at the rate of 90 kg/ha (N_1), 120 kg/ha (N_2), 150 kg/ha (N_3) and 180 kg/ha (N_4) are used in the experiment. The design used is split-plot, where whole-plot treatment is irrigation and sub-plot treatment is nitrogen. The experiment is replicated three times. The productions of maize (kg/plot) are shown below :

Replication	whole-plot treatment	Sub-plot treatment				Total y_{ij}	Total $y_{i..}$
		N_1	N_2	N_3	N_4		
1	I_1	1.5	1.6	1.8	1.8	6.7	22.7
	I_2	1.8	2.0	2.0	2.4	8.2	
	I_3	2.0	1.8	1.8	2.2	7.8	
2	I_1	1.6	1.8	1.8	2.0	7.2	25.0
	I_2	2.0	2.2	2.4	2.2	8.8	
	I_3	2.0	2.2	2.2	2.6	9.0	
3	I_1	1.8	2.0	2.0	2.4	8.2	26.3
	I_2	2.2	2.2	2.3	2.4	9.1	
	I_3	2.4	2.0	2.0	2.6	9.0	
	Total $y_{.i}$	17.3	17.8	18.3	20.6		74.0

- (i) Analyse the data and comment on the use of irrigation and nitrogen.
- (ii) Compare the levels of irrigation and nitrogen, if possible.
- (iii) Compare I_1 and I_2 in presence of N_4 .
- (iv) compare N_3 and N_4 in presence of I_3 .
- (v) Find the efficiency of the design compared to randomized block design.

Solution : (i) We have $p = 3$, $q = 4$, $r = 3$, $G = 74.0$

$$C.T. = \frac{G^2}{pqr} = \frac{(74.0)^2}{3 \times 4 \times 3} = 152.1111.$$

$$SS \text{ (Total)} = \sum \sum \sum y_{ijl}^2 - C.T. = 154.78 - 152.1111 = 2.6689.$$

$$SS \text{ (Replications)} = \sum \frac{y_{i..}^2}{qp} - C.T. = \frac{1831.98}{4 \times 3} - 152.1111 = 0.5539.$$

Productions of irrigation and nitrogen ($y_{.jl}$)

Irrigation	Nitrogen				Total $y_{.j}$
	N_1	N_2	N_3	N_4	
I_1	4.9	5.4	5.6	6.2	22.1
I_2	6.0	6.4	6.7	7.0	26.1
I_3	6.4	6.0	6.0	7.4	25.8

$$SS(\text{Irrigation}) = \sum \frac{y_{.j}^2}{qr} - \text{C.T.} = \frac{1835.26}{4 \times 3} - 152.1111 = 0.8272.$$

$$SS(\text{Error-1}) = \sum \sum \frac{y_{ij}^2}{q} - \text{C.T.} - SS(\text{Irrigation}) - SS(\text{Replications})$$

$$= \frac{614.3}{4} - 152.1111 - 0.8272 - 0.5539 = 0.0828.$$

$$SS(\text{Nitrogen}) = \sum \frac{y_{.l}^2}{pr} - \text{C.T.} = \frac{1375.38}{3 \times 3} - 152.1111 = 0.7089.$$

$$SS(\text{Irrigation} \times \text{Nitrogen}) = \sum \sum \frac{y_{.jl}^2}{r} - \text{C.T.} - SS(\text{Irrigation}) - SS(\text{Nitrogen})$$

$$= \frac{461.54}{3} - 152.1111 - 0.8272 - 0.7089 = 0.1995.$$

$$SS(\text{Error-2}) = SS(\text{Total}) - SS(\text{Replications}) - SS(\text{Irrigation}) - SS(\text{Error-1})$$

$$- SS(\text{Nitrogen}) - SS(\text{Irrigation} \times \text{Nitrogen})$$

$$= 2.6689 - 0.5529 - 0.8272 - 0.0828 - 0.7089 - 0.1995 = 0.2966.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{.05}$	P-value
Replications	2	0.5539	0.27695	13.38	6.94	< 0.05
Irrigation (A)	2	0.8272	0.4136	19.98	"	0.00
Error-1	4	0.0828	0.0207	—	—	—
Nitrogen (B)	3	0.7089	0.2363	14.34	3.16	0.00
Irrigation \times Nitrogen (AB)	6	0.1995	0.03325	2.02	2.66	> 0.05
Error-2	18	0.2966	0.01648	—	—	—
Total	35					

The levels of irrigation are highly significant. The production of maize increases with the increase in the levels of irrigation. Similar is the case with the levels of nitrogen.

(ii) The levels of irrigation can be compared by Duncan's multiple range test, where the test statistic

$$D_k = d_{0.05, k, f_1} \sqrt{\frac{s_3}{qr}}, \quad k = 2, 3; \quad f_1 = 4$$

$$D_2 = 3.93 \sqrt{\frac{0.0207}{4 \times 3}} = 0.163, \quad D_3 = 4.01 \sqrt{\frac{0.0207}{4 \times 3}} = 0.166.$$

The means due to irrigation are in ascending order as follows :

$$\bar{I}_1 = 1.842, \bar{I}_3 = 2.15, \bar{I}_2 = 2.175$$

$$\bar{I}_2 - \bar{I}_1 = 2.175 - 1.842 = 0.333 > D_3, \therefore \text{the means are significantly different.}$$

$$\bar{I}_3 - \bar{I}_1 = 2.15 - 1.842 = 0.308 > D_2, I_1 \text{ and } I_3 \text{ are different.}$$

$$\bar{I}_2 - \bar{I}_3 = 2.175 - 2.15 = 0.025 < D_2, I_2 \text{ and } I_3 \text{ are similar.}$$

To compare the means of levels of nitrogen the test statistic is

$$D_k = d_{0.05, k, f_2} \sqrt{\frac{s_6}{pr}}, \quad k = 2, 3, 4; \quad f_2 = 18.$$

$$D_2 = 2.97 \sqrt{\frac{0.01648}{3 \times 3}} = 0.127, \quad D_3 = 3.12 \sqrt{\frac{0.01648}{3 \times 3}} = 0.133,$$

$$D_4 = 3.21 \sqrt{\frac{0.01648}{3 \times 3}} = 0.137.$$

The means due to the levels of nitrogen are in ascending order as follows :

$$\bar{N}_1 = 1.922, \bar{N}_2 = 1.978, \bar{N}_3 = 2.033, \bar{N}_4 = 2.289.$$

$$\bar{N}_4 - \bar{N}_1 = 2.289 - 1.922 = 0.367 > D_4, \text{ the means differ significantly.}$$

$$\bar{N}_4 - \bar{N}_2 = 2.289 - 1.978 = 0.32 > D_3, N_2 \text{ and } N_4 \text{ are different.}$$

$$\bar{N}_3 - \bar{N}_1 = 2.033 - 1.922 = 0.111 < D_3, N_1 \text{ and } N_3 \text{ are similar.}$$

$$\bar{N}_3 - \bar{N}_2 = 2.033 - 1.978 = 0.055 < D_2, N_2 \text{ and } N_3 \text{ are similar.}$$

$$\bar{N}_4 - \bar{N}_3 = 2.289 - 2.033 = 0.256 > D_2, N_3 \text{ and } N_4 \text{ are different.}$$

The underlined means are similar.

$$\bar{I}_1, \bar{I}_3, \bar{I}_2; \quad \bar{N}_1, \bar{N}_2, \bar{N}_3, \bar{N}_4$$

(iii) We need to test $H_0 : \beta_1^4 = \beta_2^4$ against $H_A : \beta_1^4 \neq \beta_2^4$. Here β_j^4 = effect of j -th level of irrigation in presence of 4-th level of nitrogen ($j = 1, 2$).

The production of maize using N_4 are shown below :

Replications	Irrigation (y_{ij4})			Total $y_{i.4}$
	I_1	I_2	I_3	
1	1.8	2.4	2.2	6.4
2	2.0	2.2	2.6	6.8
3	2.4	2.4	2.6	7.4
Total $y_{.j4}$	6.2	7.0	7.4	20.6 = G_4

$$C.T_4 = \frac{C^2}{pr} = \frac{(20.6)^2}{3 \times 3} = 47.151.$$

$$SS(\text{Total})_4 = \sum \sum y_{ij4}^2 - C.T_4 = 47.72 - 47.151 = 0.569.$$

$$SS(\text{Replications})_4 = \frac{\sum y_{i.4}^2}{p} - C.T_4 = \frac{141.96}{3} - 47.151 = 0.169.$$

$$SS(\text{Irrigation})_4 = \frac{\sum y_{j4}^2}{r} - C.T_4 = \frac{142.2}{3} - 47.151 = 0.249.$$

$$S^2 = [SS(\text{Total})_4 - SS(\text{Replications})_4 - SS(\text{Irrigation})_4] / (p-1)(r-1) \\ = \frac{1}{4}[0.569 - 0.169 - 0.249] = 0.03775.$$

The test statistic is

$$F = \frac{r(\bar{y}_{.14} - \bar{y}_{.24})^2}{2S^2} = \frac{3(2.067 - 2.333)^2}{2 \times 0.03775} = 2.81.$$

$F < F_{0.05;1,4} = 7.71$, H_0 is accepted. Therefore, the irrigation levels 1 and 2 in presence of N_4 are similar.

(iv) We need to test $H_0 : \gamma_3^3 = \gamma_4^3$ against $H_A : \gamma_3^3 \neq \gamma_4^3$. Here γ_i^3 = effect of nitrogen in presence of $I_3(l = 3, 4)$.

The test statistic is

$$t = \frac{\bar{y}_{.33} - \bar{y}_{.34}}{\sqrt{\frac{2S_6}{r}}} = \frac{2.0 - 2.467}{\sqrt{\frac{2 \times 0.01648}{3}}} = -4.45.$$

$|t| > t_{0.025,18} = 2.101$, H_0 is rejected. The levels of nitrogen N_3 and N_4 are different in presence of I_3 .

(v) The efficiency of split-plot design compared to randomized block design is

$$\frac{(p-1)s_3 + p(q-1)s_6}{(pq-1)s_6} = \frac{(3-1)0.0207 + 3(4-1)0.01648}{(3 \times 4 - 1)0.01648} = 104.65\%.$$

4.12 Estimation of Missing Value in Case of Split-Plot Design

Let the observation of l -th sub-plot treatment in j -th whole-plot of i -th replication be missing. Let this observation be x . The missing value x is to be estimated in such a way that the sum of squares due to error-2 is minimum.

This sum of squares in presence of x is written as

$$Q = SS(\text{Error-2}) = \sum_{i'}^r \sum_{j'}^p \sum_{l'}^q y_{i'j'l'}^2 + x^2 - \frac{1}{q} \sum \sum y_{i'j'}^2 - \frac{1}{q}(y_{i.j.} + x)^2 \\ - \frac{1}{r} \sum \sum y_{j'l'}^2 - \frac{1}{r}(y_{.jl} + x)^2 + \frac{1}{qr} \sum y_{j'}^2 + \frac{1}{qr}(y_{.j.} + x)^2,$$

where $i \neq i', j \neq j', l \neq l'$.

The value of x is to be found out in such a way that $\frac{\partial \phi}{\partial x} = 0$. Thus, we have

$$x = \frac{ry_{ij.} + qy_{.jl} - y_{.j.}}{(q-1)(r-1)}.$$

The analysis of data of split-plot experiment is performed in a usual manner after replacing the missing value by the estimated value of x . However, one degree of freedom is subtracted from the total d.f. and hence, from d.f. of error-2 if there is one missing observation.

The variances of the differences in two means are calculated as follows :

$$V(\bar{y}_{.j.} - \bar{y}_{.j'.}) = \frac{2}{qr}[s_3 + ks_6], \quad V(\bar{y}_{.l.} - \bar{y}_{.l'.}) = \frac{2s_6}{pr} \left[1 + \frac{kq}{p} \right]$$

$$V(\bar{y}_{.jl} - \bar{y}_{.j'l'}) = \frac{2s_6}{r} \left[1 + \frac{kq}{p} \right]$$

$$V(\bar{y}_{.jl} - \bar{y}_{.j'l'}) = V(\bar{y}_{.jl} - \bar{y}_{.j'l'}) = \frac{2s_3}{qr} + \frac{2s_6}{qr} [(q-1) + kq^2]$$

where $k = 1/2(r-1)(q-1)$, if there is one missing observation.

The value of k depends on number of missing observations and their positions. G. S. Watson has proposed the value of k , where $k = m/2(r-d)(q-m+c-1)$. Here m = number of missing observations, C = number of replications of one or more missing observations, d = number of missing observations of sub-plot treatment. In calculating these values the missing observations related to the means to be compared are considered.

4.13 Split-Split-Plot Design

In split-plot design the entire experimental materials for a replication are divided into several whole-plots and each whole-plot is sub-divided into several sub-plots. In some instances the sub-plots are again divided into several split-split-plots so that p levels of a factor A can be randomly allocated to p whole-plots, q levels of sub-plot factor B can be allocated randomly to q sub-plots and m levels of sub-sub-plot factor C can be randomly allocated to m split-split-plots. The process of randomization is replicated r times in r blocks. The resultant design is known as split-split-plot design.

Let y_{ijkl} be the observation of l -th level of a factor C corresponding to the k -th level of a factor B and j -th level of another factor A in i -th replication [$i = 1, 2, \dots, r; j = 1, 2, \dots, p; k = 1, 2, \dots, q; l = 1, 2, \dots, m$]. The linear model for this observation is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + \delta_l + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + (\beta\gamma\delta)_{jkl} + e_{ijkl},$$

where μ = general mean, α_i = effect of i -th replication, β_j = effect of j -th level of A , γ_k = effect of k -th level of B , $(\beta\gamma)_{jk}$ = interaction of j -th level of A with k -th level of B , δ_l = effect of l -th level of C , $(\beta\delta)_{jl}$ = interaction of j -th level of A with l -th level of C , $(\gamma\delta)_{kl}$ = interaction of k -th level of B with l -th level of C , $(\beta\gamma\delta)_{jkl}$ = interaction of j -th level of A with k -th level of B and l -th level of C , e_{ijkl} = random error.

Let us consider that the model is a fixed effect model. Therefore, for the analytical purpose one can put the following restrictions :

$$\begin{aligned} \sum \alpha_i &= \sum \beta_j = \sum \gamma_k = \sum \delta_l = \sum_j (\beta\gamma)_{jk} = \sum_k (\beta\gamma)_{jk} = \sum_j (\beta\delta)_{jl} = \sum_l (\beta\delta)_{jl} \\ &= \sum_k (\gamma\delta)_{kl} = \sum_l (\gamma\delta)_{kl} = \sum_j (\beta\gamma\delta)_{jkl} = \sum_k (\beta\gamma\delta)_{jkl} = \sum_l (\beta\gamma\delta)_{jkl} = 0. \end{aligned}$$

The assumption for the analysis is

$$\begin{aligned} E(e_{ijkl}) &= 0 \\ E(e_{ijkl}, e_{i'j'k'l'}) &= \sigma^2, \text{ if } i = i', j = j', k = k', l = l' \\ &= \rho_1 \sigma^2, \text{ if } i = i', j = j', k = k', l \neq l' \\ &= \rho_2 \sigma^2, \text{ if } i = i', j = j', k \neq k' \\ &= 0, \text{ otherwise.} \end{aligned}$$

The above assumptions on random component indicate that the errors and hence the observation vector is non-orthogonal. To perform the theoretical analysis, we need to perform

orthogonal transformation to observation vector. The non-orthogonality of data arises due to allocation of levels of sub-sub-plot factor to the levels of a sub-plot factor and also to the levels of a whole-plot factor. Therefore, we need two separate orthogonal transformations.

Let us first consider the following orthogonal transformation :

$$\begin{bmatrix} U_{ijk} \\ Z_{ijk1} \\ Z_{ijk2} \\ \vdots \\ Z_{ijk(m-1)} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{m} & 1/\sqrt{m} & \cdots & 1/\sqrt{m} \\ a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{(m-1)1} & a_{(m-1)2} & \cdots & a_{(m-1)m} \end{bmatrix} \begin{bmatrix} y_{ijk1} \\ y_{ijk2} \\ \vdots \\ y_{ijkm} \end{bmatrix}.$$

Here $\sum_l a_{ul} = 0, \sum_l a_{ul}^2 = 1, \sum_l a_{ul}a_{vl} = 0, u \neq v = 1, 2, \dots, (m-1)$

$$\sum_{u=1}^{m-1} a_{ul}^2 = 1 - \frac{1}{m}, \sum_u a_{ul}a_{ul'} = -\frac{1}{m}, l \neq l' = 1, 2, \dots, m.$$

This orthogonal transformation provides

$$U_{ijk} = \frac{1}{\sqrt{m}}y_{ijk} = \sqrt{m}(\mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk}) + \Pi_{ijk}, \text{ where } \Pi_{ijk} = \frac{1}{\sqrt{m}}e_{ijk}.$$

We have $\text{Cov}[U_{ijk}, U_{ijk'}] = \text{Cov}\left[\frac{1}{\sqrt{m}}y_{ijk}, \frac{1}{\sqrt{m}}y_{ijk'}\right]$
 $= \frac{1}{m}[\rho_2\sigma^2 + \rho_2\sigma^2 + \cdots + \rho_2\sigma^2] = m\rho_2\sigma^2$

and $V(U_{ijk}) = \frac{1}{m}V(y_{ijk1} + y_{ijk2} + \cdots + y_{ijkm})$
 $= \frac{1}{m}[m\sigma^2 + m(m-1)\rho_1\sigma^2] = \sigma^2[1 + (m-1)\rho_1].$

It is observed that the parameters $\mu_1, \alpha_i, \beta_j, \gamma_k$ and $(\beta\gamma)_{jk}$ are to be estimated from the observations $U_{ijk} (i = 1, 2, \dots, r; j = 1, 2, \dots, p; k = 1, 2, \dots, q)$. In practice, these observations are obtained by adding the observations of split-split-plots. Again, these observations are non-orthogonal. Hence, another orthogonal transformation is needed at least for the theorital aspect of the analysis. Let this orthogonal transformation be as follows :

$$\begin{bmatrix} W_{ij} \\ M_{ij1} \\ M_{ij2} \\ \vdots \\ M_{ij(q-1)} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{q} & 1/\sqrt{q} & \cdots & 1/\sqrt{q} \\ b_{11} & b_{12} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2q} \\ \dots & \dots & \dots & \dots \\ b_{(q-1)1} & b_{(q-1)2} & \cdots & b_{(q-1)q} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{m}}y_{ij1} \\ \frac{1}{\sqrt{m}}y_{ij2} \\ \vdots \\ \frac{1}{\sqrt{m}}y_{ijq} \end{bmatrix}$$

Here $\sum_k b_{hk} = 0, \sum_k b_{hk}^2 = 1, \sum_k b_{hk}b_{h'k} = 0 (h \neq h' = 1, 2, \dots, q-1)$

$$\sum_{h=1}^{q-1} b_{hk}^2 = 1 - \frac{1}{q}, \sum_{h=1}^{q-1} b_{hk}b_{hk'} = -\frac{1}{q}, k \neq k' = 1, 2, \dots, q.$$

We have $W_{ij} = \frac{1}{\sqrt{mq}}y_{ij..} = \sqrt{mq}(\mu + \alpha_i + \beta_j) + \varphi_{ij}, \text{ where } \varphi_{ij} = \frac{1}{\sqrt{mq}}e_{ij..}$

where $V(\varphi_{ij}) = \sigma^2[1 + (m-1)\rho_1 + m(q-1)\rho_2].$

From the first orthogonal transformation, we have

$$\begin{aligned} z_{ijk_u} &= a_{u1}y_{ijk1} + a_{u2}y_{ijk2} + \cdots + a_{um}y_{ijk_m} \\ &= \sum_{l=1}^m a_{ul} \{ \delta_l + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + (\beta\gamma\delta)_{jkl} \} + \epsilon_{ijk_u}, \quad \text{where } \epsilon_{ijk_u} = \sum_l a_{ul} e_{ijkl}. \end{aligned}$$

We have $\text{Cov}[Z_{ijk_u}, Z_{ijk_{u'}}] = 0$ and $V[Z_{ijk_u}] = V[\sum a_{ul}y_{ijkl}] = \sigma^2(1 - \rho_1)$.

Under orthogonal transformation the sub-sub-plot observations transform to Z'_{ijk_u} and these are orthogonal. Using these observations, we have to estimate the parameters δ_l , $(\beta\gamma)_{jl}$, $(\gamma\delta)_{kl}$ and $(\beta\gamma\delta)_{jkl}$. Since these are the observations from the third steps of randomization, their analysis will generate an error sum of squares which is known as SS (Error-3). The variance of this error is $\sigma^2(1 - \rho_1)$.

Again, we have

$$M_{ijh} = \frac{1}{\sqrt{m}} \sum_{k=1}^q b_{hk}y_{ijk} = \frac{1}{\sqrt{m}} \sum_k b_{hk}(\gamma_k + (\beta\gamma)_{jk}) + \phi_{ijh}, \quad \text{where } \phi_{ijh} = \frac{1}{\sqrt{m}} \sum_{k=1}^q b_{hk}e_{ijk}.$$

$$V(\phi_{ijh}) = V(M_{ijh}) = \sigma^2[1 + (m-1)\rho_1 - m\rho_2].$$

This error variance arises due to the observations of sub-plot factor. It is the second kind of error variance. The parameters γ_k and $(\beta\gamma)_{jk}$ are to be estimated theoretically from M_{ijh} observations.

The observations W_{ij} are due to replications and whole-plot factor. These observations arise from the first step of randomization. Hence, the analysis of these observations will provide an estimate of first kind of error variance. The parameters μ , α_i and β_j are to be estimated using these observations.

We have observed three different errors arisen from orthogonal transformation. These are φ_{ij} , ϕ_{ijh} and ϵ_{ijk_u} . Let the reciprocal of the variances of these errors be

$$W_1 = \frac{1}{\sigma^2[1 + (m-1)\rho_1 + m(q-1)\rho_2]}$$

$$W_2 = \frac{1}{\sigma^2[1 + (m-1)\rho_1 + m\rho_2]} \quad \text{and} \quad W_3 = \frac{1}{\sigma^2(1 - \rho_1)}.$$

Now, the weighted error sum of squares using W_{ij} , M_{ijh} and Z_{ijk_u} is written as

$$\begin{aligned} \phi &= W_1 \sum_i \sum_j \{W_{ij} - \sqrt{qm}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)\}^2 \\ &\quad + W_2 \sum \sum \sum \sum \left\{ M_{ijh} - \frac{1}{\sqrt{m}} \sum_k b_{hk}(\hat{\gamma}_k + (\hat{\beta}\hat{\gamma})_{jk}) \right\}^2 \\ &\quad + W_3 \sum \sum \sum \sum \sum \left\{ Z_{ijk_u} - \sum_l a_{ul}(\hat{\delta}_l + (\hat{\beta}\hat{\delta})_{jl} + (\hat{\gamma}\hat{\delta})_{kl} + (\hat{\beta}\hat{\gamma}\hat{\delta})_{jkl}) \right\}^2. \end{aligned}$$

The estimates of the parameters are obtained solving the equations

$$\begin{aligned} \frac{\partial \phi}{\partial \hat{\alpha}_i} &= 0, \quad \frac{\partial \phi}{\partial \hat{\mu}} = 0, \quad \frac{\partial \phi}{\partial \hat{\beta}_j} = 0, \quad \frac{\partial \phi}{\partial \hat{\gamma}_k} = 0, \quad \frac{\partial \phi}{\partial \hat{\delta}_l} = 0, \quad \frac{\partial \phi}{\partial (\hat{\beta}\hat{\gamma})_{jk}} = 0, \\ \frac{\partial \phi}{\partial (\hat{\beta}\hat{\delta})_{jl}} &= 0, \quad \frac{\partial \phi}{\partial (\hat{\gamma}\hat{\delta})_{kl}} = 0 \quad \text{and} \quad \frac{\partial \phi}{\partial (\hat{\beta}\hat{\gamma}\hat{\delta})_{jkl}} = 0. \end{aligned}$$

The estimates are

$$\begin{aligned} \hat{\mu} &= \bar{y}_{\dots}, \hat{\alpha}_i = \bar{y}_{i\dots} - \bar{y}_{\dots}, \hat{\beta}_j = \bar{y}_{.j\dots} - \bar{y}_{\dots}, \hat{\gamma}_k = \bar{y}_{\dots k} - \bar{y}_{\dots} \\ \hat{\delta}_l &= \bar{y}_{\dots l} - \bar{y}_{\dots}, (\hat{\beta\gamma})_{jk} = \bar{y}_{.jk\dots} - \bar{y}_{.j\dots} - \bar{y}_{\dots k} + \bar{y}_{\dots} \\ (\hat{\beta\delta})_{jl} &= \bar{y}_{.jl\dots} - \bar{y}_{.j\dots} - \bar{y}_{\dots l} + \bar{y}_{\dots}, (\hat{\gamma\delta})_{kl} = \bar{y}_{\dots kl} - \bar{y}_{\dots k} - \bar{y}_{\dots l} + \bar{y}_{\dots} \\ (\hat{\beta\gamma\delta})_{jkl} &= \bar{y}_{.jkl\dots} - \bar{y}_{.jk\dots} - \bar{y}_{.j\dots l} - \bar{y}_{\dots kl} + \bar{y}_{.j\dots} + \bar{y}_{\dots k} + \bar{y}_{\dots l} - \bar{y}_{\dots} \end{aligned}$$

Since the estimates are free of reciprocals of error variances, the analysis, in practice, is done usually with the observations. The whole plot analysis is done with $y_{ij\dots}$ observations, the sub-plot analysis is done using $y_{ijk\dots}$ observations. The whole-plot error sum of squares is obtained during whole-plot analysis, sub-plot error sum of squares is obtained during sub-plot analysis. The sub-sub-plot error sum of squares is obtained during the analysis of original observations.

The different sum of squares are :

$$\begin{aligned} S_1 &= SS(\text{Replications}) = pqm \sum (\bar{y}_{i\dots} - \bar{y}_{\dots})^2 \\ S_2 &= SS(A) = qrm \sum (\bar{y}_{.j\dots} - \bar{y}_{\dots})^2 \\ S_3 &= SS(\text{Error-1}) = qm \sum \sum (\bar{y}_{ij\dots} - \bar{y}_{i\dots} - \bar{y}_{.j\dots} + \bar{y}_{\dots})^2 \\ S_4 &= SS(B) = prm \sum (\bar{y}_{\dots k} - \bar{y}_{\dots})^2 \\ S_5 &= SS(AB) = rm \sum (\bar{y}_{.jk\dots} - \bar{y}_{.j\dots} - \bar{y}_{\dots k} + \bar{y}_{\dots})^2 \\ S_6 &= SS(\text{Error-2}) = m \sum \sum (\bar{y}_{ijk\dots} - \bar{y}_{ij\dots} - \bar{y}_{.jk\dots} + \bar{y}_{.j\dots})^2 \\ S_7 &= SS(C) = pqr \sum (\bar{y}_{\dots l} - \bar{y}_{\dots})^2 \\ S_8 &= SS(AC) = qr \sum \sum (\bar{y}_{.jl\dots} - \bar{y}_{.j\dots} - \bar{y}_{\dots l} + \bar{y}_{\dots})^2 \\ S_9 &= SS(BC) = pr \sum \sum (\bar{y}_{\dots kl} - \bar{y}_{\dots k} - \bar{y}_{\dots l} + \bar{y}_{\dots})^2 \\ S_{10} &= SS(ABC) = r \sum \sum \sum (\bar{y}_{.jkl\dots} - \bar{y}_{.jk\dots} - \bar{y}_{.j\dots l} - \bar{y}_{\dots kl} + \bar{y}_{.j\dots} + \bar{y}_{\dots k} \\ &\quad + \bar{y}_{\dots l} - \bar{y}_{\dots})^2 \\ S_{11} &= SS(\text{Error-3}) = \sum \sum \sum (y_{ijkl} - \bar{y}_{.ijkl})^2. \end{aligned}$$

The main objectives of this analysis are to test the significance of the hypotheses :

- (i) $H_0 : \alpha_i = 0$ (ii) $H_0 : \beta_j = 0$ (iii) $H_0 : \gamma_k = 0$ (iv) $H_0 : (\beta\gamma)_{jk} = 0$
- (v) $H_0 : \delta_l = 0$ (vi) $H_0 : (\beta\delta)_{jl} = 0$ (vii) $H_0 : (\gamma\delta)_{kl} = 0$ and
- (viii) $H_0 : (\beta\gamma\delta)_{jkl} = 0$.

Here $\sigma_1^2 = \sigma^2[1 + (m-1)\rho_1 + m(q-1)\rho_2]$, $\sigma_2^2 = \sigma^2[1 + (m-1)\rho_1 - m\rho_2]$ and $\sigma_3^2 = \sigma^2(1 - \rho_1)$.

The F -statistic for h -th hypothesis is F_h ($h = 1, 2, \dots, 8$) as shown in ANOVA table. The conclusion regarding a particular hypothesis will be drawn as usual [$F_h \geq F_{0.05; f_1, f_2}$, H_0 is rejected; otherwise it is accepted. Here f_1 and f_2 are the numerator and denominator d.f., respectively of F_h].

One of the important aspect of the analysis of such experimental data is the comparison of two split-plot treatments in presence of a particular whole-plot treatment. The hypothesis for this is

$$H_0 : \gamma_k^j = \gamma_{k'}^j$$

against $H_A : \gamma_k^j \neq \gamma_{k'}^j; j = 1, 2, \dots, p; k \neq k' = 1, 2, \dots, q$.

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	$E(MS)$	F
Replications	$r - 1$	S_1	s_1	$\sigma_1^2 + \frac{pqm}{r-1} \sum \alpha_i^2$	$F_1 = \frac{s_1}{s_3}$
A	$p - 1$	S_2	s_2	$\sigma_1^2 + \frac{qrm}{p-1} \sum \beta_j^2$	$F_2 = \frac{s_2}{s_3}$
Error-1	$(p - 1)(r - 1)$	S_3	s_3	σ_1^2	—
B	$q - 1$	S_4	s_4	$\sigma_2^2 + \frac{prm}{q-1} \sum \gamma_k^2$	$F_3 = \frac{s_4}{s_6}$
AB	$(p - 1)(q - 1)$	S_5	s_5	$\sigma_2^2 + \frac{rm}{(p-1)(q-1)} \sum \sum (\beta\gamma)_{jk}^2$	$F_4 = \frac{s_5}{s_6}$
Error-2	$p(r - 1)(q - 1)$	S_6	s_6	σ_2^2	—
C	$(m - 1)$	S_7	s_7	$\sigma_3^2 + \frac{pqr}{(m-1)} \sum \delta_l^2$	$F_5 = \frac{s_7}{s_{11}}$
AC	$(p - 1)(m - 1)$	S_8	s_8	$\sigma_3^2 + \frac{qr}{(p-1)(m-1)} \sum \sum (\beta\delta)_{jl}^2$	$F_6 = \frac{s_8}{s_{11}}$
BC	$(q - 1)(m - 1)$	S_9	s_9	$\sigma_3^2 + \frac{pr}{(q-1)(m-1)} \sum \sum (\gamma\delta)_{kl}^2$	$F_7 = \frac{s_9}{s_{11}}$
ABC	$(p - 1)(q - 1)(m - 1)$	S_{10}	s_{10}	$\sigma_3^2 + \frac{r}{(p-1)(q-1)(m-1)} \sum \sum \sum (\beta\gamma\delta)_{jkl}^2$	$F_8 = \frac{s_{10}}{s_{11}}$
Error-3	$pq(r - 1)(m - 1)$	S_{11}	s_{11}	σ_3^2	—
Total	$pqrm - 1$				

Here γ_k^j = effect of k -th sub-plot treatment in j -th whole plot. The test statistic for this hypothesis is.

$$t = \frac{\bar{y}_{.jk} - \bar{y}_{.jk'}}{\sqrt{\frac{2s_6}{rm}}}$$

This t has $p(r - 1)(q - 1)$ d.f. This t test is used to compare a particular pair of means of the type $\bar{y}_{.jk}$ and $\bar{y}_{.jk'}$. All pairs of means can be compared by Duncan's multiple range test, where the test statistic is

$$D_h = d_{0.05, h, f} \sqrt{\frac{s_6}{rm}}, \quad h = 2, 3, \dots, q : f = p(r - 1)(q - 1).$$

However, if we need to compare two means of the types $\bar{y}_{.jk}$ and $\bar{y}_{.j'k'}$, then the test statistic is

$$T_1 = \frac{(\bar{y}_{.jk} - \bar{y}_{.j'k'})^2}{S_{01}^2 \left(\frac{2}{rm}\right)},$$

where $S_{01}^2 = \frac{m}{(r - 1)(p - 1)} \sum_j \sum_k (\bar{y}_{ijk} - \bar{y}_{.jk} - \bar{y}_{i.k} + \bar{y}_{..k})^2$.

The statistic T_1 is distributed as variance ratio distribution with 1 and $(r - 1)(p - 1)$ d.f. under $H_0 : \gamma_k^j = \gamma_{k'}^{j'} (j \neq j' = 1, 2, \dots, p; k \neq k' = 1, 2, \dots, q)$. The non-null distribution of T_1 is non-central F with non-centrality parameter

$$\lambda_1 = \frac{rm}{2\sigma^2[1 + (m - 1)\rho_1]} [\beta_j - \beta_{j'} + \gamma_k - \gamma_{k'} + (\beta\gamma)_{jk} - (\beta\gamma)_{j'k'}]^2.$$

The hypothesis to compare to split-split-plot treatment means in presence of a particular split-plot factor is

$$H_0 : \delta_l^k = \delta_{l'}^k$$

against $H_A : \delta_l^k \neq \delta_{l'}^k, l \neq l' = 1, 2, \dots, m; k = 1, 2, \dots, q.$

The test statistic for this hypothesis is

$$t = \frac{\bar{y}_{..kl} - \bar{y}_{..kl'}}{\sqrt{\frac{2s_{11}}{pr}}}$$

This t has $pq(m - 1)(r - 1)$ d.f. The Duncan's multiple range test statistic for the comparison of all pairs of above said means is

$$D_h = d_{\alpha,h,f} \sqrt{\frac{s_{11}}{pr}}, h = 2, 3, \dots, m; f = pq(m - 1)(r - 1).$$

However, if comparison of means $\bar{y}_{..kl}$ and $\bar{y}_{..k'l'}$ is needed, the test statistic is

$$T_2' = \frac{(\bar{y}_{..kl} - \bar{y}_{..k'l'})^2}{S_{02}^2 \left(\frac{2}{pr}\right)},$$

where $S_{02}^2 = \frac{1}{(r - 1)(q - 1)} \sum \sum (y_{ijkl} - \bar{y}_{.jkl} - \bar{y}_{ij.l} + \bar{y}_{.jkl})^2.$

The statistic T_2 is distributed as variance ratio with 1 and $(r - 1)(q - 1)$ d.f. under null hypothesis

$$H_0 : \delta_l^k = \delta_{l'}^k.$$

The non-null distribution of T_2 is non-central F with non-centrality parameter :

$$\lambda_2 = \frac{pr}{2(1 - \rho_2)\sigma^2} [\gamma_k - \gamma_{k'} + \delta_l - \delta_{l'} + (\gamma\delta)_{kl} - (\gamma\delta)_{k'l'}]^2.$$

The comparison of two split-split-plot treatment means at the same level of whole-plot treatment or at different levels of whole-plot treatment are done using the test statistics

$$T_3 = \frac{(\bar{y}_{.j.l} - \bar{y}_{.j.l'})^2}{S_{03}^2 \left(\frac{2}{qr}\right)} \text{ and } T_4 = \frac{(\bar{y}_{.j.l} - \bar{y}_{.j'.l'})^2}{S_{03}^2 \left(\frac{2}{qr}\right)}, \text{ respectively.}$$

The null distributions of both the statistics are variance ratio distribution with 1 and $(p - 1)(r - 1)$ d.f. Here

$$S_{03}^2 = \frac{q}{(p - 1)(r - 1)} \sum \sum (\bar{y}_{ij.l} - \bar{y}_{.j.l} - \bar{y}_{i..l} + \bar{y}_{..l})^2.$$

The non-null distributions of T_3 and T_4 are non-central F -distribution with non-centrality parameter :

$$\lambda_3 = \frac{qr}{2\sigma^2[1 + (q - 1)\rho_2]} [\beta_j - \beta_{j'} + (\beta\delta)_{jl} - (\beta\delta)_{j'l'}]^2$$

and $\lambda_4 = \frac{qr}{2\sigma^2[1 + (q - 1)\rho_2]} [\beta_j - \beta_{j'} + \delta_l - \delta_{l'} + (\beta\delta)_{jl} - (\beta\delta)_{j'l'}]^2.$

The multiple comparison of the means mentioned above can also be performed using S_{03}^2 , where the Duncan's multiple range test statistic is

$$D_h = d_{\alpha,h,f} \sqrt{\frac{S_{03}^2}{qr(p - 1)(r - 1)}}, h = 2, 3, \dots, m; f = (p - 1)(r - 1).$$

The analysis of data of split-split-plot experiment is also performed as usual in presence of one missing observation of l -th split-split-plot treatment corresponding to k -th split-plot treatment and j -th whole-plot treatment in i -th replication except that 1 is subtracted from total d.f. and hence, from error-3 d.f. The missing observation is estimated by minimising the error-3 sum of squares, where the estimate is

$$x = \frac{ry_{ijl} + my_{jkl} - y_{jk}}{(m-1)(r-1)}$$

Efficiency of split-split-plot design : The split-split-plot experiment may be considered as a $p \times q \times m$ asymmetrical factorial experiment, where number of treatments is pqm . The experiment with pqm treatments can also be performed through randomized block design. Allocating qm levels of two factors B and C in p levels of A , where p levels of A are randomly allocated to p whole-plots and qm levels are allocated to qm sub-plots of each whole-plot like split-plot experiment. Therefore, split-split-plot design can be compared with randomized block design and with split-plot design.

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$
Replications	$r - 1$	S_1	s_1
A	$p - 1$	S_2	s_2
Error-1	$(p - 1)(r - 1)$	S_3	s_3
B	$q - 1$	S_4	s_4
AB	$(p - 1)(q - 1)$	S_5	s_5
Error-2	$p(r - 1)(q - 1)$	S_6	s_6
C	$(m - 1)$	S_7	s_7
AC	$(p - 1)(m - 1)$	S_8	s_8
BC	$(q - 1)(m - 1)$	S_9	s_9
ABC	$(p - 1)(q - 1)(m - 1)$	S_{10}	s_{10}
Error-3	$pq(r - 1)(m - 1)$	S_{11}	s_{11}
Total	$pqr - 1$		

Let the effects of C , AC , BC and ABC be insignificant and these are on an average, equal to error-3 variance. The analysis of variance table transforms to the following type :

ANOVA Table

Sources of variation	d.f.	SS
Replications	$r - 1$	$(r - 1)s_1$
A	$p - 1$	$(p - 1)s_2$
Error-1	$(p - 1)(r - 1)$	$(p - 1)(r - 1)s_3$
B	$q - 1$	$(q - 1)s_4$
AB	$(p - 1)(q - 1)$	$(p - 1)(q - 1)s_5$
Error-2	$p(q - 1)(r - 1)$	$p(q - 1)(r - 1)s_6$
Error-3	$pqr(m - 1)$	$pqr(m - 1)s_{11}$
Total	$pqr - 1$	

The above analysis is similar to the analysis of observations in whole-plots and split-plots having r replications. Here total error sum of squares of split-plot analysis is $p(q - 1)(r - 1)s_6 +$

$pqr(m - 1)s_{11}$. This error sum of squares has $p(qrm - r - q + 1)$ d.f. Therefore, the efficiency of C, AC, BC and ABC of split-split-plot experiment compared to split-plot experiment is

$$\frac{(r - 1)(q - 1)s_6 + qr(m - 1)s_{11}}{(qrm - r - q + 1)s_{11}}$$

and the efficiency of B and AB is

$$\frac{(r - 1)(q - 1)s_6 + qr(m - 1)s_{11}}{(qrm - r - q + 1)s_6}$$

Again, let us consider that the effects of A, B and AB are insignificant. Assume that the effect of A , on an average, equals the error variance-1 and the effects of B and interaction AB , on an average, equal the error variance-2. The analysis of variance table then becomes

ANOVA Table

Sources of variation	d.f.	SS
Replications	$r - 1$	$(r - 1)s_1$
Error-1	$r(p - 1)$	$r(p - 1)s_3$
Error-2	$pr(q - 1)$	$pr(q - 1)s_6$
Error-3	$pqr(m - 1)$	$pqr(m - 1)s_{11}$
Total	$pqrm - 1$	

The total error sum of squares is $r(p - 1)s_3 + pr(q - 1)s_6 + pqr(m - 1)s_{11}$ and its d.f. is $r(pqm - 1)$. Therefore, compared to randomized block design the efficiency of split-split-plot design in estimating the effects and interactions C, AC, BC and AB is

$$\frac{(p - 1)s_3 + p(q - 1)s_6 + pq(m - 1)s_{11}}{(pqm - 1)s_{11}}$$

The efficiency of B and AB is

$$\frac{(p - 1)s_3 + p(q - 1)s_6 + pq(m - 1)s_{11}}{(pqm - 1)s_6}$$

Example 4.10 : An experiment is conducted to study the productivity of 3 different varieties of maize using 4 doses of nitrogen as urea. The seeds of maize varieties are sown in rows of 4 different types and these are sown at different distances in a row. The distance of plants in rows and within a row is considered as split-split-plot factors, nitrogen is a sub-plot factor and variety of maize is considered as whole-plot factor. The four doses of nitrogen are $N_1 = 150$ kg/ha; $N_2 = 200$ kg/ha, $N_3 = 250$ kg/ha and $N_4 = 300$ kg/ha. The distances of plants are :

$S_1 =$ row to row distance is 37.5 cm and plant to plant distance is 20 cm

$S_2 =$ row to row distance is 50 cm and plant to plant distance is 20 cm

$S_3 =$ row to row distance is 50 cm and plant to plant distance is 30 cm

$S_4 =$ row to row distance is 70 cm and plant to plant distance is 30 cm.

The productions of maize (kg/plot of 150 cm \times 120 cm) are recorded for analysis. The experiment is replicated 3 times.

- (i) Analyse the data and comment on the use of nitrogen fertilizer.
- (ii) Group the varieties of maize, if possible.

- (iii) Is there any difference in the use of N_4 in producing first two varieties of maize?
 (iv) Compare the levels of nitrogen in producing first variety of maize.
 (v) Compare the plantation plans for N_4 .
 (vi) Do you think that the plantation plans S_3 and S_4 are different for N_3 and N_4 ?
 (vii) Do you think that the first two varieties of maize are different in plantation plan S_4 ?
 (viii) Do you think that the first two varieties are giving similar results in plantation plans S_3 and S_4 ?
 (ix) Find the efficiency of split-split-plot design compared to split-plot design and randomized block design.

Production of Maize (y_{ijkl} , kg/plot)

Maize variety	Doses of nitrogen	Spaces in replications											
		Replications											
		1				2				3			
		S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
V_1	N_1	0.63	0.55	1.00	1.05	0.62	0.54	1.00	1.03	0.60	0.55	0.99	1.02
	N_2	0.68	0.65	1.00	1.06	0.67	0.66	1.02	1.05	0.67	0.66	1.00	1.06
	N_3	0.75	0.70	1.08	1.10	0.74	0.70	1.06	1.10	0.75	0.70	1.06	1.08
	N_4	0.80	0.67	1.02	1.08	0.80	0.68	1.00	1.10	0.81	0.68	1.02	1.08
V_2	N_1	0.66	0.58	1.00	1.04	0.72	0.60	1.00	1.02	0.70	0.59	1.00	1.01
	N_2	0.70	0.66	1.03	1.07	0.69	0.70	1.03	1.03	0.72	0.62	1.02	1.04
	N_3	0.76	0.72	1.10	1.05	0.77	0.71	1.06	1.06	0.77	0.71	1.03	1.00
	N_4	0.84	0.72	1.04	1.04	0.85	0.72	1.05	1.01	0.85	0.73	1.05	1.02
V_3	N_1	0.75	0.77	0.95	1.00	0.76	0.78	1.00	1.01	0.76	0.78	0.98	1.00
	N_2	0.76	0.77	1.00	1.00	0.77	0.76	1.02	1.02	0.77	0.79	1.02	1.01
	N_3	0.88	0.90	1.02	1.04	0.86	0.95	1.02	1.05	0.90	0.90	1.03	1.04
	N_4	0.86	0.92	1.08	1.06	0.85	0.92	1.06	1.06	0.92	0.90	1.08	1.06
Total	$y_{i...}$	42.59				42.68				42.53			

Solution : (i) Here $r = 3$, $p = 3$, $q = m = 4$.

The production of replications and varieties are ($y_{ij..}$) :

Varieties	Replications			Total y_{j00}
	1	2	3	Y_{0j00}
V_1	13.82	13.77	13.73	41.32
V_2	14.01	14.02	13.86	41.89
V_3	14.76	14.89	14.94	44.59
Total $y_{i...}$	42.59	42.68	42.53	127.80

The production of varieties and doses of nitrogen ($y_{.jk.}$) are :

Varieties	Doses of nitrogen			
	N_1	N_2	N_3	N_4
V_1	9.58	10.18	10.82	10.74
V_2	9.92	10.31	10.74	10.92
V_3	10.54	10.69	11.59	11.77
Total $y_{.k.}$	30.04	31.18	33.15	33.43

The productions of varieties, nitrogen and spaces are ($y_{.jkl}$) :

Doses of nitrogen	Varieties of maize											
	V_1				V_2				V_3			
	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
N_1	1.85	1.64	2.99	3.10	2.08	1.77	3.00	3.07	2.27	2.33	2.93	3.01
N_2	2.02	1.97	3.02	3.17	2.11	1.98	3.08	3.14	2.30	2.32	3.04	3.03
N_3	2.24	2.10	3.20	3.28	2.30	2.14	3.19	3.11	2.64	2.75	3.07	3.13
N_4	2.41	2.03	3.04	3.26	2.54	2.17	3.14	3.07	2.63	2.74	3.22	3.18
Total $y_{.j.l}$	8.52	7.74	12.25	12.81	9.03	8.06	12.41	12.39	9.84	10.14	12.26	12.35

Doses of nitrogen	Spaces			
	S_1	S_2	S_3	S_4
N_1	6.20	5.74	8.92	9.18
N_2	6.43	6.27	9.14	9.34
N_3	7.18	6.99	9.46	9.52
N_4	7.58	6.94	9.40	9.51
Total $y_{...l}$	27.39	25.94	36.92	37.55

$$G = 127.80$$

$$C.T. = \frac{G^2}{pqrm} = \frac{(127.80)^2}{3 \times 4 \times 3 \times 4} = 113.4225.$$

$$SS(\text{Total}) = \sum \sum \sum \sum y_{ijkl}^2 - C.T. = 117.2782 - 113.4225 = 3.8557.$$

$$SS(\text{Replications}) = \frac{1}{pqm} \sum y_{i...}^2 - C.T. = \frac{5444.2914}{3 \times 4 \times 4} - 113.4225 = 0.0002.$$

$$SS(\text{Varieties}) = \frac{1}{qrm} \sum y_{.j..}^2 - C.T. = \frac{5450.3826}{4 \times 3 \times 4} - 113.4225 = 0.1271.$$

$$SS(\text{Error-1}) = \frac{1}{qm} \sum \sum y_{ij..}^2 - C.T. - SS(\text{Replications}) - SS(\text{Varieties})$$

$$= \frac{1816.8316}{4 \times 4} - 113.425 - 0.0002 - 0.1271 = 0.0022.$$

$$SS(\text{Nitrogen}) = \frac{1}{prm} \sum y_{.k.}^2 - C.T. = \frac{4091.0814}{3 \times 3 \times 4} - 113.4225 = 0.2186$$

$$SS(\text{Varieties} \times \text{Nitrogen}) = \frac{1}{rm} \sum \sum y_{.jk.}^2 - C.T. - SS(\text{Varieties}) - SS(\text{Nitrogen})$$

$$= \frac{1365.354}{3 \times 5} - 113.4225 - 0.1271 - 0.2186 = 0.0113.$$

The productions of varieties, nitrogen in replications (y_{ijk} .)

Varieties	Replications											
	1				2				3			
	Doses of nitrogen											
	N_1	N_2	N_3	N_4	N_1	N_2	N_3	N_4	N_1	N_2	N_3	N_4
V_1	3.23	3.39	3.63	3.57	3.19	3.40	3.60	3.58	3.16	3.39	3.59	3.59
V_2	3.28	3.46	3.63	3.64	3.34	3.45	3.60	3.63	3.30	3.40	3.51	3.65
V_3	3.47	3.53	3.84	3.92	3.55	3.57	3.88	3.89	3.52	3.59	3.87	3.96

$$\begin{aligned}
 SS(\text{Error-2}) &= \frac{1}{m} \sum \sum \sum y_{ijk}^2 - \text{C.T.} - SS(\text{Replications}) - SS(\text{Varieties}) \\
 &\quad - SS(\text{Error-1}) - SS(\text{Nitrogen}) - SS(\text{Variety} \times \text{Nitrogen}) \\
 &= \frac{455.142}{4} - 113.4225 - 0.002 - 0.1271 - 0.0022 - 0.2186 - 0.0113 \\
 &= 0.0036.
 \end{aligned}$$

$$SS(\text{Spaces}) = \frac{1}{pqr} \sum \sum \sum y_{i..l}^2 - \text{C.T.} = \frac{4196.1846}{3 \times 4 \times 3} - 113.4225 = 3.1382.$$

$$\begin{aligned}
 SS(\text{Variety} \times \text{Spaces}) &= \frac{1}{cr} \sum \sum y_{j..l}^2 - \text{C.T.} - SS(\text{Variety}) - SS(\text{Spaces}) \\
 &= \frac{1403.1566}{4 \times 3} - 113.4225 - 0.1271 - 3.1382 = 0.2419.
 \end{aligned}$$

$$\begin{aligned}
 SS(\text{Nitrogen} \times \text{Spaces}) &= \frac{1}{pr} \sum \sum y_{..kl}^2 - \text{C.T.} - SS(\text{Nitrogen}) - SS(\text{Spaces}) \\
 &= \frac{1051.614}{3 \times 3} - 113.4225 - 0.2186 - 3.1382 = 0.0667.
 \end{aligned}$$

$$\begin{aligned}
 SS(\text{Variety} \times \text{Nitrogen} \times \text{Spaces}) &= \frac{1}{r} \sum \sum \sum y_{jkl}^2 - \text{C.T.} - SS(\text{Variety}) \\
 &\quad - SS(\text{Nitrogen}) - SS(\text{Spaces}) - SS(\text{Variety} \times \text{Nitrogen}) - SS(\text{Variety} \times \text{Spaces}) \\
 &\quad - SS(\text{Nitrogen} \times \text{Spaces}) \\
 &= \frac{351.7634}{3} - 113.4225 - 0.1271 - 0.2186 \\
 &\quad - 3.1382 - 0.0113 - 0.2419 - 0.0667 \\
 &= 0.0282.
 \end{aligned}$$

$$\begin{aligned}
 SS(\text{Error-3}) &= SS(\text{Total}) - SS(\text{Replications}) - SS(\text{Varieties}) - SS(\text{Error-1}) \\
 &\quad - SS(\text{Nitrogen}) - SS(\text{Variety} \times \text{Nitrogen}) - SS(\text{Error-2}) \\
 &\quad - SS(\text{Spaces}) - SS(\text{Variety} \times \text{Spaces}) - SS(\text{Nitrogen} \times \text{Spaces}) \\
 &\quad - SS(\text{Variety} \times \text{Nitrogen} \times \text{Spaces}) \\
 &= 3.8557 - 0.0002 - 0.1271 - 0.0022 - 0.2186 - 0.0113 \\
 &\quad - 0.0036 - 3.1382 - 0.2419 - 0.0667 - 0.0282 \\
 &= 0.0177.
 \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F	F ₀₅
Replications	2	0.0002	0.0001	0.18	6.04
Variety (V)	2	0.1271	0.06355	115.54	"
Error-1	4	0.0022	0.00055	—	—
Nitrogen (N)	3	0.2186	0.07287	364.35	3.16
VN	6	0.0113	0.00188	9.40	2.66
Error-2	18	0.0036	0.0002	—	—
Space (S)	3	3.1382	1.04607	4558.62	2.74
VS	6	0.2419	0.04032	168.00	2.23
NS	9	0.0667	0.00741	30.87	2.02
VNS	18	0.0282	0.00157	6.54	1.75
Error-3	72	0.0177	0.00024	—	—
Total	143				

It is observed that the varieties, the doses of nitrogen and the different spaces are highly significantly different. The mean productions due to the use of different doses of nitrogen are :

$$\bar{N}_1 = 0.83, \bar{N}_2 = 0.87, \bar{N}_3 = 0.92, \bar{N}_4 = 0.93.$$

With the increase in the doses of nitrogen the production of maize increases and this is true for every dose. All doses are significantly different. This is observed by Duncan's multiple range test, where the test statistic is

$$D_h = d_{0.05,h,f} \sqrt{\frac{s_6}{prm}}, \quad h = 2, 3, 4; \quad f = 18.$$

Here $D_2 = 0.0070$, $D_3 = 0.0073$, and $D_4 = 0.0076$.

(ii) The variety means are, in ascending order, as follows :

$$\bar{V}_1 = 0.86, \bar{V}_2 = 0.87, \bar{V}_3 = 0.93.$$

The Duncan's multiple range test statistic for grouping variety means is

$$D_h = d_{0.05,h,f} \sqrt{\frac{s_3}{qrm}}, \quad h = 2, 3; \quad f = 4.$$

Here $D_2 = 0.0133$, $D_3 = 0.0136$.

It is observed that V_1 and V_2 are in the same group, since $\bar{V}_2 - \bar{V}_1 < D_2$.

(iii) We need to test significance of the hypothesis $H_0 : \beta_1^4 = \beta_2^4$ against $H_A : \beta_1^4 \neq \beta_2^4$, where $\beta_1^4 =$ effect of variety V_1 in presence of N_4 and $\beta_2^4 =$ effect of variety V_2 in presence of N_4 . The test statistic for this hypothesis is

$$T_1 = \frac{(\bar{y}_{.14.} - \bar{y}_{.24.})^2}{S_{01}^2 \left(\frac{2}{rm}\right)}, \quad \text{where } r = 3, m = 4.$$

$$S_{01}^2 = \frac{m}{(r-1)(p-1)} \sum \sum (\bar{y}_{ij4.} - \bar{y}_{.j4.} - \bar{y}_{i.1.} + \bar{y}_{..4.})^2.$$

$$\begin{aligned} \text{Here } m \sum \sum (\bar{y}_{ij4.} - \bar{y}_{.j4.} - \bar{y}_{i.4.} + \bar{y}_{..4.})^2 \\ = \frac{1}{m} \sum \sum y_{ij4.}^2 - C.T_1 - SS(\text{Replication})_1 - SS(\text{Varieties})_1. \end{aligned}$$

This sum of squares are calculated using the data of only N_4 in different replications, whole-plots and in different spaces.

The productions of replications and varieties ($y_{ij4.}$) are given below :

Replications	Varieties			Total, R_i
	V_1	V_2	V_3	
1	3.57	3.64	3.92	11.13
2	3.58	3.63	3.89	11.10
3	3.59	3.65	3.96	11.20
Total V_j	10.74	10.92	11.77	33.43 = G_1

$$C.T_1 = \frac{G_1^2}{prm} = \frac{(33.43)^2}{3 \times 3 \times 4} = 31.0435.$$

$$SS(\text{Replications})_1 = \frac{1}{pm} \sum R_i^2 - C.T_1 = \frac{372.5269}{3 \times 4} - 31.0435 = 0.0004.$$

$$SS(\text{Varieties})_1 = \frac{1}{rm} \sum V_j^2 - C.T_1 = \frac{373.1269}{3 \times 4} - 31.0435 = 0.0504.$$

$$\begin{aligned} S_{01}^2 &= \frac{1}{(r-1)(p-1)} \left[\frac{1}{m} \sum \sum y_{ij4.}^2 - C.T_1 - SS(\text{Replications})_1 - SS(\text{Varieties})_1 \right] \\ &= \frac{1}{2 \times 2} \left[\frac{124.3785}{4} - 31.0435 - 0.0004 - 0.0504 \right] = 0.00008. \end{aligned}$$

$$T_1 = \frac{(0.895 - 0.91)^2}{0.00008 \times \frac{2}{3 \times 4}} = 16.87.$$

Since $T_1 > F_{0.05; 1,4} = 7.41$, H_0 is rejected. The averages of production of variety-1 and variety-2 in presence of N_4 are significantly different.

(iv) We need to compare the means $\bar{y}_{.1k.}$ ($k = 1, 2, 3, 4$). The Duncan's multiple range test statistic for the comparison is :

$$D_h = d_{05, h, f} \sqrt{\frac{s_6}{rm}}, \quad h = 2, 3, 4; \quad f = 18 \text{ (d.f. of } s_6\text{)}.$$

$$D_2 = 2.97 \sqrt{\frac{0.0002}{3 \times 4}} = 0.0121, \quad D_3 = 3.12 \sqrt{\frac{0.0002}{3 \times 4}} = 0.0127,$$

$$D_4 = 3.21 \sqrt{\frac{0.0002}{3 \times 4}} = 0.0131.$$

The means to be compared are :

$$\bar{N}_{11} = \frac{y_{.11.}}{rm} = \frac{9.58}{12} = 0.798, \quad \bar{N}_{12} = \frac{y_{.12.}}{rm} = \frac{10.18}{12} = 0.848,$$

$$\bar{N}_{13} = \frac{y_{.13.}}{rm} = \frac{10.82}{12} = 0.902, \quad \bar{N}_{14} = \frac{y_{.14.}}{rm} = \frac{10.74}{12} = 0.895.$$

$\bar{N}_{13} - \bar{N}_{11} = 0.104 > D_4$, all means are significantly different.

$\bar{N}_{13} - \bar{N}_{12} = 0.054 > D_3$, N_{13} and N_{12} are different.

$N_{14} - \bar{N}_{11} = 0.097 > D_3$, N_{11} and N_{14} are different.

$\bar{N}_{12} - \bar{N}_{11} = 0.05 > D_2$, N_{11} and N_{12} are different.

$\bar{N}_{14} - \bar{N}_{12} = 0.047 > D_2$, N_{12} and N_{14} are different.

$\bar{N}_{13} - \bar{N}_{14} = 0.007 < D_2$, N_{13} and N_{14} are similar.

The first variety of maize is produced in similar amount using N_3 and N_4 .

(v) We need to compare the means of S_l ($l = 1, 2, 3, 4$) in presence of N_4 . The means are $\bar{y}_{..4l}$, where

$$\bar{y}_{..41} = \frac{y_{..41}}{pr} = \frac{7.58}{9} = 0.842, \bar{y}_{..42} = \frac{y_{..42}}{pr} = \frac{6.94}{9} = 0.771.$$

$$\bar{y}_{..43} = \frac{y_{..43}}{pr} = \frac{9.40}{9} = 1.044, \bar{y}_{..44} = \frac{y_{..44}}{pr} = \frac{9.51}{9} = 1.057.$$

The Duncan's multiple range test statistic for the comparison of the above means is

$$D_h = d_{0.05,h,f} \sqrt{\frac{s_{11}}{pr}}; h = 2, 3, 4; f = 72 \text{ (d.f. of } s_{11}\text{)}.$$

$$D_2 = 2.821 \sqrt{\frac{0.00024}{9}} = 0.0145, D_3 = 2.971 \sqrt{\frac{0.00024}{9}} = 0.0153,$$

$$D_4 = 3.071 \sqrt{\frac{0.00024}{9}} = 0.0158.$$

$\bar{y}_{..44} - \bar{y}_{..42} = 0.0286 > D_4$, all means are significantly different.

$\bar{y}_{..44} - \bar{y}_{..41} = 0.215 > D_3$, S_1 and S_4 in presence of N_4 are different.

$\bar{y}_{..43} - \bar{y}_{..42} = 0.273 > D_3$, S_2 and S_3 in presence of N_4 are different.

$\bar{y}_{..41} - \bar{y}_{..42} = 0.071 > D_2$, S_1 and S_2 in presence of N_4 are different.

$\bar{y}_{..43} - \bar{y}_{..41} = 0.202 > D_2$, S_1 and S_3 in presence of N_4 are different.

$\bar{y}_{..44} - \bar{y}_{..43} = 0.013 < D_2$, S_3 and S_4 in presence of N_4 are similar.

(vi) We need to compare the means $\bar{y}_{..34}$ and $\bar{y}_{..43}$. The test statistic to compare such means ($\bar{y}_{..kl}$ and $\bar{y}_{..k'l'}$) is

$$T_2 = \frac{(\bar{y}_{..34} - \bar{y}_{..43})^2}{S_{02}^2 \left(\frac{2}{pr}\right)},$$

where
$$S_{02}^2 = \frac{1}{(r-1)(q-1)} \sum_i \sum_k (y_{ijkl} - \bar{y}_{.jkl} - \bar{y}_{ij.l} + \bar{y}_{.j.l})^2.$$

This means that sum of squares is to be calculated using the observations of any particular variety and the observations of any particular space. Since we are comparing the means of N_3 related to S_4 , we can use the observations related to factor level S_4 in presence of V_1 .

The productions in replications corresponding to V_1 and S_4 (y_{i1k4}) are shown below :

Replications	Doses of nitrogen				Total R_k
	N_1	N_2	N_3	N_4	
1	1.05	1.06	1.10	1.08	4.29
2	1.03	1.05	1.10	1.10	4.28
3	1.02	1.06	1.08	1.08	4.24
Total N_k	3.10	3.17	3.28	3.26	12.81 = G_2

$$C.T_2 = \frac{G_2^2}{rq} = \frac{(12.81)^2}{3 \times 4} = 13.6747.$$

$$SS(\text{Total})_2 = \sum \sum y_{i1k4}^2 - C.T_2 = 13.6827 - 13.6747 = 0.008.$$

$$SS(\text{Replications})_2 = \frac{1}{q} \sum R_k^2 - C.T_2 = \frac{54.7001}{4} - 13.6747 = 0.0003.$$

$$SS(\text{Nitrogen})_2 = \frac{1}{r} \sum N_k^2 - C.T_2 = \frac{41.0449}{3} - 13.6747 = 0.0069.$$

$$S_{02}^2 = \frac{1}{(r-1)(q-1)} [SS(\text{Total})_2 - SS(\text{Replications})_2 - SS(\text{Nitrogen})_2]$$

$$= \frac{1}{2 \times 3} [0.008 - 0.0003 - 0.0069] = 0.00013.$$

$$T_2 = \frac{(1.058 - 1.044)^2}{\frac{2}{3 \times 3} \times 0.00013} = 6.78.$$

This $T_2 > F_{0.05;1,6} = 5.99$. The two means are significantly different.

(vii) We need to compare the means $\bar{y}_{.1.4}$ and $\bar{y}_{.2.4}$. The test statistic to compare these means is :

$$T_3 = \frac{(\bar{y}_{.1.4} - \bar{y}_{.2.4})^2}{S_{03}^2 \left(\frac{2}{qr} \right)},$$

$$\text{where } S_{03}^2 = \frac{q}{(p-1)(r-1)} \sum_i \sum_j (\bar{y}_{ij.4} - \bar{y}_{.j.4} - \bar{y}_{i..4} + \bar{y}_{...4})^2.$$

This S_{03}^2 is to be calculated using the observations of S_4 only for different replications, whole-plots and sub-plots.

The productions of S_4 in replications ($y_{ij.4}$) are tabulated below :

Replications	Varieties			Total, R_l
	V_1	V_2	V_3	
1	4.29	4.20	4.10	12.59
2	4.28	4.12	4.14	12.54
3	4.24	4.07	4.11	12.42
Total V_l	12.81	12.39	12.35	37.55 = G_3

$$C.T_3 = \frac{G_3^2}{pqr} = \frac{(37.55)^2}{3 \times 4 \times 3} = 39.1667.$$

$$SS(\text{Total})_3 = \frac{1}{q} \sum \sum y_{ij.4}^2 - C.T_3 = \frac{156.7211}{4} - 39.1667 = 0.013575.$$

$$SS(\text{Replications})_3 = \frac{1}{pq} \sum R_i^2 - C.T_3 = \frac{470.0161}{3 \times 4} - 39.1667 = 0.001308.$$

$$SS(\text{Varieties})_3 = \frac{1}{qr} \sum V_i^2 - C.T_3 = \frac{470.1307}{4 \times 3} - 39.1667 = 0.010858.$$

$$S_{03}^2 = \frac{1}{(p-1)(r-1)} [SS(\text{Total})_3 - SS(\text{Replications})_3 - SS(\text{Varieties})_3]$$

$$= \frac{1}{2 \times 2} [0.013575 - 0.001308 - 0.010858] = 0.00035.$$

$$T_3 = \frac{(1.0675 - 1.0325)^2}{\frac{2}{4 \times 3}(0.00035)} = 21.00.$$

Since $T_3 > F_{0.05; 1,4} = 7.71$, the means are significantly different.

(viii) We need to compare the means $\bar{y}_{.1.3}$ and $\bar{y}_{.2.4}$ and the means $\bar{y}_{.1.4}$ and $\bar{y}_{.2.3}$. The test statistics for the comparison of two groups of means are :

$$T_4 = \frac{(\bar{y}_{.1.3} - \bar{y}_{.2.4})^2}{S_{03}^2 \left(\frac{2}{qr}\right)} = \frac{(1.0208 - 1.0325)^2}{\frac{2}{4 \times 3}(0.00035)} = 2.35$$

and $T_5 = \frac{(\bar{y}_{.1.4} - \bar{y}_{.2.3})^2}{S_{03}^2 \left(\frac{2}{qr}\right)} = \frac{(1.0675 - 1.0342)^2}{\frac{2}{4 \times 3}(0.00035)} = 19.01.$

Since $T_4 < F_{0.05; 1,4} = 7.71$, the means $\bar{y}_{.1.3}$ and $\bar{y}_{.2.4}$ are similar. But the means $\bar{y}_{.1.4}$ and $\bar{y}_{.2.3}$ are significantly different since $T_5 > F_{0.05; 1,4} = 7.71$.

(ix) The efficiency of split-split-plot design compared to split-plot design in estimating the effects and interactions S, VS, NS and VNS is

$$\frac{(r-1)(q-1)s_6 + qr(m-1)s_{11}}{(qrm - r - q + 1)s_{11}} = \frac{(3-1)(4-1)0.0002 + 4 \times 3(4-1)0.00024}{(4 \times 3 \times 4 - 3 - 4 + 1)0.00024} = 97.62\%.$$

This efficiency in estimating the effects and interactions of N and VN is

$$\frac{(r-1)(q-1)s_6 + qr(m-1)s_{11}}{(qrm - r - q + 1)s_6} = \frac{(3-1)(4-1)0.0002 + 4 \times 3(4-1)0.00024}{(4 \times 3 \times 4 - 3 - 4 + 1)0.0002} = 117.14\%.$$

The same two efficiencies compared to randomized block design are respectively :

$$\frac{(p-1)s_3 + p(q-1)s_6 + pq(m-1)s_{11}}{(pqm - 1)s_{11}}$$

$$= \frac{(3-1)0.00055 + 3(4-1)0.0002 + 3 \times 4(4-1)0.00024}{(3 \times 4 \times 4 - 1)0.00024} = 102.3\%$$

and $\frac{(p-1)s_3 + p(q-1)s_6 + pq(m-1)s_{11}}{(pqm - 1)s_6}$

$$= \frac{(3-1)0.00055 + 3(4-1)0.0002 + 3 \times 4(4-1)0.00024}{(3 \times 4 \times 4 - 1)0.0002} = 122.8\%.$$

4.14 Split-Block Design (Split-Plot Design with Sub-units in Strips)

Let there be two factors A and B . The factor A has p levels and the factor B has q levels. The p levels of A are randomly allocated to p whole-plots of a block and each level of B is randomly allocated to all whole-plots of a block. The arrangement of p levels of A and q levels of B in a block (replications) is as follows :

A	A_1	A_2	...	A_p
B				
B_1				
B_2				
\vdots				
B_q				

Here the levels of B are perpendicularly allocated over the levels of A so that each level of B is allocated to a block of p plots (p levels of A). If this arrangement of p levels of A and q levels of B are repeated in r blocks, the resultant design is called split-block design or strip-plot design where sub-units are allocated in a strip of p plots.

The design is used in agricultural experiment as well as in laboratory experiment. For example, let us consider that in a zoological experiment the death rate of earth-worm kept in different pesticides is under investigation. Let there be four types of earth-worms. These are kept in three different pesticides in such a way that groups of earth-worm of each type are kept in each pesticide separately. If B_1, B_2 and B_3 are the types of pesticides and A_1, A_2, A_3 and A_4 are the types of earth-worms; then the earth-worms can be kept according to a plan as shown above. The advantage of such experiment is that it helps us to study the joint impacts of the type of earth-worm and pesticide, where pesticide can be allocated to experimental units of smaller size. The interaction AB is estimated more efficiently.

Let y_{ijl} be the result of l -th level of B in j -th level of A and in i -th replication ($i = 1, 2, \dots, r; j = 1, 2, \dots, p; l = 1, 2, \dots, q$). The linear model for y_{ijl} observation is

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l + (\beta\gamma)_{jl} + e_{ijl},$$

where μ = general mean, α_i = effect of i -th replication, β_j = effect of j -th level of A , γ_l = effect of l -th level of B , $(\beta\gamma)_{jl}$ = interaction of j -th level of A with l -th level of B , e_{ijl} = random component.

Assume that the model is a fixed effect model with restrictions

$$\sum \alpha_i = \sum \beta_j = \sum_j (\beta\gamma)_{jl} = \sum_l (\beta\gamma)_{jl} = 0.$$

Further assumptions are :

$$\begin{aligned} E(e_{ijl}) &= 0, \quad E[e_{ijl}, e_{i'j'l'}] = \sigma^2, \quad \text{if } i = i', j = j', l = l' \\ &= \rho_1 \sigma^2, \quad \text{if } i = i', j = j', l \neq l' \\ &= \rho_2 \sigma^2, \quad \text{if } i = i', j \neq j', l = l' \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

The assumption on errors indicates that the data are non-orthogonal. Therefore, orthogonal transformation on data vector are needed at least in the theoretical aspects of the analysis. Let the first transformation be

$$\begin{bmatrix} U_{ij} \\ Z_{ij1} \\ Z_{ij2} \\ \vdots \\ Z_{ijq-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{q}} & \frac{1}{\sqrt{q}} & \cdots & \frac{1}{\sqrt{q}} \\ a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \dots & \dots & \dots & \dots \\ a_{q-1,1} & a_{q-1,2} & \cdots & a_{q-1,q} \end{bmatrix} \begin{bmatrix} y_{ij1} \\ y_{ij2} \\ \vdots \\ y_{ijq} \end{bmatrix}$$

Thus, $U_{ij} = \frac{1}{\sqrt{q}} \sum_{l=1}^q y_{ijl} = \sqrt{q}(\mu + \alpha_i + \beta_j) + \frac{1}{\sqrt{q}}e_{ij}$.

and $Z_{iju} = \sum_{l=1}^q a_{ul}[\gamma_l + (\beta\gamma)_{jl}] + \sum_l a_{ul}e_{ijl}$, $u \neq l = 1, 2, \dots, (q - 1)$.

But U_{ij} or Z_{iju} are not orthogonal. Further, orthogonal transformations are :

$$\begin{bmatrix} V_i \\ W_{i1u} \\ W_{i2u} \\ \vdots \\ W_{i(p-1)u} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{pq}} & \frac{1}{\sqrt{pq}} & \cdots & \frac{1}{\sqrt{pq}} \\ b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \dots & \dots & \dots & \dots \\ b_{p-1,1} & b_{p-1,2} & \cdots & b_{p-1,p} \end{bmatrix} \begin{bmatrix} U_{i1} \\ U_{i2} \\ \vdots \\ U_{ip} \end{bmatrix} \quad \text{and}$$

$$\begin{bmatrix} M_{i \cdot u} \\ M_{i1u} \\ M_{i2u} \\ \vdots \\ M_{i(p-1)u} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{pq}} & \frac{1}{\sqrt{pq}} & \cdots & \frac{1}{\sqrt{pq}} \\ b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \dots & \dots & \dots & \dots \\ b_{p-1,1} & b_{p-1,2} & \cdots & b_{p-1,p} \end{bmatrix} \begin{bmatrix} Z_{i1u} \\ Z_{i2u} \\ Z_{i3u} \\ \vdots \\ Z_{ipu} \end{bmatrix}$$

Thus, we have $V_i = \sqrt{pq}(\mu + \alpha_i) + \frac{1}{\sqrt{pq}}e_{i \cdot \cdot}$.

$$\begin{aligned} W_{iku} &= \sum_j b_{kj}U_{ij}, \quad k \neq j = 1, 2, \dots, (p - 1) \\ &= \sum b_{kj}\beta_j + \frac{1}{\sqrt{q}} \sum_j b_{kj}e_{ij} \cdot \end{aligned}$$

$$M_{i \cdot u} = \frac{1}{\sqrt{p}}Z_{i \cdot u} = \sqrt{p} \sum_l a_{ul}\gamma_l + \frac{1}{\sqrt{p}} \sum_l a_{ul}e_{i \cdot l}$$

$$M_{iku} = \sum_j b_{kj}Z_{iju} = \sum_j \sum_l a_{ul}b_{kj}(\beta\gamma)_{jl} + \sum_j \sum_l a_{ul}b_{kj}e_{ijl}$$

The random component e_{ijl} is divided into several components, where

$$e_{ijl} = \frac{1}{\sqrt{pq}}e_{i \cdot \cdot} + \frac{1}{\sqrt{q}} \sum_j b_{kj}e_{ij} \cdot + \frac{1}{\sqrt{p}} \sum_l a_{ul}e_{i \cdot l} + \sum_j \sum_l a_{ul}b_{kj}e_{ijl}$$

We have

$$V(V_i) = V\left(\frac{1}{pq}e_{i \cdot \cdot}\right) = \sigma^2[1 + (q - 1)\rho_1 + (p - 1)\rho_2].$$

But the sum of squares of the random error component ($e_{i..}/\sqrt{pq}$) is not available. Here $e_{i..}$ is related to the observations of i -th replication. This is available by adding all observations in i -th replication ($i = 1, 2, \dots, r$). But these added observations give sum of squares of replications. Therefore, in analysing data of split-block design, we have three different errors. These errors and their variances are shown below :

$$V \left[\frac{1}{\sqrt{q}} \sum b_{kj} e_{ij} \right] = \sigma^2 [1 + (q-1)\rho_1 - \rho_2]$$

$$V \left[\frac{1}{\sqrt{p}} \sum a_{ul} e_{il} \right] = \sigma^2 [1 - \rho_1 + (p-1)\rho_2]$$

$$V \left[\sum_j \sum_l a_{ul} b_{kj} e_{ijl} \right] = \sigma^2 [1 - \rho_1 - \rho_2].$$

$$\text{Let } W_2 = \frac{1}{\sigma^2 [1 + (q-1)\rho_1 - \rho_2]}, \quad W_3 = \frac{1}{\sigma^2 [1 - \rho_1 + (p-1)\rho_2]}$$

$$\text{and } W_4 = \frac{1}{\sigma^2 [1 - \rho_1 - \rho_2]}, \quad W_1 = \frac{1}{\sigma^2 [1 + (q-1)\rho_1 + (p-1)\rho_2]}.$$

Therefore, the estimates of parameters are to be found out by minimizing weighted error sum of squares ϕ , where

$$\begin{aligned} \phi = & W_1 \sum_i [V_i - \sqrt{pq}(\hat{\mu} + \hat{\alpha}_i)]^2 + W_2 \sum_i \sum_k \sum_u [W_{iku} - \sum b_{kj} \hat{\beta}_j]^2 \\ & + W_3 \sum_i \sum_u [M_{i..u} - \sqrt{p} \sum_l a_{ul} \hat{\gamma}_l]^2 + W_4 \sum_i \sum_k \sum_u \left[M_{iku} = \sum_j \sum_l a_{ul} b_{kj} (\hat{\beta}\hat{\gamma})_{jl} \right]^2 \end{aligned}$$

The estimates are

$$\begin{aligned} \hat{\mu} &= \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad \hat{\gamma}_l = \bar{y}_{..l} - \bar{y}_{...}, \\ (\hat{\beta}\hat{\gamma})_{jl} &= \bar{y}_{.jl} - \bar{y}_{.j.}\bar{y}_{..l} - \bar{y}_{..l} + \bar{y}_{...} \end{aligned}$$

It is observed that the estimates are free of weights. Hence, the sum of squares due to estimates are also of free of weights. The sum of squares are :

$$SS(\text{Replications}) = S_1 = pq \sum (\bar{y}_{i..} - \bar{y}_{...})^2,$$

$$SS(A) = S_2 = qr \sum (\bar{y}_{.j.} - \bar{y}_{...})^2.$$

$$SS(\text{Error-1}) = S_3 = q \sum \sum (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2.$$

$$SS(B) = S_4 = pr \sum (\bar{y}_{..l} - \bar{y}_{...})^2,$$

$$SS(\text{Error-2}) = p \sum \sum (y_{i.l} - \bar{y}_{i..} - \bar{y}_{..l} + \bar{y}_{...})^2 = S_5$$

$$SS(AB) = S_6 = r \sum \sum (\bar{y}_{.jl} - \bar{y}_{.j.} - \bar{y}_{..l} + \bar{y}_{...})^2$$

$$SS(\text{Error-3}) = \sum \sum \sum (y_{ijl} - \bar{y}_{ij.} - \bar{y}_{i.l} - \bar{y}_{.jl} + \bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{..l} - \bar{y}_{...})^2 = S_7.$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F
Replications	$r - 1$	S_1	s_1	
A	$p - 1$	S_2	s_2	$F_1 = s_2/s_3$
Error-1	$(p - 1)(r - 1)$	S_3	s_3	—
B	$q - 1$	S_4	s_4	$F_2 = s_4/s_5$
Error-2	$(q - 1)(r - 1)$	S_5	s_5	—
AB	$(p - 1)(q - 1)$	S_6	s_6	$F_3 = s_6/s_7$
Error-3	$(p - 1)(q - 1)(r - 1)$	S_7	s_7	—
Total	$pqr - 1$			

Here the test statistics F_1, F_2 and F_3 are calculated to test the significance of the hypotheses

(i) $H_0 : \beta_j = 0$ against $H_A : \beta_j \neq 0$

(ii) $H_0 : \gamma_l = 0$ against $H_A : \gamma_l \neq 0$

and (iii) $H_0 : (\beta\gamma)_{jl} = 0$ against $H_A : (\beta\gamma)_{jl} \neq 0$

respectively. The multiple comparison to compare the means of different levels of A and B are usually found out. However, the comparison of two means of B in presence of a particular level of A ($\bar{y}_{.jl}$ and $\bar{y}_{.j'l}$) is done using the test statistic :

$$T_1 = \frac{r(\bar{y}_{.jl} - \bar{y}_{.j'l})^2}{2S_{01}}, \text{ where } S_{01} = \frac{1}{(q - 1)(r - 1)} \sum \sum (y_{ijl} - \bar{y}_{.jl} - \bar{y}_{ij.} + \bar{y}_{.j.})^2.$$

The null distribution of T_1 is central F with 1 and $(q - 1)(r - 1)$ d.f. The non-null distribution of T_1 is non-central F with non-centrality parameter

$$\lambda_1 = \frac{r}{2\sigma^2(1 - \rho_1)} [\gamma_l - \gamma_{l'} + (\beta\gamma)_{jl} - (\beta\gamma)_{j'l}]^2.$$

Again, the comparison of two means of the types $\bar{y}_{.jl}$ and $\bar{y}_{.j'l}$ can be made using the test statistic

$$T_2 = \frac{r(\bar{y}_{.jl} - \bar{y}_{.j'l})^2}{2S_{02}} \text{ where } S_{02} = \frac{1}{(p - 1)(r - 1)} \sum \sum (y_{ijl} - \bar{y}_{.jl} - \bar{y}_{i.l} + \bar{y}_{..l})^2$$

The null distribution of T_2 is central F with 1 and $(p - 1)(r - 1)$ d.f. The non-null distribution of T_2 is non-central F with non-centrality parameter :

$$\lambda_2 = \frac{r}{2\sigma^2(1 - \rho_2)} [\beta_j - \beta_{j'} + (\beta\gamma)_{jl} - (\beta\gamma)_{j'l}]^2.$$

Since $EMS(\text{replications}) = \sigma^2[1 + (q - 1)\rho_1 + (p - 1)\rho_2] + \frac{pq}{r-1} \sum \alpha_i^2$, the exact F statistic is not available to test the significance of $H_0 : \alpha_i = 0$. It is seen that

$$E(s_1 + s_7) = E(s_3 + s_5)$$

under H_0 . Therefore, approximate F-statistic is

$$F = \frac{s_1 + s_7}{s_3 + s_5}$$

According to Satterthwaite (1946), the above F has

$$\frac{(s_1 + s_7)^2}{\frac{s_1^2}{r-1} + \frac{s_7^2}{(p-1)(r-1)(q-1)}}, \text{ and } \frac{(s_3 + s_5)^2}{\frac{s_3^2}{(p-1)(r-1)} + \frac{s_5^2}{(q-1)(r-1)}} \text{ d.f.}$$

Example 4.11 : In an agricultural research station an experiment is conducted to study the germination rate of a variety of maize using three doses of phosphorus as P_2O_5 . The doses of phosphorus are $P_1 = 90$ kg/ha, $P_2 = 120$ kg/ha and $P_3 = 150$ kg/ha. The doses of phosphorus are used randomly in 3 groups of plots. During land preparation the doses of phosphorus are applied. In the plots, where a dose of phosphorus is applied, three doses of nitrogen as urea are also applied during land preparation. The design used in the experiment is split-block design. The experiment is replicated 3 times. The number of germinated seeds per plot (out of 25 seeds per plot) are shown below :

Number of germinated seeds in different replications (y_{ijl})

Doses of phosphorus	Replications								
	1			2			3		
	Levels of nitrogen N_1 N_2 N_3			Levels of nitrogen N_1 N_2 N_3			Levels of nitrogen N_1 N_2 N_3		
P_1	18	18	20	17	19	19	18	20	18
P_2	21	20	22	20	20	21	22	21	21
P_3	22	23	23	22	23	24	23	23	24

- (i) Analyse the data and group the doses of phosphorus.
- (ii) Compare the means of P_2 and P_3 in presence of N_3 .
- (iii) Compare the means of N_2 and N_3 in presence of P_3 .
- (iv) Find the efficiency of this design compared to randomized block design in estimating the interaction of nitrogen and phosphorus.

Solution : (i) We have $r = 3$, $p = 3$, $q = 3$

The observations of phosphorus and replications (y_{ij})

Replications	Levels of phosphorus			Total $y_{i..}$
	P_1	P_2	P_3	
1	56	63	68	187
2	55	61	69	185
3	56	64	70	190
Total $y_{.j}$	167	188	207	562 = G

$$C.T. = \frac{G^2}{pqr} = \frac{(562)^2}{3 \times 3 \times 3} = 11697.9259.$$

$$SS(\text{Total}) = \sum \sum \sum y_{ijl}^2 - C.T. = 11804 - 11697.9259 = 106.0741.$$

$$SS(\text{Replications}) = \frac{1}{pq} \sum y_{i..}^2 - \text{C.T.} = \frac{105294}{3 \times 3} - 11697.9259 = 1.4074.$$

$$SS(\text{Phosphorus}) = \frac{1}{qr} \sum y_{.j.}^2 - \text{C.T.} = \frac{106082}{3 \times 3} - 11697.9259 = 88.9630.$$

$$\begin{aligned} SS(\text{Error-1}) &= \frac{1}{q} \sum \sum y_{ij.}^2 - \text{C.T.} - SS(\text{replications}) - SS(\text{Phosphorus}) \\ &= \frac{35368}{3} - 11697.9259 - 1.4074 - 88.9630 = 1.0370. \end{aligned}$$

The observations of nitrogen and replications ($y_{i.l}$)

Replications	Levels of nitrogen		
	N_1	N_2	N_3
1	61	61	65
2	59	62	64
3	63	64	63
Total $y_{.l}$	183	187	192

$$SS(\text{Nitrogen}) = \frac{1}{pr} \sum y_{.l}^2 - \text{C.T.} = \frac{105322}{3 \times 3} - 11697.9259 = 4.5185.$$

$$\begin{aligned} SS(\text{Error-2}) &= \frac{1}{p} \sum \sum y_{i.j}^2 - \text{C.T.} - SS(\text{Replications}) - SS(\text{Nitrogen}) \\ &= \frac{35122}{3} - 11697.9259 - 1.4074 - 4.5185. \\ &= 3.4815. \end{aligned}$$

The observations of phosphorus and nitrogen ($y_{.jl}$)

Levels of phosphorus	Levels of nitrogen		
	N_1	N_2	N_3
P_1	53	57	57
P_2	63	61	64
P_3	67	69	71

$$\begin{aligned} SS(\text{Phosphorus} \times \text{Nitrogen}) &= \frac{1}{r} \sum \sum y_{.jl}^2 - \text{C.T.} - SS(\text{Phosphorus}) - SS(\text{Nitrogen}) \\ &= \frac{35384}{3} - 11697.9259 - 88.9630 - 4.5185 = 3.2623. \end{aligned}$$

$$\begin{aligned} SS(\text{Error-3}) &= SS(\text{Total}) - SS(\text{Replications}) - SS(\text{Phosphorus}) - SS(\text{Error-1}) \\ &\quad - SS(\text{Nitrogen}) - SS(\text{Error-2}) - SS(\text{Phosphorus} \times \text{Nitrogen}) \\ &= 106.0741 - 1.4074 - 88.9630 - 1.0370 - 4.5185 - 3.4815 - 3.2623 \\ &= 3.4044. \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{.05}$
Replications	2	1.4074	0.7037	—	—
Phosphorus	2	88.9630	44.4815	171.58	6.94
Error-1	4	1.0370	0.25925	—	—
Nitrogen	2	4.5185	2.25925	2.60	6.94
Error-2	4	3.4815	0.8704	—	—
Phosphorus × nitrogen	4	3.2623	0.8165	1.92	3.84
Error-3	8	3.4044	0.42555	—	—
Total	26				

It is observed that the levels of phosphorus differ significantly in influencing the germination rate of a variety of maize. The means of germinated seeds under different levels of phosphorus are :

$$\bar{P}_1 = 18.556, \bar{P}_2 = 20.889, \bar{P}_3 = 23.000.$$

These means can be grouped using Duncan's multiple range test, where the test statistic is

$$D_k = d_{0.05, k, f} \sqrt{\frac{s_3}{qr}}, \quad k = 2, 3; \quad f = 4.$$

$$\text{We have } D_2 = 3.93 \sqrt{\frac{0.25925}{9}} = 0.113, \quad D_3 = 4.01 \sqrt{\frac{0.25925}{9}} = 0.115.$$

All the means are significantly different since pairwise difference in means are greater than D_2 . The three levels of phosphorus are in three different groups.

(ii) We need to compare the means $\bar{y}_{.23}$ and $\bar{y}_{.33}$. These means are $\bar{y}_{.23} = 21.33$ and $\bar{y}_{.33} = 23.67$. The test statistic to compare these means is

$$T_2 = \frac{r(\bar{y}_{.23} - \bar{y}_{.33})^2}{2S_{02}}, \quad \text{where}$$

$$S_{02} = \frac{1}{(p-1)(r-1)} \sum_i \sum_j (y_{ijl} - \bar{y}_{.jl} - \bar{y}_{i.l} + \bar{y}_{..l})^2; \quad l = 3$$

$$= \frac{1}{(p-1)(r-1)} \left[\sum \sum y_{ij3}^2 - C.T_2 - SS(\text{Replications})_2 - SS(\text{Phosphorus})_2 \right]$$

The observations of replications and phosphorus in presence of $N_3(y_{i.j3})$

Replications	Levels of phosphorus			Total $y_{i.3}$
	P_1	P_2	P_3	
1	20	22	23	65
2	19	21	24	64
3	18	21	24	63
Total $y_{.j3}$	57	64	71	$192 = G_2$

$$C.T_2 = \frac{(G_2)^2}{pr} = \frac{(192)^2}{3 \times 3} = 4096.00.$$

$$SS(\text{Replications})_2 = \frac{1}{p} \sum y_{i.3}^2 - C.T_2 = \frac{12290}{3} - 4096.00 = 0.6667.$$

$$SS(\text{Phosphorus})_2 = \frac{1}{r} \sum y_{.j3}^2 - C.T_2 = \frac{12386}{3} - 4096.00 = 32.6667.$$

$$S_{02} = \frac{1}{(p-1)(r-1)} [\sum \sum y_{ij3}^2 - C.T_2 - SS(\text{Replications})_2 - SS(\text{Phosphorus})_2]$$

$$= \frac{1}{2 \times 2} [4132 - 4096.00 - 0.6667 - 32.6667] = 0.66665.$$

Therefore, $T_2 = \frac{3(21.33 - 23.67)^2}{2 \times 0.66665} = 12.32.$

This $T_2 > F_{0.5;1,4} = 7.71$. The two means are significantly different.

(iii) We need to compare the means $\bar{y}_{.32} = 23.00$ and $\bar{y}_{.33} = 23.67$. It is observed that the levels of nitrogen do not differ significantly and hence, the comparison of means of different levels of nitrogen is not needed. However, we can compare the means in presence of a particular dose of phosphorus. The test statistic for this is

$$T_1 = \frac{r(\bar{y}_{.j1} - \bar{y}_{.j1'})^2}{2S_{01}}, \text{ where } S_{01} = \frac{1}{(q-1)(r-1)} \sum \sum (y_{ijt} - \bar{y}_{.jt} - \bar{y}_{ij.} + \bar{y}_{.j.})^2; j = 3.$$

The observations of nitrogen in presence of P_3 in different replications (y_{i3t})

Replications	Levels of nitrogen			Total y_{i3} .
	N_1	N_2	N_3	
1	22	23	23	68
2	22	23	24	69
3	23	23	24	70
Total $y_{.3t}$	67	69	71	207 = G_1

$$C.T_1 = \frac{G_1^2}{pq} = \frac{(207)^2}{3 \times 3} = 4761.00.$$

$$SS(\text{Replications})_1 = \frac{1}{q} \sum y_{i.3}^2 - C.T_1 = \frac{14285}{3} - 4761.00 = 0.6667$$

$$SS(\text{Nitrogen})_1 = \frac{1}{r} \sum y_{.3t}^2 - C.T_1 = \frac{14291}{3} - 4761.00 = 2.6667.$$

$$S_{01} = \frac{1}{(q-1)(r-1)} [\sum \sum y_{i3t}^2 - C.T_1 - SS(\text{Replications})_1 - SS(\text{Nitrogen})_1]$$

$$= \frac{1}{2 \times 2} [4765 - 4761.00 - 0.6667 - 2.6667] = 0.16665.$$

Therefore, $T_1 = \frac{3(23.00 - 23.67)^2}{2 \times 0.16665} = 4.04.$

But $T_1 < F_{0.05;1,4} = 7.71$. The two means do not differ significantly.

(iv) The efficiency of split-block design in estimating interaction (AB) of phosphorus and nitrogen compared to randomized block design is

$$\begin{aligned} & \frac{(p-1)s_3 + (q-1)s_5 + (p-1)(q-1)s_7}{(pq-1)s_7} \\ &= \frac{(3-1)0.25925 + (3-1)0.8704 + (3-1)(3-1)0.42555}{(9-1)0.42555} \\ &= \frac{3.9615}{3.4044} = 116.36\%. \end{aligned}$$

Chapter 5

Incomplete Block Design

5.1 Introduction

One of the disadvantages of randomized block design is that the block heterogeneity arises if number of plots are needed to allocate a large number of treatments in a block. This problem of allocation of a large number of treatments in a block is obviated introducing confounding technique in the experiment. The fractional replication of factorial experiment is also used to remove the problem of block heterogeneity in the experiment. The confounded technique or application of fractional replication of treatment does not provide information on some effects and interactions. Therefore, those interactions which are least important are confounded with blocks so that block contrast within a replication represents a higher order interaction.

In varietal trial experiment if number of variety is large, a portion of it cannot be used in the experiment and there is no scope to loose information on any effect of a variety. The experiment is to be conducted in such a way that all effects of varieties are estimated with equal efficiency without allocating all treatments in plots of a block. The design of such varietal trial experiment in the field of agriculture has first been introduced by Yates (1936). Later on, Yates, Fisher and Bose (1939) have developed the technique to allocate large number of treatments, in smaller number of plots of a block.

Let there be v treatments which are to be allocated in $k(k < v)$ plots of a block. Since number of plots is less than number of treatments, all treatments are not allocated within plots of a block and the block is called incomplete. However, the treatments are allocated in blocks in such a way that each pair of treatments is compared with same efficiency. Yates, Fisher and Bose (1939) have first introduced the technique of analysis of data obtained from such experiment.

The main objective of incomplete block design is to estimate the treatment effect without allocating all treatments in a block. Therefore, the two factor effects, viz., treatment effect and block effect are of interest. The observations are influenced by these two factors except the influence of uncontrolled source of variation (error). Therefore, the model to represent the data obtained from incomplete block design is

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl}.$$

$$i = 1, 2, \dots, b; j = 1, 2, \dots, v; l = 1, 2, \dots, n_{ij}.$$

Here y_{ijl} = l th observation of j -th treatment in i -th block, μ = general mean, α_i = effect of i -th block, β_j = effect of j -th treatment, $(\alpha\beta)_{ij}$ = interaction of j -th treatment with i -th block and e_{ijl} = random error.

In such experiment, if any $n_{ij} = 0$, the block can be considered incomplete. The analysis of such experimental data can be performed using the technique of analysis of randomized block design with missing observations. The general form of analysis of such experimental data is shown in section 2.4.

The value of $n_{ij} = 0$ indicates that in i -th block j -th treatment is missing and $n_{ij} = 1$ indicates that j -th treatment is allocated once in i -th block. Therefore, the block is incomplete if $n_{ij} = 0$ or 1. However, the j -th treatment is allocated in i -th block in such a way that all j -th and j' -th ($j \neq j' = 1, 2, \dots, v$) treatments can be compared with equal efficiency.

5.2 Balanced Incomplete Block (BIB) Design

Let there be v treatments which are to be allocated in b blocks of k plots ($k < v$) each. The allocation of v treatments in b blocks of k plots each is called BIB design if the treatments are allocated in such a way that

- (i) each treatment is replicated r times ($r < b$)
- (ii) each pair of treatments is replicated in λ blocks ($\lambda \leq r$).

In such an arrangement of treatments in blocks, if $b = v$, the design is called symmetric BIB design.

The design is introduced first for varietal trial experiment in the field of agriculture. However, the design is not suitable for varietal trial experiment if the treatments need more replications. Moreover, the design is not suitable for any number of treatments.

Relation among Parameters of BIB Design

It has already been mentioned that the number of observations of j -th treatment in i -th block is n_{ij} ($i = 1, 2, \dots, b; j = 1, 2, \dots, v$), where for BIB design $n_{ij} = 0$ or 1. The observations of all treatments in all blocks can be represented by a matrix N called incidence matrix, where

$$N = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1v} \\ n_{21} & n_{22} & \cdots & n_{2j} & \cdots & n_{2v} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} & \cdots & n_{iv} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ n_{b1} & n_{b2} & \cdots & n_{bj} & \cdots & n_{bv} \end{bmatrix}_{b \times v}$$

Here $\sum_j n_{ij} = N_{i.} = k$, $\sum_i n_{ij} = N_{.j} = r$, $\sum_i n_{ij}n_{is} = \lambda$, ($j \neq s$)

$$\sum_i \sum_j n_{ij} = n = \sum_i N_{i.} = \sum_j N_{.j} = bk = vr.$$

The values b , v , r , k and λ related to n_{ij} observations are called parameters of BIB design. Let us now discuss the relations of these parameters.

(i) $bk = vr$

Proof : The design has b blocks and each block has k plots. Hence, total number of plots are bk . The design is used to study the effects of v treatments. Each treatment is replicated r times. Therefore, total number of observations of v treatments are vr . These vr observations are obtained from bk plots. Since each plot contains maximum 1 treatment, $bk = vr$.

(ii) $\lambda(v - 1) = r(k - 1)$

Proof : From the incidence matrix, we have

$$N'N = \begin{bmatrix} r & \lambda & \lambda & \cdots & \lambda \\ \lambda & r & \lambda & \cdots & \lambda \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda & \lambda & \lambda & \cdots & r \end{bmatrix}_{v \times v}$$

Here $\sum_j n_{ij}^2 = k$, $\sum_i n_{ij}^2 = r$, $\sum_i n_{ij}n_{is} = \lambda$, ($j \neq s$).

Post multiplying $N'N$ by $\mathbf{1} = (1, 1, \dots, 1)'_{1 \times v}$, we have

$$N'N \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{v \times 1} = \begin{bmatrix} r + \lambda(v-1) \\ r + \lambda(v-1) \\ \dots \\ r + \lambda(v-1) \end{bmatrix}_{v \times 1}$$

Again, $N \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{v \times 1} = \begin{bmatrix} k \\ k \\ \vdots \\ k \end{bmatrix}_{b \times 1}$

and $N'N \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} n_{11} & n_{21} & \dots & n_{b1} \\ n_{12} & n_{22} & \dots & n_{b2} \\ \dots & \dots & \dots & \dots \\ n_{1v} & n_{2v} & \dots & n_{bv} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} rk \\ rk \\ \vdots \\ rk \end{bmatrix}_{v \times 1}$

From both the results, we have

$$r + \lambda(v-1) = rk \quad \text{or,} \quad \lambda(v-1) = r(k-1).$$

(iii) $r > \lambda$

Proof: We have $\lambda(v-1) = r(k-1)$. From the condition of BIB design, we have $k-1 < v-1$ or, $v-1 > k-1$. Therefore, to maintain equality $\lambda(v-1) = r(k-1)$ λ must be less than r .

$\therefore r > \lambda$.

(iv) $b \geq v$ [Fisher's inequality]

Proof: We have

$$N'N = \begin{bmatrix} r & \lambda & \lambda & \dots & \lambda \\ \lambda & r & \lambda & \dots & \lambda \\ \dots & \dots & \dots & \dots & \dots \\ \lambda & \lambda & \lambda & \dots & r \end{bmatrix}_{v \times v} \quad \text{and} \quad |N'N| = \begin{vmatrix} r & \lambda & \lambda & \dots & \lambda \\ \lambda & r & \lambda & \dots & \lambda \\ \dots & \dots & \dots & \dots & \dots \\ \lambda & \lambda & \lambda & \dots & r \end{vmatrix}.$$

Subtracting first column from other columns, we have

$$|N'N| = \begin{vmatrix} r & \lambda-r & \lambda-r & \dots & \lambda-r \\ \lambda & r-\lambda & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \lambda & 0 & 0 & \dots & r-\lambda \end{vmatrix}.$$

Adding the sum of all rows except first row with the first row, we get

$$|N'N| = \begin{vmatrix} r + \lambda(v-1) & 0 & 0 & 0 \\ \lambda & r-\lambda & 0 & 0 \\ \lambda & 0 & r-\lambda & 0 \\ \dots & \dots & \dots & \dots \\ \lambda & 0 & 0 & r-\lambda \end{vmatrix} = [r + \lambda(v-1)][r-\lambda]^{v-1}.$$

$$\therefore |N'N| \neq 0, \quad \because r > \lambda.$$

$$\text{Hence, } r(N'N) = r(N') = r(N) = v.$$

$$\therefore b \geq v \text{ [since row rank is equal to column rank].}$$

$$(v) \quad r \geq k$$

Proof : We know $bk = vr$ and $b \geq v$.

$$\therefore k \leq r \text{ or, } r \geq k.$$

$$(vi) \quad b \geq v + r - k$$

Proof : We have $b \geq v, r \geq k$.

$$\therefore v(r - k) \geq k(r - k) \quad [\because k < v]$$

$$vr \geq vk + rk - k^2$$

$$bk \geq vk + rk - k^2.$$

$$\therefore b \geq v + r - k.$$

(vii) For a symmetric BIB design $(r - \lambda)$ is a perfect square.

$$\begin{aligned} \text{Proof : We have } |N'N| &= [r + \lambda(v - 1)][r - \lambda]^{v-1} = [r + r(k - 1)][r - \lambda]^{v-1} \\ &= rk(r - \lambda)^{v-1}. \end{aligned}$$

For a symmetric BIB design $b = v$ and hence, $r = k$ ($\because bk = vr$).

$$\text{Again, } |N'N| = |N|^2 = r^2(r - v)^{v-1}.$$

$$\therefore |N| = \pm r(r - \lambda)^{(v-1)/2}.$$

But $|N|$ is an integer [since the elements in N are either 0 or 1] and r is also an integer. Therefore, $(r - \lambda)^{(v-1)/2}$ must be an integer and hence, $(r - \lambda)^{1/2}$ is an integer. Hence, $(r - \lambda)$ is a perfect square.

(viii) For a BIB design if b is divisible by r , then $b \geq v + r - 1$.

Proof : If b is divisible by r , let us take $b = nr$, where n is an integer and greater than 1. We know

$$\lambda(v - 1) = r(k - 1)$$

$$\begin{aligned} \text{or, } r &= \frac{\lambda(v - 1)}{k - 1} = \frac{\lambda(nk - 1)}{k - 1} = \frac{\lambda(n - 1)}{k - 1} + \lambda n \\ &= +ve \text{ quantity } (\because n > 1, k > 1). \end{aligned}$$

$$\text{Again, } b \geq v, \quad \therefore b \geq v + r - 1.$$

If this is not true, $b < v + r - 1$.

$$\therefore nr < v + r - 1, \quad r(n - 1) < v - 1, \quad r(n - 1) < \frac{r}{\lambda}(k - 1)$$

$$\Rightarrow \frac{\lambda(n - 1)}{k - 1} < 1.$$

This is not true and hence, $b < v + r - 1$ is not true. Therefore, $b \geq v + r - 1$.

(ix) For a symmetric BIB design $\gamma_{ii'} = \lambda$, where $\gamma_{ii'} = \sum_j n_{ij}n_{i'j}$, $i \neq i' = 1, 2, \dots, b$.

Proof : From the incidence matrix, we have

$$\begin{aligned}
 NN' &= \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1v} \\ n_{21} & n_{22} & \cdots & n_{2j} & \cdots & n_{2v} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ n_{b1} & n_{b2} & \cdots & n_{bj} & \cdots & n_{bv} \end{bmatrix} \begin{bmatrix} n_{11} & n_{21} & \cdots & n_{b1} \\ n_{12} & n_{22} & \cdots & n_{b2} \\ \cdots & \cdots & \cdots & \cdots \\ n_{1v} & n_{2v} & \cdots & n_{bv} \end{bmatrix} \\
 &= \begin{bmatrix} k & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1b} \\ \gamma_{21} & k & \gamma_{23} & \cdots & \gamma_{2b} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{b1} & \gamma_{b2} & \gamma_{b3} & \cdots & k \end{bmatrix}_{b \times b}
 \end{aligned}$$

Since $b = v$ and $r(N) = v$, N^{-1} is available and $NN^{-1} = I$.

$$\therefore NN' = NN'NN^{-1} = N(N'N)N^{-1} \tag{a}$$

$$\text{Again, } N'N = \begin{bmatrix} r & \lambda & \lambda & \cdots & \lambda \\ \lambda & r & \lambda & \cdots & \lambda \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda & \lambda & \lambda & \cdots & r \end{bmatrix} = (r - \lambda)I_{v \times v} + \lambda \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{v \times v}$$

$$\begin{aligned}
 \therefore N(N'N) &= (r - \lambda)N + N \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [1 \ 1 \ \cdots \ 1]_{1 \times v} \lambda \\
 &= (r - \lambda)N + \lambda \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [1 \ 1 \ \cdots \ 1] k.
 \end{aligned}$$

$$\begin{aligned}
 N(N'N)N^{-1} &= (r - \lambda)NN^{-1} + \lambda \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [1 \ 1 \ \cdots \ 1] kN^{-1} \\
 &= (r - \lambda)I_v + \lambda \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [1 \ 1 \ \cdots \ 1] NN^{-1} \quad [\because r = k] \\
 &= \begin{bmatrix} r & \lambda & \lambda & \cdots & \lambda \\ \lambda & r & \lambda & \cdots & \lambda \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda & \lambda & \lambda & \cdots & r \end{bmatrix} \tag{b}
 \end{aligned}$$

$$\therefore \gamma_{ii'} = \lambda \text{ [from (a) and (b)]}$$

5.3 Analysis of BIB Design

The BIB design is constructed in such a way that a group of k treatments are allocated in plots of a block. Again, the group of k treatments are allocated to a block randomly. Thus, there are two steps of randomisation in allocating treatments to the plots of blocks. Since a group of treatments are randomly selected for a block, the block effect may be random. The block effect may also be considered as fixed effect. Thus, there are two types of analysis of the data of BIB design. The analysis of data assuming fixed block effect is called intra-block analysis and if block effect is considered as random variable, the analysis is called inter-block analysis.

Since all blocks are not of same types, the treatment effect varies with the variation in block effect. Therefore, some information of treatments are retained in blocks. The inter-block analysis is done to recover the information which is retained with block. Thus, the treatment effect is also estimated using intra-block and inter-block estimate. This estimate is known as combined intra- and inter-block estimate with recovery of inter-block information. In the present section, all three types of analysis will be discussed.

Intra-block Analysis of BIB Design : The model for this analysis is

$$y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl}, \quad (A)$$

where y_{ijl} = l th observation of j -th treatment in i -th block, μ = general mean, α_i = effect of i -th block, β_j = effect of j -th treatment and e_{ijl} = random error [$i = 1, 2, \dots, b$; $j = 1, 2, \dots, v$; $l = n_{ij} = 0$ or 1]. Since $n_{ij} = 0$ or 1 , the interaction $(\alpha\beta)_{ij}$ considered in the model is dropped.

Assumption : The error term e_{ijl} is distributed normally with mean zero and variance σ^2 .

The normal equations to estimate the parameters are :

$$y_{...} = bk\hat{\mu} + k \sum \hat{\alpha}_i + r \sum \hat{\beta}_j \quad (a)$$

$$y_{i..} = k\hat{\mu} + k\hat{\alpha}_i + \sum n_{ij}\hat{\beta}_j \quad (b)$$

$$y_{.j.} = r\hat{\mu} + \sum n_{ij}\hat{\alpha}_i + r\hat{\beta}_j. \quad (c)$$

Replacing the value of $\hat{\alpha}_i = \frac{y_{i..}}{k} - \hat{\mu} - \frac{1}{k} \sum n_{ij}\hat{\beta}_j$ from (b) in (c), we get

$$C_{j1}\hat{\beta}_1 + C_{j2}\hat{\beta}_2 + \dots + C_{jj}\hat{\beta}_j + \dots + C_{jv}\hat{\beta}_v = Q_j; j = 1, 2, \dots, v,$$

where $C_{jj} = N_{.j} - \sum_{i0}^b \frac{n_{ij}^2}{N_i} = r - \frac{r}{k} = \frac{r}{k}(k-1) = \frac{\lambda(v-1)}{k}$

$$C_{js} = - \sum \frac{n_{ij}n_{is}}{N_i} = -\frac{\lambda}{k}, j \neq s = 1, 2, \dots, v.$$

$$\therefore \frac{\lambda}{k}(v-1)\hat{\beta}_j + \frac{\lambda}{k}\hat{\beta}_j - \frac{\lambda}{k} \sum_j^v \hat{\beta}_j = Q_j, Q_j = y_{.j.} - \frac{1}{k} \sum n_{ij}y_{i..}$$

Putting the restriction $\sum \hat{\beta}_j = 0$, we get

$$\frac{1}{k}\hat{\beta}_j[\lambda(v-1) + \lambda] = Q_j \Rightarrow \hat{\beta}_j = \frac{Q_j}{rE}, \text{ where } r = \frac{\lambda v}{rk}.$$

The sum of squares due to estimate is :

$$SS \text{ (Estimates)} = \frac{1}{k} \sum y_{i..}^2 + \frac{1}{rE} \sum Q_j^2 \tag{d}$$

and $S_3 = SS \text{ (Error)} = \sum \sum \sum y_{ijl}^2 + \frac{1}{k} \sum y_{i..}^2 - \frac{1}{rE} \sum Q_j^2.$

The objective of the analysis is to test the significance of $H_0 : \beta_1 = \beta_2 = \dots = \beta_v.$

Under H_0 , the model (A) stands

$$y_{ijl} = \mu + \alpha_i + e_{ijl}. \tag{B}$$

The sum of squares due to estimates of this model is

$$SS \text{ (Estimates)} = \frac{1}{k} \sum y_{i..}^2. \tag{e}$$

Hence, sum of squares due to treatment under H_0 is (d) - (e), where

$$S_2 = SS \text{ (Treatments)}_{\text{adjusted}} = \frac{1}{rE} \sum Q_j^2.$$

The block sum of squares unadjusted (ignoring treatment) is

$$S_1 = SS \text{ (Blocks)} = \frac{1}{k} \sum y_{i..}^2 - \text{C.T.}, \quad \text{C.T.} = \frac{G^2}{bk}.$$

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{\text{d.f.}}$	F
Blocks (unadjusted)	$b - 1$	S_1	s_1	$F_1 = \frac{s_1}{s_3}$
Treatments (adjusted)	$v - 1$	S_2	s_2	$F_2 = \frac{s_2}{s_3}$
Intra-block error	$bk - b - v + 1$	S_3	s_3	—
Total	$bk - 1$			

If $F_2 \geq F_{0.05; v-1, bk-b-v+1}$, H_0 is rejected.

The variance of the estimate of contrast $\beta_j - \beta_s$ is

$$V(\hat{\beta}_j - \hat{\beta}_s) = V \left[\frac{1}{rE} (Q_j - Q_s) \right] = \frac{2\sigma^2}{rE},$$

where $V(\theta_j) = C_{jj}\sigma^2$ and $\text{Cov}(\theta_j, \theta_s) = C_{js}\sigma^2$. The estimate of this variance is $2s_3/rE$, where s_3 is the mean square intra-block error. The estimate of variance of the estimate of contrast $\sum d_i\beta_j$ is $\frac{s_3}{rE} \sum d_j^2$, where $\sum d_j = 0$. Therefore, usual t -test can be performed to test the significance of any contrast of treatment effects.

The variance of $\hat{\beta}_j - \hat{\beta}_s$ in BIB design is

$$V(\hat{\beta}_j - \hat{\beta}_s) = \frac{2\sigma^2}{rE}, \quad \text{where } \sigma^2 = \text{intra-block error variance}$$

Let the variance of the estimated contrast $\hat{\beta}_j - \hat{\beta}_s$ be

$$V(\hat{\beta}_j - \hat{\beta}_s) = \frac{2\sigma_R^2}{r},$$

if the experiment is conducted through randomized block design, where σ_R^2 is the error variance in case of randomized block design. Therefore, the efficiency of BIB design compared to randomized block design is

$$\text{Efficiency of BIB design} = \frac{2\sigma_R^2}{r} / \frac{2\sigma^2}{rE} = E \frac{\sigma_R^2}{\sigma^2}.$$

$$\text{Here } E = \frac{\lambda v}{rk} = \frac{\lambda v - \lambda + \lambda}{rk - r + r} = \frac{\lambda(v-1) + \lambda}{r(k-1) + r} = \frac{r(k-1) + \lambda}{r(k-1) + r} < 1;$$

since $\lambda > r$. The efficiency of BIB design is less than that of RBD.

Inter-block Analysis : Yates (1940) has proposed this analysis using the block total, where block total is

$$y_{i..} = k\mu + k\alpha_i + \sum n_{ij}\beta_j + e_{i..}$$

The original model for BIB design is

$$y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl},$$

$i = 1, 2, \dots, b; j = 1, 2, \dots, v; l = n_{ij} = 0 \text{ or } 1$

The block effect α_i is assumed random variable and hence, the assumption for this analysis is

$$E(\alpha_i) = 0, \quad E(\alpha_i, \alpha_{i'}) = \sigma_\alpha^2, \quad \text{if } i = i' = 1, 2, \dots, b \\ = 0, \quad \text{otherwise}$$

and $E(\alpha_i, e_{ij}) = 0$.

It is observed that $V(y_{i..}) = k(\sigma^2 + k\sigma_\alpha^2)$, where $E(e_{ij}) = 0$ and $V(e_{ij}) = \sigma^2$.

Since block effect is random variable, the parameters μ and β_j are to be estimated by minimizing the error sum of squares.

$$\phi = \sum (y_{i..} - k\hat{\mu} - \sum n_{ij}\hat{\beta}_j)^2.$$

The normal equations are :

$$y_{...} = bk\hat{\mu} + r \sum \hat{\beta}_j \tag{f}$$

$$\sum n_{ij}y_{i..} = rk\hat{\mu} + \sum_i n_{ij} \sum_j n_{ij}\hat{\beta}_j. \tag{g}$$

The equation (g) can be written as

$$\sum n_{ij} \left(\frac{y_{i..}}{k} - \frac{y_{...}}{bk} \right) = \hat{\beta}_j \left(\frac{1}{k} - \frac{\lambda}{k} \right) + \frac{\lambda}{k} \sum \hat{\beta}_j.$$

Under the restriction $\sum \hat{\beta}_j = 0$, we have

$$\hat{\beta}_j = \frac{P_j}{\frac{1}{k}(r-\lambda)}, \quad \text{where } P_j = \sum_i n_{ij} \left(\frac{y_{i..}}{k} - \frac{y_{...}}{bk} \right).$$

The inter-block adjusted treatment sum of squares is

$$S_1 = SS(\text{Treatment})_{\text{adjusted}} = \frac{k}{r-\lambda} \sum P_j^2.$$

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	$E(MS)$	F
Treatment (adjusted)	$v - 1$	S_1	s_1	$\sigma^2 + k\sigma_\alpha^2 + \frac{r-\lambda}{k(v-1)} \sum \beta_j^2$	$\frac{s_1}{s_2}$
Inter-block error	$b - v$	S_2	s_3	$\sigma^2 + k\sigma_\alpha^2$	
Total	$b - 1$				

The analysis is done if $b \neq v$ and in that case, the usual F -test statistic is calculated to test the significance of $H_0 : \beta_j = 0$. The variance of $\hat{\beta}_j - \hat{\beta}_s$ is given by :

$$V(\hat{\beta}_j - \hat{\beta}_s) = \frac{2k(\sigma^2 + k\sigma_\alpha^2)}{r - \lambda}$$

where inter-block error mean square s_2 is the estimate of $(\sigma^2 + k\sigma_\alpha^2)$. Therefore, the significance of the contrast of the type $\beta_j - \beta_s$ ($j \neq s = 1, 2, \dots, v$) can be tested by t -test.

Combined Intra- and Inter-block Analysis of BIB Design : It has already been mentioned that a group of k treatments selected from v treatments are randomly allocated to blocks and hence, block effect is random variable. Due to variability in blocks the treatment behaves differently in different blocks. In such a situation only intra-block estimate of treatment effect is not sufficient to get the information on treatment. Some information on treatment is retained in the blocks. Thus, we need the combined intra- and inter-block estimate of treatment effect. The combined estimate is known as estimate of treatment effect after recovery of inter-block information, where the recovered estimate of treatment effect is

$$\hat{\beta}_j \text{ (inter-block)} = \frac{P_j}{\frac{1}{k}(r - \lambda)}$$

The combined estimate is found out as follows :

The model for BIB design is

$$y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl}$$

$i = 1, 2, \dots, b; j = 1, 2, \dots, v; l = n_{ij} = 0$ or 1 ,

We have already observed that the intra-block and inter-block error variances are σ^2 and $\sigma^2 + k\sigma_\alpha^2$, respectively. Let $W = \frac{1}{\sigma^2}$ and $W_1 = \frac{1}{(\sigma^2 + k\sigma_\alpha^2)}$. Then weighted error sum of squares for combined intra- and inter-block analysis is

$$\phi = W \sum \sum \sum (y_{ijl} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 + \frac{W_1}{k} \sum \left[y_{i..} - k\hat{\mu} - \sum_j n_{ij}\hat{\beta}_j \right]^2$$

The $\hat{\beta}_j$ is found out from the equation $\frac{\partial \phi}{\partial \hat{\beta}_j} = 0$. We have

$$W \left[y_{.j} - r\hat{\mu} - \sum n_{ij}\hat{\alpha}_i - r\hat{\beta}_j \right] + \frac{W_1}{k} \left[\sum n_{ij}y_{i..} - rk\hat{\mu} - \sum_i n_{ij} \sum_j n_{ij}\hat{\beta}_j \right] = 0$$

From intra-block analysis, we have

$$\hat{\alpha}_i = \frac{1}{k}y_{i..} - \hat{\mu} - \frac{1}{k} \sum n_{ij}\hat{\beta}_j$$

Putting the value of $\hat{\alpha}_i$ in the above equation and on simplification, we get

$$WQ_j + W_1P_j = \left[rEW + \frac{W_1}{k}(r - \lambda) \right] \hat{\beta}_j.$$

$$\therefore \hat{\beta}_j = \frac{WQ_j + W_1P_j}{rEW + \frac{W_1}{k}(r - \lambda)} = \frac{Q_j + \rho P_j}{rE + \frac{\rho}{k}(r - \lambda)}, \quad \rho = \frac{W_1}{W}.$$

If $b \neq v$, the combined estimate of β_j can also be obtained from weighted estimates of intra- and inter-block estimates. The variances of intra- and inter-block estimates of β_j are respectively

$$V(\hat{\beta}_j) \text{ intra-block} = \frac{(v-1)\sigma^2}{vrE} \quad \text{and} \quad V(\hat{\beta}_j) \text{ inter-block} = \frac{(v-1)(\sigma^2 + k\sigma_\alpha^2)}{\frac{v}{k}(r-\lambda)}.$$

Let $w = \frac{vrE}{(v-1)\sigma^2}$ and $w_1 = \frac{v(r-\lambda)}{k(v-1)(\sigma^2 + k\sigma_\alpha^2)}$. Then

$$w + w_1 = \frac{v}{v-1} \left[\frac{rE}{\sigma^2} + \frac{r-\lambda}{k(\sigma^2 + k\sigma_\alpha^2)} \right] = \frac{Wv}{v-1} \left[rE + \frac{\rho}{k}(r-\lambda) \right].$$

Therefore, the weighted estimate of β_j is :

$$\begin{aligned} \hat{\beta}_j \text{ (weighted)} &= \frac{1}{w + w_1} \left[\frac{Q_j}{rE} \frac{vrE}{(v-1)\sigma^2} + \frac{kP_j}{(r-\lambda)} \frac{v(r-\lambda)}{k(v-1)(\sigma^2 + k\sigma_\alpha^2)} \right] \\ &= \frac{Q_j + \rho P_j}{rE + \frac{\rho}{k}(r-\lambda)}. \end{aligned}$$

However, the value of W and W_1 and hence, ρ are unknown. These values can be estimated. Let the estimate of W be $\hat{W} = \frac{1}{s^2}$, where s^2 is the intra-block error mean square. Let the estimate of σ_α^2 be s_α^2 , where

$$s_\alpha^2 = \frac{1}{b-1} S_\alpha^2$$

and $S_\alpha^2 = SS$ (Treatment) adjusted + SS (Blocks) unadjusted - SS (Treatment) unadjusted.

Here $E(s_\alpha^2) = \sigma^2 + \frac{v(r-1)}{b-1}\sigma_\alpha^2$. Also $E(s^2) = \sigma^2$.

$$E(s_\alpha^2 - s^2) = \frac{v(r-1)}{b-1}\sigma_\alpha^2.$$

$$\therefore \hat{\sigma}_\alpha^2 = \frac{(b-1)(s_\alpha^2 - s^2)}{v(r-1)}.$$

Therefore, the estimate of W_1 is

$$\hat{W}_1 = \frac{v(r-1)}{k(b-1)s_\alpha^2 - (v-k)s^2}.$$

Hence, ρ can be estimated and the estimate of combined treatment effect is

$$\hat{\beta}_j \text{ (combined)} = \frac{Q_j + \hat{\rho}P_j}{rE + \frac{\hat{\rho}}{k}(r-\lambda)}.$$

The above estimate of β_j is not unbiased since it is found out using $\hat{\rho}$. The value w and w_1 can be estimated independently (if $b \neq v$). Hence, the bias of order $\sum \frac{1}{f_i}$ can be removed from the estimate of β_j where $f_1 = bk - b - v + 1$ and $f_2 = b - v$. The adjusted estimate of β_j [Meir (1953)] is

$$\hat{\beta}_j \text{ (combined) adjusted} = \frac{\hat{w}\hat{\beta}_j \text{ (intra)} + \hat{w}_1\hat{\beta}_j \text{ (inter)}}{\hat{w} + \hat{w}_1} - \sum \frac{1}{f_i} \left[\frac{\partial^2 R}{\partial x_i^2} \right] \text{ for all } x_i.$$

Here $R = \frac{\hat{w}\hat{\beta}_j \text{ (intra)} + \hat{w}_1\hat{\beta}_j \text{ (inter)}}{\hat{w} + \hat{w}_1}$, $x_1 = \frac{\hat{w}}{w}$, $x_2 = \frac{\hat{w}_1}{w}$.

Here the adjusted combined intra- and inter-block estimate of treatment effect is suggested. However, such estimate is not needed if $s_\alpha^2 \leq s^2$.

The inter-block analysis is possible if $b > v$. Let F_2 be the test statistic to test the significance of $H_0 : \beta_j = 0$. Again, let F_1 be the test statistic for the same hypothesis in intra-block analysis. Using these two test statistics we can infer about the significance of combined effect of β_j . This is possible by combining the two test statistics result F_1 and F_2 . We have

$$p_i = P \left[F \geq \frac{F_i}{H_0} \right], \quad i = 1, 2.$$

It is known that intra-block analysis is more powerful than the inter-block analysis. Therefore, the combination of tests is to be done in such a way that F_1 gets maximum weight. Thus, the critical region for the combined test is to be found out so that

$$\omega : \{p_1 p_2^\theta \leq C\}, \quad 0 \leq \theta \leq 1.$$

where ω = critical region. The null hypothesis of insignificance of combined treatment effect is rejected if $p_1 p_2^\theta \leq C_\alpha$. The value of C_α is to be found out in such a way that

$$p\{p_1 p_2^\theta \leq C_\alpha\} = \alpha$$

If $\theta = 0$, only the intra-block test statistic F_1 is to be used. If $\theta = 1$, then the test statistic is

$$\chi^2 = -2 \sum_{i=1}^2 \ln p_i$$

This χ^2 has 2×2 d.f. In practice,

$$\theta = \frac{1 - E}{E} \left(\frac{\sigma^2}{\sigma^2 + k\sigma_\alpha^2} \right),$$

where σ^2 is to be replaced by intra-block error mean square and $(\sigma^2 + k\sigma_\alpha^2)$ is to be replaced by inter-block error mean square.

Example 5.1 : In a laboratory an experiment is conducted to study the impacts of pesticides on heart rate of one kind of slug. The pesticides are applied on different slugs according to the plan of BIB design with parameter $b = 12$, $v = 9$, $k = 3$, $r = 4$ and $\lambda = 1$. The arrangement of pesticides and the heart rate per minute of slugs under different pesticides are shown below :

Block	Treatment	Heart rate under treatment	Block total $y_{i..}$	Block	Treatment	Heart rate under treatment	Block total $y_{i..}$
1	1 2 3	10 10 8	28	7	1 6 8	11 8 9	28
2	4 5 6	10 9 8	27	8	2 4 9	10 9 8	27
3	7 8 9	11 10 10	31	9	3 5 7	8 8 10	26
4	1 4 7	9 9 10	28	10	1 5 9	10 8 9	27
5	2 5 8	11 10 9	30	11	2 6 7	10 8 10	28
6	3 6 9	9 8 10	27	12	3 4 8	8 8 9	25

- (i) Group the pesticides, if possible, from intra-block analysis.
(ii) Find combined intra- and inter-block estimate of pesticide effect.
(iii) Find the efficiency of BIB design compared to randomized block design.

Solution : (i) We have $b = 12$, $v = 9$, $r = 4$, $k = 3$, $\lambda = 1$.

Table to estimate treatment effect (intra-block analysis)

Treatment	Total of treatment $y_{.j}$	Total of block totals in which j th treatment is present $\sum_i n_{ij}y_{i..}$	Adjusted treatment total $Q_j = y_{.j} - \frac{1}{k} \sum_i n_{ij}y_{i..}$	Treatment effect $\hat{\beta}_j = Q_j/rE$ $E = \frac{\lambda v}{rk} = 0.75$ $rE = 3.00$
1	40	111	3.00	1.0000
2	41	113	3.3333	1.1111
3	33	106	-2.3333	-0.7778
4	36	107	0.3333	0.1111
5	35	110	-1.6666	-0.5555
6	32	110	-4.6666	-1.5555
7	41	113	3.3333	1.1111
8	37	114	-1.0000	-0.3333
9	37	112	-0.3333	-0.1111
Total	332		0.00	

$$G = 332, \text{C.T.} = \frac{G^2}{bk} = \frac{(332)^2}{12 \times 3} = 3061.7778.$$

$$SS (\text{Total}) = \sum \sum \sum y_{ijl}^2 - \text{C.T.} = 3096 - 3061.7778 = 34.2222.$$

$$SS (\text{Blocks}) = \frac{1}{k} \sum y_{i..}^2 - \text{C.T.} = \frac{9214}{3} - 3061.7778 = 9.5555.$$

$$SS (\text{Treatment}) \text{ adjusted} = \frac{1}{rE} \sum Q_j^2 = 20.8143.$$

$$SS (\text{Intra-block error}) = SS (\text{Total}) - SS (\text{Blocks}) - SS (\text{Treatment}) \text{ adjusted} \\ = 34.2222 - 9.5555 - 20.8143 = 3.8524.$$

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$
Blocks (unadjusted)	11	9.5555	0.8687	3.61	2.45
Treatment (adjusted)	8	20.8143	2.6018	10.84	2.59
Intra-block error	16	3.8524	0.2408	—	—
Total	35				

It is observed that the pesticides differ significantly in influencing the heart beat of slugs. The pesticides can be grouped using Duncan's multiple range test, where the test statistic is :

$$D_h = d_{0.05,h,f} \sqrt{\frac{s_3}{rE}}, \quad h = 2, 3, \dots, 9; \quad f = 16, \quad s_3 = 0.2408$$

$$D_2 = 3.00 \sqrt{\frac{0.2408}{3}} = 0.8499, \quad D_3 = 3.15 \sqrt{\frac{0.2408}{3}} = 0.8924,$$

$$D_4 = 3.23 \sqrt{\frac{0.2408}{3}} = 0.9151, \quad D_5 = 3.30 \sqrt{\frac{0.2408}{3}} = 0.9343,$$

$$D_6 = 3.34 \sqrt{\frac{0.2408}{3}} = 0.9463, \quad D_7 = 3.37 \sqrt{\frac{0.2408}{3}} = 0.9548,$$

$$D_8 = 3.39 \sqrt{\frac{0.2408}{3}} = 0.9604, \quad D_9 = 3.41 \sqrt{\frac{0.2408}{3}} = 0.9661.$$

$\hat{\beta}_7 - \hat{\beta}_6 = 2.6666 > D_9$, all the effects are significantly different.

$\hat{\beta}_1 - \hat{\beta}_6 = 2.5555 > D_7$, T_1 and T_6 are different.

$\hat{\beta}_4 - \hat{\beta}_3 = 1.8889 > D_8$, T_3 and T_7 are different.

$\hat{\beta}_4 - \hat{\beta}_6 = 1.6666 > D_6$, T_4 and T_6 are different.

$\hat{\beta}_1 - \hat{\beta}_3 = 1.7778 > D_6$, T_1 and T_3 are different.

$\hat{\beta}_1 - \hat{\beta}_5 = 1.5555 > D_5$, T_1 and T_5 are different.

$\hat{\beta}_4 - \hat{\beta}_3 = 0.8889 < D_5$, T_3 and T_4 are not different.

In a similar way all pairs can be compared. The treatments which do not differ are underlined below :

$\bar{T}_6, \bar{T}_3, \bar{T}_5, \bar{T}_8, \bar{T}_9, \bar{T}_4, \bar{T}_1, \bar{T}_2, \bar{T}_7$, where T_j is the j -th treatment (pesticide).

(ii) Inter-block estimate of treatment effects are as follows :

Treatment	$P_j = \sum_i n_{ij} \left(\frac{y_{i..}}{k} - \frac{y_{...}}{bk} \right)$	$\hat{\beta}_j = \frac{P_j}{\frac{1}{k}(r-\lambda)}$
1	0.1111	0.1111
2	0.7778	0.7778
3	-1.5555	-1.5555
4	-1.2222	-1.2222
5	-0.2222	-0.2222
6	-0.2222	-0.2222
7	0.7778	0.7778
8	1.1111	1.1111
9	0.4444	0.4444

$$SS \text{ (Treatment) adjusted from inter-block analysis} = \frac{k}{r-\lambda} \sum P_j^2 = 6.6664.$$

$$SS \text{ (Inter-block error)} = \frac{1}{k} \sum y_{i00}^2 - \text{C.T.} - SS \text{ (Treatment) adjusted} \\ = 9.5555 - 6.6664 = 2.8891.$$

Now, $\hat{W} = 1/MS \text{ (Intra-block error)} = 4.1528.$

Again, $\hat{W}_1 = \frac{v(r-1)}{k(b-1)s_\alpha^2 - (v-k)s^2},$

where $s_\alpha^2 = \frac{S_\alpha^2}{b-1}, s^2 = MS \text{ (Intra-block error)} = 0.2408.$

$$S_\alpha^2 = SS \text{ (Treatment) adjusted} + SS \text{ (Blocks) unadjusted} - SS \text{ (Treatment)}$$

$$SS \text{ (Treatment)} = \frac{1}{r} \sum y_{.j.}^2 - \text{C.T.} = \frac{12334}{4} - 3061.7778 = 21.7222.$$

$$S_\alpha^2 = 20.8143 + 9.5555 - 21.7222 = 8.6476, s_\alpha^2 = 0.7861.$$

$$\therefore \hat{W}_1 = \frac{9(4-1)}{3(12-1)0.7861 - (9-3)0.2408} = \frac{27}{24.4965} = 1.1022.$$

Therefore, $\hat{\rho} = \frac{\hat{W}_1}{\hat{W}} = \frac{1.1022}{4.1528} = 0.2654.$

The combined intra- and inter-block estimate of treatment effects are as follows :

Treatment	$\hat{\beta}_j = \frac{Q_j + \hat{\rho}P_j}{rE + \frac{\hat{\rho}}{k}(r-\lambda)}$	Treatment	$\hat{\beta}_j = \frac{Q_j + \hat{\rho}P_j}{rE + \frac{\hat{\rho}}{k}(r-\lambda)}$
1	0.9277	6	-1.4472
2	1.0840	7	1.0840
3	-0.8410	8	-0.2159
4	0.0027	9	-0.0659
5	-0.5284		

(iii). SS (Error) from randomized block design analysis

$$= SS \text{ (Total)} - SS \text{ (Treatment)} - SS \text{ (Blocks)} = 34.2222 - 21.7222 - 9.5555 = 2.9445.$$

$$\hat{\sigma}_R = \frac{SS \text{ (Error) from RBD}}{\text{d.f. (Error)}} = 0.1840.$$

Again, $\hat{\sigma}_R = 0.2408$ (from BIB design).

Therefore, efficiency of BIB design compared to RB design is :

$$E \frac{\hat{\sigma}_R^2}{\sigma^2} = 0.75 \frac{0.1840}{0.2408} = 57.31\%$$

Estimation of Missing Value in BIB Design.

Let the observation of j -th treatment in i -th block be missing. Let this l -th observation of j -th treatment be x . We need to estimate the value of x so that the intra-block error sum of squares is minimum.

During intra-block analysis it is observed that the adjusted total of j -th treatment is

$$Q_j = y_{.j} - \frac{1}{k} \sum n_{ij} y_{i..} = y_{.j} - \frac{1}{k} B_j,$$

where B_j = total of block totals of those blocks in which j -th treatment is present. Since j -th treatment is missing in i -th block, the adjusted total of j -th treatment is :

$$Q_j = y'_{.j} + (y_{.j} + x) - \frac{1}{k} \beta'_j - \frac{1}{k} (y_{i..} + x).$$

Here B'_j = total of block totals of blocks except i -th block in which j -th treatment is present.

Due to missing value in i -th block, the adjusted total of j' -th treatments ($j \neq j'$) which are present in i -th block are also affected. Let these adjusted total be $Q_{j'}$, where

$$\begin{aligned} Q_{j'} &= y_{.j'} - \frac{1}{k} B'_{j'} - \frac{1}{k} (y_{i..} + x); \quad j' = 1, 2, \dots, k-1 \\ &= y_{.j'} - \frac{1}{k} B_{j'} - \frac{x}{k}. \end{aligned}$$

Here $B'_{j'}$ = total of block totals except i -th block in which j' -th treatment is present, $\beta_{j'} =$ total of block totals of those blocks in which j' -th treatment is missing.

Therefore, we have

$$\sum Q_j^2 = \text{constant} + \left(Q_1 - \frac{x}{k}\right)^2 + \left(Q_2 - \frac{x}{k}\right)^2 + \dots + \left(Q_{k-1} - \frac{x}{k}\right)^2 + \left(Q_j + \frac{x(k-1)}{k}\right)^2.$$

Here constant is used to indicate terms free of x .

On simplification, we get

$$\sum Q_j^2 = \text{constant} + \frac{(k-1)x^2}{k^2} - \frac{2xQ'_j}{k} + \frac{x^2(k-1)^2}{k^2} + 2xQ_j.$$

Here $Q'_j =$ adjusted total of all j' -th treatment present in i -th block.

Therefore, the intra-block error sum of squares is given by

$$\phi = \text{constant} + x^2 - \frac{1}{k}(y_{i..} + x)^2 - \frac{(k-1)x^2}{rEk^2} + \frac{2xQ'_j}{rEk} - \frac{x^2(k-1)^2}{rEk^2} - \frac{2Q_jx}{rE}.$$

Now, $\frac{\partial \phi}{\partial x} = 0$ gives

$$x = \frac{rEy_{i..} - Q'_j + kQ_j}{(k-1)(rE-1)}.$$

Example 5.2 : Estimate the missing value of x of the following BIB design and analyse the data.

Block	Treatment	Result of treatments, (y_{ijl})	Block total $y_{i..}$	Block total with x
1	1 2 3	10 12 10	32	32
2	1 4 5	11 13 9	33	33
3	1 6 7	10 x 8	$18 + x$	24
4	2 4 6	11 12 10	33	33
5	2 5 7	12 10 9	31	31
6	3 4 7	11 13 10	34	34
7	3 5 6	12 11 10	33	33

Here the observation of 6th treatment in third block is missing. We have $y_{i..} = y_{3..} = 18$, $k = 3$, $r = 3$, $\lambda = 1$, $b = v = 7$, $rE = 7/3$.

The other treatments in the third block are 1 and 7.

Treatment	Total of treatment $y_{.j}$	Total of block totals in which j -th treatment is present, $\sum n_{ij}y_{i..}$	$Q_j = y_{.j} - \frac{1}{k} \sum n_{ij}y_{i..}$	Q_j with x
1	31	$83 + x$	$3.33 - \frac{x}{3}$	1.33
2	35	96	3	3
3	33	99	0	0
4	38	100	4.67	4.67
5	30	97	-2.33	-2.33
6	$20 + x$	$84 + x$	$-8 + x - \frac{x}{3}$	-4.00
7	27	$83 + x$	$-0.67 - \frac{x}{3}$	-2.67

The missing observation \hat{x} is

$$\begin{aligned} x &= \frac{rEy_{i..} - Q'_j + Q_j}{(k-1)(rE-1)}, \quad Q'_j = Q_1 + Q_7 = 3.33 + (-0.67) = 2.66, \quad Q_j = -8 \\ &= \frac{\frac{7}{3}18 - 2.66 - 3 \times 8}{(3-1)\left(\frac{7}{3}-1\right)} = 6. \end{aligned}$$

$$\text{C.T.} = \frac{G^2}{bk} = \frac{(220)^2}{7 \times 3} = 2304.762$$

$$SS \text{ (Total)} = \sum \sum \sum y_{ijl}^2 - \text{C.T.} = 2360 - 2304.762 = 55.238.$$

$$SS \text{ (Blocks)} = \frac{1}{k} \sum y_{i..}^2 - \text{C.T.} = \frac{6984}{3} - 2304.762 = 23.238.$$

$$SS \text{ (Treatment) adjusted} = \frac{1}{rE} \sum Q_j^2 = 26.201.$$

$$SS \text{ (Intra-block error)} = SS \text{ (Total)} - SS \text{ (Blocks)} - SS \text{ (Treatment) adjusted} \\ = 55.238 - 23.238 - 26.201 = 5.799.$$

ANOVA Table

Sources of Variation	d.f.	SS	MS = $\frac{SS}{\text{d.f.}}$	F	F _{0.05}
Blocks (unadjusted)	6	23.238	3.873	4.68	3.87
Treatment (adjusted)	6	26.201	4.367	5.27	3.87
Intra-block error	7	5.799	0.828	—	—
Total	19				

The treatments are significantly different.

5.4 Partially Balanced Incomplete Block (PBIB) Design

It has already been mentioned that BIB design is not available for any number of treatments. Even, if a BIB design is available for any number of treatments, its number of replications becomes large. To avoid the problem, Bose and Nair (1939) have evolved a group of designs which are known as PBIB design. However, the variances of the estimate of contrasts for the proposed designs are not same. The modified form of PBIB design has been proposed by Bose and Nair, Nair and Rao (1942) and later on Bose and Shimamoto (1952).

The definition of PBIB design depends on association scheme of treatments. Let there be v treatments for an experiment. Each treatment is associated with other treatments in different blocks differently. This association of treatments is the basis of association scheme of PBIB design.

Association Scheme of PBIB Design :

- (i) Any two treatments A and B is either first, second, third, or m th associates and if A is i -th associates of B , then B is also i -th associate of A .
- (ii) For any treatment A the number of treatments which are i -th associate with A is n_i . The value of n_i does not depend on the choice of A .
- (iii) If A and B are mutually i -th associate for any pair of treatments A and B , then the number of treatments which are simultaneously i -th associate of A and k th associate of B is P_{jk}^i , where P_{jk}^i does not depend on i -th associate of A and B . Here v, n_i, P_{jk}^i ($i, j = 1, 2, \dots, m$) are the parameters of m -class association scheme. The parameters are related as follows :

$$\sum_{i=1}^m n_i = v - 1, \quad \sum_{k=1}^m P_{jk}^i = n_j - \delta_{ij}, \quad n_i P_{jk}^i = n_j P_{jk}^i,$$

where $\delta_{ij} = 1$, if $i = j$ and $\delta_{ij} = 0$, otherwise.

Let A and B be two treatments which are mutually in i -th association scheme. In such a case k -th associate of A ($k = 1, 2, \dots, m$) must be all numbers of j -th associate of B ($j \neq i$). Therefore,

$$\sum_{k=1}^m P_{jk}^i = n_j.$$

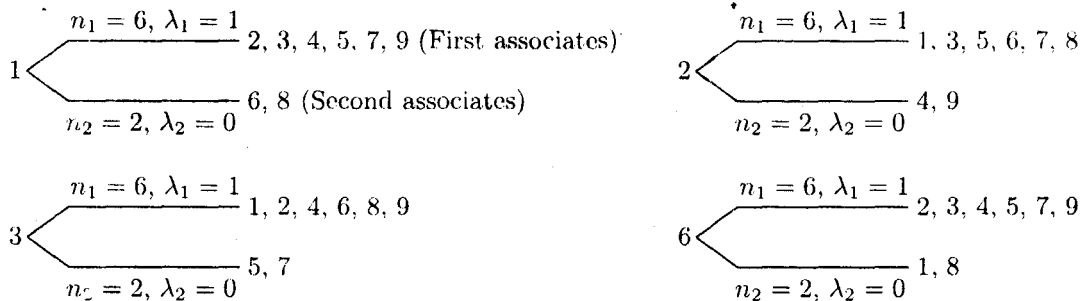
If $j = i$, A itself is j -th associate of B and for that reason all numbers of j -th associate of B become the numbers of k th associate of A . Therefore,

$$\sum_{k=1}^m P_{jk}^i = n_j - 1.$$

Under the association scheme discussed above, PBIB design is defined as an arrangement of v treatments in b blocks of k plots each ($k < v$) if the following conditions are satisfied :

- (i) In each block there are k number of ($k < v$) separate treatments.
- (ii) Each treatment appears in r blocks.
- (iii) If A and B are i -th associate in association scheme, then A and B appear together in λ_i blocks. Here λ_i does not depend on A and B . But they are i -th associate. Here λ_i 's are not equal and b, v, r, k, λ_i are the parameters of PBIB design.

Let us explain the association scheme of a PBIB design with parameters $b = v = 9, r = k = 3, n_1 = 6, n_2 = 2, \lambda_1 = 1, \lambda_2 = 0$. The treatments in different blocks are (1, 2, 3), (4, 5, 6), (7, 8, 9), (1, 4, 7), (2, 5, 8), (3, 6, 9), (1, 5, 9), (7, 2, 6) and (4, 8, 3). The association scheme of treatments 1, 2, 3 and 6 are as follows :



It is observed that treatments 1 and 2 are in first association scheme. The value $P_{11}^1 = 3$ for these two treatments. Here there are 3 treatments (3, 5 and 7) which are common in first association scheme of 2 and 1. Similarly,

$$P_{12}^1 = 2 = P_{21}^1, P_{22}^1 = 0.$$

Here $P_{12}^1 = 2$ indicates the number of treatments which are common in first association scheme of 1 and second association scheme of 2. Both 1 and 2 are first associate. Treatment 1 and 6 are second associate. The number of common treatments in first association scheme of 1 and first association scheme of 6 are 6 (2, 3, 4, 5, 7, 9). This number is $P_{11}^2 = 6$. Similarly, other common treatment numbers are :

$$P_{12}^1 = 0 = P_{21}^2, P_{22}^2 = 1.$$

Classification of PBIB Designs with Two Associate Classes

The design shown above is a PBIB design of two associate classes. The two associate classes PBIB designs can be classified into five classes. These are : (a) Group Divisible (GD) Scheme, (b) Triangular Scheme, (c) Latin Square Scheme, (d) Cyclic Scheme, (e) Simple PBIB Design.

The simple PBIB design is that one in which treatments are in two association scheme and for this scheme, either $\lambda_1 \neq 0, \lambda_2 = 0$ or, $\lambda_1 = 0, \lambda_2 \neq 0$.

5.5 Analysis of PBIB Design

The model for this design is

$$y_{hjl} = \mu + \alpha_h + \beta_j + e_{hjl}, \tag{1}$$

$h = 1, 2, \dots, b; j = 1, 2, \dots, v; l = n_{hj} = 0$ or 1 . Here y_{hjl} = the l -th result of j -th treatment in h -th block, μ = general mean, α_h = effect of h -th block, β_j = effect of j -th treatment and e_{hjl} = random error;

$$\begin{aligned} n_{hj} &= 0, \text{ if } j\text{-th treatment is absent in } h\text{-th block} \\ &= 1, \text{ if } j\text{-th treatment is present in } h\text{-th block.} \end{aligned}$$

$\sum_h n_{hj} = r, \sum_j n_{hj} = k, \sum_h n_{hj}n_{hj'} = \lambda_j$, if j -th and j' -th treatments are i -th associates ($i = 1, 2, \dots, m; m = 2$, if the design is simple PBIB design).

Intra-block Analysis for Simple PBIB Design

The normal equations to estimate the treatment effect β_j is

$$\begin{aligned} r(k-1)\hat{\beta}_j - \lambda_1 S_1(\hat{\beta}_j) - \lambda_2 S_2(\hat{\beta}_j) &= kQ_j \\ r(k-1)S_1(\hat{\beta}_j) - S_1(\hat{\beta}_j)(\lambda_1 P_{11}^1 + \lambda_2 P_{12}^2) - n_1 \lambda_1 \hat{\beta}_j - S_2(\hat{\beta}_j)(\lambda_1 P_{11}^2 + \lambda_2 P_{12}^2) &= kS_1(Q_j) \\ r(k-1)S_2(\hat{\beta}_j) - S_1(\hat{\beta}_j)(\lambda_1 P_{21}^2 + \lambda_2 P_{22}^1) - S_2(\hat{\beta}_j)(\lambda_1 P_{21}^1 + \lambda_2 P_{22}^2) - n_2 \lambda_2 \hat{\beta}_j &= kS_2(Q_j). \end{aligned}$$

Here, $Q_j = y_{.j} - \frac{1}{k} \sum_h n_{hj} y_{h.}$

$S_i(\hat{\beta}_j)$ = sum of n_i treatments which are i -th associates with j -th treatment

$S_i(Q_j)$ = sum of adjusted total of above mentioned n_i treatments.

Solving the first two equations, we get the value of $\hat{\beta}_j$ under the restriction

$$\hat{\beta}_j + S_1(\hat{\beta}_j) + S_2(\hat{\beta}_j) = 0,$$

where the estimate is

$$\hat{\beta}_j = \frac{k\{B_2 Q_j - A_2 S_1(Q_j)\}}{A_1 B_2 - A_2 B_1}.$$

Here $A_1 = r(k-1) + \lambda_2, A_2 = \lambda_2 - \lambda_1,$

$$B_1 = (\lambda_2 - \lambda_1)P_{12}^2, B_2 = r(k-1) + \lambda_2 + (\lambda_2 - \lambda_1)(P_{11}^1 - P_{11}^2).$$

The variance of the estimate of contrast of the type $\beta_j - \beta_{j'}$ is :

$$V_1 = V(\hat{\beta}_j - \hat{\beta}_{j'}) = \frac{2k(B_2 + A_2)\sigma^2}{A_1 B_2 - A_2 B_1}, \text{ if } j\text{-th and } j'\text{-th treatments are first associates}$$

$$V_2 = V(\hat{\beta}_j - \hat{\beta}_{j'}) = \frac{2k B_2 \sigma^2}{A_1 B_2 - A_2 B_1}, \text{ if } j\text{-th and } j'\text{-th treatments are second associates.}$$

Here σ^2 is the intra-block error variance. Its estimate is the intra-block error mean square, where

$$SS \text{ (Intra-block error)} = \sum \sum \sum y_{hjl}^2 - \text{C.T.} - SS \text{ (Blocks) unadjusted} - \sum \hat{\beta}_j Q_j.$$

The variances of the estimate of treatment contrast depend on the association scheme. These variances are averaged to compare with the variance of estimate of treatment contrast of randomized block design. The average variance (A.V.) is given by

$$\text{A.V. } (\hat{\beta}_j - \hat{\beta}_{j'}) = \frac{n_1 V_1 + n_2 V_2}{n_1 + n_2}.$$

Therefore, the efficiency factor of this design compared to randomized block design is :

$$\frac{(v - 1)(A_1 B_2 - A_2 B_1)}{rk\{(v - 1)B_2 + n_1 A_2\}}$$

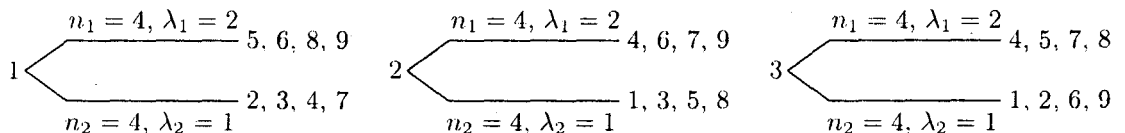
Example 5.3 : An experiment is conducted to study the loss of body weight of one type of slug kept under 9 different pesticides. The experiment is conducted through PBIB design having parameters $b = v = 9, r = k = 4, n_1 = n_2 = 4, \lambda_1 = 2, \lambda_2 = 1$, where the treatments are 9 different pesticides. The slugs are kept in different pesticides for seven days at room temperature. After 7 days the loss in body weights (in gm) are recorded. The plan of treatments and the loss of body weight under different pesticides are shown below :

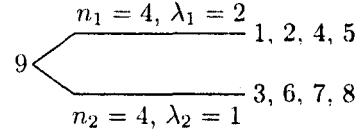
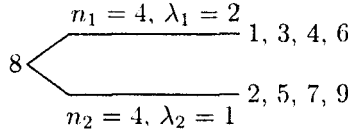
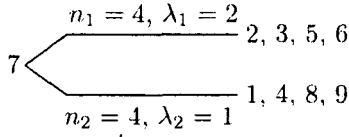
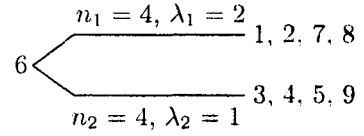
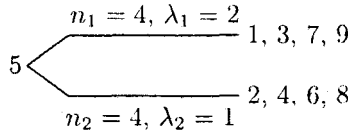
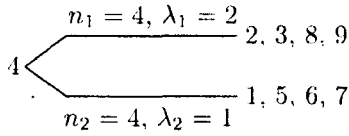
Block	(Treatments) and loss in body weight (in gms)	Block total $y_{h..}$
1	(1) 0.2 (3) 0.5 (5) 0.5 (8) 0.2	1.4
2	(2) 0.2 (3) 0.6 (4) 0.4 (7) 0.3	1.5
3	(3) 0.6 (6) 0.8 (7) 0.5 (8) 0.3	2.2
4	(1) 0.4 (2) 0.2 (6) 0.7 (9) 0.5	1.8
5	(1) 0.3 (5) 0.4 (6) 0.6 (7) 0.4	1.7
6	(3) 0.5 (4) 0.5 (5) 0.5 (9) 0.7	2.2
7	(2) 0.4 (4) 0.4 (6) 0.5 (8) 0.4	1.7
8	(1) 0.5 (4) 0.6 (8) 0.5 (9) 0.6	2.2
9	(2) 0.3 (5) 0.4 (7) 0.6 (9) 0.5	1.8

- (i) Analyse the data and group the pesticides, if possible.
- (ii) Find the combined intra- and inter-block estimate of effects of pesticides.
- (iii) Is there any difference between pesticide-8 and pesticide-9?

Compare these two treatments using combined intra- and inter-block information.

Solution : (i) The association scheme of the design is





$$P_{jk}^1 = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, P_{jk}^2 = \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix}, G = 16.5, \text{C.T.} = \frac{G^2}{bk} = \frac{(16.5)^2}{9 \times 4} = 7.5625.$$

$$SS (\text{Total}) = \sum \sum \sum y_{hjl}^2 - \text{C.T.} = 8.33 - 7.5625 = 0.7675.$$

Table to estimate Intra-block Effect of Treatment

Treatment	Total of treatment $y_{.j}$	Total of block totals in which j -th treatment is present, $\sum_h n_{hj}y_{h..}$	$Q_j = y_{.j} - \frac{1}{k} \sum_h n_{hj}y_{h..}$	$\hat{\beta}_j = \frac{k[B_2Q_j - A_2S_1(Q_j)]}{A_1B_2 - A_2B_1}$
1	1.4	7.1	-0.375	-0.103
2	1.1	6.8	-0.600	-0.163
3	2.2	7.3	0.375	0.107
4	1.9	7.6	0.000	-0.009
5	1.8	7.1	0.025	0.014
6	2.6	7.4	0.750	0.205
7	1.8	7.2	0.000	0.012
8	1.4	7.5	-0.475	-0.131
9	2.3	8.0	0.300	0.072

$$B_2 = r(k - 1) + \lambda_2 + (\lambda_2 - \lambda_1)(P_{11}^1 - P_{11}^2) = 14.$$

$$B_1 = (\lambda_2 - \lambda_1)P_{12}^2 = -2, A_1 = r(k - 1) + \lambda_2 = 13,$$

$$A_2 = \lambda_2 - \lambda_1 = -1, A_1B_2 - A_2B_1 = 180.$$

$$SS (\text{Treatment}) \text{ adjusted} = \sum \hat{\beta}_j Q_j = 0.4516.$$

$$SS (\text{Block}) \text{ unadjusted} = \frac{1}{k} \sum y_{i..}^2 - \text{C.T.} = \frac{30.99}{4} - 7.5625 = 0.185.$$

$$SS (\text{Intra-block error}) = SS (\text{Total}) - SS (\text{Treatment}) \text{ adjusted} - SS (\text{Block}) \text{ unadjusted} = 0.7675 - 0.4516 - 0.185 = 0.1309.$$

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$
Blocks (unadjusted)	8	0.185	0.02312	3.35	2.48
Treatment (adjusted)	8	0.4516	0.05645	8.19	2.48
Intra-block error	19	0.1309	0.00689	—	—
Total	35				

It is observed that the pesticides are significantly different in reducing the body weight of slugs.

Now, the variance of the estimate of contrast of the type $\beta_j - \beta_{j'}$ ($j \neq j' = 1, 2, \dots, v$), where j -th and j' -th treatments in first association scheme is :

$$v(\hat{\beta}_j - \hat{\beta}_{j'})_1 = \frac{2k(B_2 + A_2)\hat{\sigma}^2}{A_1B_2 - A_2B_1} = \frac{2 \times 4(14 - 1)0.00689}{180} = 0.00398, \text{ where } \hat{\sigma}^2 = 0.00689.$$

When j -th and j' -th treatments are in the second association scheme,

$$v(\hat{\beta}_j - \hat{\beta}_{j'})_2 = \frac{2kB_2\hat{\sigma}^2}{A_1B_2 - A_2B_1} = \frac{2 \times 4 \times 14 \times 0.00689}{180} = 0.00429,$$

Therefore, to compare the effects of two pesticides, which are in first association scheme, the LSD is given by :

$$LSD = t_{0.05,19} \sqrt{v(\hat{\beta}_j - \hat{\beta}_{j'})_1} = 2.093 \sqrt{0.00398} = 0.132.$$

The same LSD when two treatments are in second association scheme is :

$$LSD = t_{0.05,19} \sqrt{v(\hat{\beta}_j - \hat{\beta}_{j'})_2} = 2.093 \sqrt{0.00429} = 0.137.$$

The treatment effects in ascending order are $\hat{\beta}_2 = -0.163$, $\hat{\beta}_8 = -0.131$, $\hat{\beta}_1 = -0.103$, $\hat{\beta}_4 = -0.009$, $\hat{\beta}_7 = 0.012$, $\hat{\beta}_5 = 0.014$, $\hat{\beta}_9 = 0.072$, $\hat{\beta}_3 = 0.107$, $\hat{\beta}_6 = 0.201$.

$\beta_2, \beta_8, \beta_1, \beta_4, \beta_7, \beta_5, \beta_9, \beta_3, \beta_5$

The underlined effects do not differ significantly.

We have

$$W = \frac{1}{MS \text{ (Intra-block error)}} = 145.13788.$$

$$W_1 = \frac{v(r-1)}{k(b-1)s_\alpha - (v-k)s}, \text{ where } s = MS \text{ (Intra-block error),}$$

$$s_\alpha = MS \text{ (Block) adjusted.}$$

$$SS \text{ (Block) adjusted} = SS \text{ (Block) unadjusted} - SS \text{ (Treatment) unadjusted} \\ + SS \text{ (Treatment) adjusted.}$$

$$SS \text{ (Treatment) unadjusted} = \frac{1}{r} \sum y_{.j}^2 - \text{C.T.} = \frac{32.11}{4} - 7.5625 = 0.465.$$

$$SS \text{ (Block) adjusted} = 0.185 - 0.465 + 0.4516 = 0.1716.$$

$$\therefore s_\alpha = \frac{1}{(k-1)} SS \text{ (Block) adjusted} = \frac{0.1716}{8} = 0.02145.$$

$$\therefore W_1 = \frac{9(4-1)}{4(9-1)0.02145 - (9-4)0.00689} = 41.41422.$$

Now, for combined intra- and inter-block estimate of treatment effects, we have

$$A'_1 = r\{W(k-1) + W_1\} + (W - W_1)\lambda_2 = 2011.0351$$

$$A'_2 = (\lambda_2 - \lambda_1) + (W - W_1) = -103.72358$$

$$B'_1 = (W - W_1)(\lambda_2 - \lambda_1)P_{12}^2 = -207.44732$$

$$B'_2 = r\{W(k-1) + W_1\} + (W - W_1)\{\lambda_2 + (\lambda_2 - \lambda_1)(P_{11}^1 - P_{11}^2)\} = 2114.75876.$$

$$A'_1 B'_2 - A'_2 B'_1 = 4231336.916.$$

Table to Estimate Combined Intra- and Inter-block Treatment Effect.

Treatment No.	$P_j = \sum n_{hj}y_{h.} - \frac{r}{b}G$	$R_j = WkQ_j + W_1P_j$	$S_1(R_j)$	$\hat{\beta}_j = \frac{[B'_2 R_j - A'_2 S_1(R_j)]}{A'_1 B'_2 - A'_2 B'_1}$
1	-0.2333	-227.36876	375.94592	-0.1044
2	-0.5333	-370.41711	645.47694	-0.1693
3	-0.0333	216.32773	-258.48172	0.1018
4	0.2667	11.04517	-221.17128	0.0001
5	-0.2333	4.85185	185.21477	0.0070
6	0.0667	438.17597	-872.16461	0.1976
7	-0.1333	-5.52052	288.93844	0.0043
8	0.1667	-268.85822	438.18011	-0.1236
9	0.6667	201.77632	-581.88885	0.0866

Using the above analysis the homogeneity of treatment effects cannot be tested by F -test statistic since W and W_1 are estimated values. However, a quantity related to treatment sum of squares is $\frac{1}{k} \sum \hat{\beta}_j R_j = 61.44$. This quantity is approximately distributed as χ^2 with $(v-1) = 8$ d.f. Since 61.44 is greater than $\chi_{0.05,8}^2$, the treatment effects are significantly different.

(iii) The pesticide-8 and pesticide-9 are in second association scheme. The variance of the treatment contrast of these two treatments is estimated by

$$v(\hat{\beta}_j - \hat{\beta}'_j) = \frac{2kB'_2}{A'_1 B'_2 - A'_2 B'_1} = 0.003998, \quad s \cdot e(\hat{\beta}_j - \hat{\beta}'_j) = 0.06323.$$

$$\text{Therefore, } t = \frac{\hat{\beta}_8 - \hat{\beta}_9}{s \cdot e(\hat{\beta}_8 - \hat{\beta}_9)} = \frac{-0.1236 - 0.0866}{0.06323} = -3.32.$$

Here $|t| > t_{0.025,19} = 2.093$. Pesticide-8 and pesticide-9 are significantly different.

If j -th and j' -th treatments are in first association scheme, then

$$v(\hat{\beta}_j - \hat{\beta}_{j'}) = \frac{2k(B'_2 + A'_2)}{A'_1 B'_2 - A'_2 B'_1} = 0.0038.$$

Chapter 6

Covariance Analysis

6.1 Definition

The local control is used to classify the experimental materials or experimental units into homogeneous groups so that the treatment is unaffected by uncontrolled source of variation when it is qualitative in character. For example, let us consider the experiment of testing the homogeneity of several doses of nitrogen when it is used to produce marigold. To study the effects of doses of nitrogen different doses of nitrogen are to be applied in agricultural plots of homogeneous fertility. If the plots used for different levels of nitrogen are not of same type in respect of fertility, the production will be affected by fertility differentials and treatment effect will be entangled with fertility effect. This fertility differential is a source of experimental error which can be reduced by local control. The production of marigold will also be affected if number of plants under different levels of nitrogen are not same. The variation in number of plants is also an external source of variation and it cannot be controlled by local control. This external source of variation is quantitative in character. In such a situation the effect of nitrogen is to be estimated after eliminating the impact of number of plants on production. Covariance analysis is a technique to estimate the treatment effect after eliminating the effect of quantitative external source of variation. The quantitative external source of variation is known as concomitant variable and it is used to control the experimental error.

The concomitant variable is denoted by x and the experimental result is denoted by y , where y is assumed to be linearly dependent on x . Therefore, to formulate the mathematical model for the observations of y a regression coefficient of y on x is introduced along with other parameters to measure the impacts of qualitative variables. As a result the analysis of data is performed using analysis of variance technique and technique of regression analysis. Since analysis of variance and regression analysis is performed simultaneously the technique is called covariance analysis.

The main objective of covariance analysis is to estimate the treatment effect eliminating the effect of concomitant variable. The effect of the concomitant variable is the regression coefficient of y variable on concomitant variable x . In practice, there may be more than one concomitant variable and hence, in the usual analysis of variance model more than one regression parameters are to be introduced.

Assumptions in covariance analysis : (i) The concomitant variables x 's are non-random variable and the values of these variables are observed without error.

(ii) The regression of y on x is linear after eliminating the effects of block and treatment. The regression effect is independent of block effect and treatment effect.

(iii) The experimental error is assumed to be normally distributed with mean zero and common variance σ^2 .

Application of analysis of covariance : The covariance analysis is used in the following aspects.

- (i) The covariance analysis is used to control the quantitative external source of variation and hence, to increase the efficiency of the experiment.
- (ii) It is used to estimate the treatment effect eliminating the effect of concomitant variable.
- (iii) It is helpful in estimating the effect of missing value. The technique of covariance analysis is used to analyse the data having missing observations.
- (iv) It helps to explain the model for treatment effect.

6.2 Covariance Analysis in Case of Completely Randomized Design (CRD) with One Concomitant Variable

The model for CRD with one concomitant variable is

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + e_{ij}, \tag{A}$$

$i = 1, 2, \dots, p; j = 1, 2, \dots, q.$

Here y_{ij} = observation of j -th replication of i -th treatment, μ = general mean, α_i = effect of i -th treatment, β = the regression coefficient of y on x , e_{ij} = random error.

The normal equations to estimate the parameters in the model (A) are :

$$\begin{aligned} y_{..} &= pq\hat{\mu} + q \sum \hat{\alpha}_i \\ y_{i.} &= q\hat{\mu} + q\hat{\alpha}_i + q\hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \\ \sum \sum y_{ij}(x_{ij} - \bar{x}_{..}) &= q \sum \hat{\alpha}_i(\bar{x}_{i.} - \bar{x}_{..}) + \hat{\beta} \sum \sum (x_{ij} - \bar{x}_{..})^2. \end{aligned}$$

Let $G_{yy} = \sum \sum (y_{ij} - \bar{y}_{..})^2$ $T_{yy} = q\sum (\bar{y}_{i.} - \bar{y}_{..})^2$
 $G_{xx} = \sum \sum (x_{ij} - \bar{x}_{..})^2$ $T_{xx} = q \sum (\bar{x}_{i.} - \bar{x}_{..})^2$
 $G_{xy} = \sum \sum (x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..})$ $T_{xy} = q \sum (\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..})$
 $E_{yy} = G_{yy} - T_{yy}, E_{xx} = G_{xx} - T_{xx}, E_{xy} = G_{xy} - T_{xy}$

Then under the restriction $\sum \hat{\alpha}_i = 0$, we have

$$\hat{\mu} = \bar{y}_{..}, \hat{\alpha}_i = (\bar{y}_{i.} - \bar{y}_{..}) - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \text{ and } \hat{\beta} = \frac{E_{xy}}{E_{xx}}.$$

The sum of squares due to estimates is :

$$\begin{aligned} SS(\text{estimates}) &= \hat{\mu}y_{i.} + \sum \hat{\alpha}_i y_{i.} + \sum \sum \hat{\beta} y_{ij}(x_{ij} - \bar{x}_{..}) \\ &= pq\bar{y}_{..}^2 + q \sum (\bar{y}_{i.} - \bar{y}_{..})\{(\bar{y}_{i.} - \bar{y}_{..}) - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..})\} + \hat{\beta}G_{xy} \\ &= pq\bar{y}_{..}^2 + T_{yy} + \hat{\beta}(G_{xy} - T_{xy}) \\ &= pq\bar{y}_{..}^2 + T_{yy} + \hat{\beta}E_{xy}. \end{aligned}$$

The d.f. of this $SS(\text{estimates})$ is $(p + 1)$. The sum of squares of error is :

$$SS(\text{error}) = \sum \sum y_{ij}^2 - pq\bar{y}_{..}^2 - T_{yy} - \hat{\beta}E_{xy} = G_{yy} - T_{yy} - \hat{\beta}E_{xy} = E_{yy} - \hat{\beta}E_{xy}. \tag{a}$$

The d.f. of this sum of squares is $(pq - p - 1)$.

The main objective of this analysis is to test the significance of the hypothesis :

$$H_0 : \alpha_i = 0, \text{ against } H_A : \alpha_i \neq 0.$$

Under the null hypothesis the model becomes

$$y_{ij} = \mu + \beta(x_{ij} - \bar{x}_{..}) + e_{ij}, \text{ where } \hat{\beta} = \frac{G_{xy}}{G_{xx}}.$$

The sum of squares of estimates is :

$$SS(\text{estimate}) = \hat{\mu}y_{..} + \hat{\beta} \sum \sum y_{ij}(x_{ij} - \bar{x}_{..}) = pq\bar{y}^2 + \hat{\beta}G_{xy}.$$

The d.f. of the above sum of squares is 2. The $SS(\text{error})$ under H_0 is :

$$SS(\text{error}) = G_{yy} - \hat{\beta}G_{xy}. \quad (b)$$

The d.f. of this error sum of squares is $pq - 2$. The sum of squares due to treatment under H_0 is :

$$SS(\hat{\alpha}_i) = (b) - (a) = G_{yy} - \hat{\beta}G_{xy} - E_{yy} + \hat{\beta}E_{xy} = T_{yy} - \hat{\beta}G_{xy} + \hat{\beta}E_{xy}.$$

This sum of squares has $(p - 1)$ d.f.

In practice, the covariance analysis is performed, if $\beta \neq 0$. Therefore, at the first step of analysis we need to test the significance of

$$H_0 : \beta = 0 \text{ against } H_A : \beta \neq 0.$$

The model under the above null hypothesis is

$$y_{ij} = \mu + \alpha_i + e_{ij},$$

where $\hat{\mu} = \bar{y}_{..}$, $\hat{\alpha}_i = \bar{y}_{.i} - \bar{y}_{..}$. The sum of squares due to error of the above model is

$$SS(\text{error}) = E_{yy}. \quad (c)$$

This sum of squares has $p(q - 1)$ d.f. Now, subtracting (a) from (c), we have

$$SS(\hat{\beta}) = \hat{\beta}E_{xy}.$$

This $SS(\hat{\beta})$ has 1 d.f. Hence, the test statistic for $H_0 : \beta = 0$ is

$$F = \frac{\hat{\beta}E_{xy}/1}{(E_{yy} - \hat{\beta}E_{xy})/(pq - p - 1)}.$$

This F has 1 and $(pq - p - 1)$ d.f. If $H_0 : \beta = 0$ is rejected, then covariance analysis is needed, otherwise usual analysis of variance is sufficient to test the significance of $H_0 : \alpha_i = 0$. Under covariance analysis, the test statistic for $H_0 : \alpha_i = 0$ is

$$F = \frac{(T_{yy} - \hat{\beta}G_{xy} + \hat{\beta}E_{xy})/(p - 1)}{(E_{yy} - \hat{\beta}E_{xy})/(pq - p - 1)}.$$

This F has $(p - 1)$ and $(pq - p - 1)$ d.f.

ANCOVA Table

Sources of variation	$SS(x)$	$SS(y)$	$Sp(xy)$	Regression coefficient	Adjusted		$SS(\hat{\alpha}_i)$ under H_0
					SS	d.f.	
Treatment	T_{xx}	T_{yy}	T_{xy}			$p - 1$	
Error	E_{xx}	E_{yy}	E_{xy}	$\hat{\beta} = \frac{E_{xy}}{E_x}$	$E_{yy} - \hat{\beta}E_{xy}$	$pq - p - 1$	$T_{yy} - \hat{\beta}G_{xy} + \hat{\beta}E_{xy}$
Total	G_{xx}	G_{yy}	G_{xy}				

The adjusted treatment means are

$$\bar{y}_i \text{ (adjusted)} = \bar{y}_i - \hat{\beta}(\bar{x}_i - \bar{x}_{..})$$

$$V[\bar{y}_i \text{ (adjusted)}] = \sigma^2 \left[\frac{1}{q} + \frac{(\bar{x}_i - \bar{x}_{..})^2}{E_{xx}} \right].$$

Also, we have $V(\bar{y}_i - \bar{y}_{i'}) \text{ adjusted} = \sigma^2 \left[\frac{2}{q} + \frac{(\bar{x}_i - \bar{x}_{i'})^2}{E_{xx}} \right]$.

Hence, the test statistic to test the significance of $H_0 : \alpha_i = \alpha_{i'} \ (i \neq i' = 1, 2, \dots, p)$ is

$$F = \frac{(\bar{y}_i - \bar{y}_{i'})^2 \text{ adjusted}}{\hat{\sigma}^2 \left[\frac{2}{q} + \frac{(\bar{x}_i - \bar{x}_{i'})^2}{E_{xx}} \right]}.$$

Here $\hat{\sigma}^2 = \frac{1}{pq - p - 1} [E_{yy} - \hat{\beta}E_{xy}]$. This F has 1 and $(pq - p - 1)$ d.f.

It is observed that $V(\bar{y}_i - \bar{y}_{i'})$ depends on $(\bar{x}_i - \bar{x}_{i'})^2$. The average value of $V(\bar{y}_i - \bar{y}_{i'})$ can be written as

$$\frac{2MS(\text{error})}{q} \left[1 + \frac{T_{xx}/(p-1)}{E_{xx}} \right].$$

This variance can be used for multiple comparison of treatment means.

The above covariance analysis is based on the assumption that the impact of concomitant variable is same for all treatments. In practice, the different treatments may be responded differently by the concomitant variable. To tackle this situation differential regression coefficient is introduced in the model. Thus, the model becomes

$$y_{ij} = \mu + \alpha_i + \beta_i(x_{ij} - \bar{x}_i) + e_{ij},$$

$i = 1, 2, \dots, p; j = 1, 2, \dots, q$.

The estimate of β_i is

$$\hat{\beta}_i = \frac{E_{xy_i}}{E_{xx_i}}, \text{ where}$$

$$E_{xy_i} = \sum_j (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i), \quad E_{xx_i} = \sum_j (x_{ij} - \bar{x}_i)^2 \text{ such that}$$

$$E_{xy} = \sum_{i=1}^p E_{xy_i}, \quad E_{xx} = \sum_{i=1}^p E_{xx_i}, \quad E_{yy} = \sum_{i=1}^p E_{yy_i} \quad \text{and} \quad E_{yy_i} = \sum_j (y_{ij} - \bar{y}_i)^2.$$

The sum of squares due to estimates for the model is :

$$SS(\text{estimates}) = \hat{\mu}y_{..} + \sum \hat{\alpha}_i y_{i.} + \sum \sum \hat{\beta}_i y_{ij}(x_{ij} - \bar{x}_i)$$

$$= pq\bar{y}_{..}^2 + q \sum (\bar{y}_i - \bar{y}_{..})^2 + \sum_{i=1}^p \hat{\beta}_i E_{xy_i}.$$

The d.f. of this sum of squares is $2p$. The error sum of squares is :

$$S_1 = SS(\text{error}) = \sum \sum y_{ij}^2 - pq\bar{y}_{..}^2 - q \sum (\bar{y}_i - \bar{y}_{..})^2 - \sum_{i=1}^p \hat{\beta}_i E_{xy_i} = E_{yy} - \sum \hat{\beta}_i E_{xy_i}.$$

Here S_1 has $p(q-1) - p$ d.f. Now the sum of squares due to the use of β_i instead of β is given by

$$S_2 = E_{yy} - \hat{\beta}E_{xy} - S_1 = \sum_{i=1}^p \hat{\beta}_i E_{xy_i} - \hat{\beta}E_{xy}.$$

The d.f. of S_2 is $(p-1)$. Now, to test the significance of the hypothesis :

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = \beta,$$

the test statistic is

$$F = \frac{S_2/(p-1)}{S_1/[p(q-1) - p]}.$$

This F is distributed as variance ratio with $(p-1)$ and $[p(q-1) - p]$ d.f.

For the covariance analysis the concomitant variable x is assumed to be non-random variable. If x is also a random variable, the regression of y on x is :

$$y_{ij} = \mu_y + \beta(x - \mu_x)$$

such that $\beta = \frac{\sigma_{xy}}{\sigma_x^2}$. Here $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$ and σ_{xy} are means, variances and covariance of the variables X and Y . Under assumption of randomness of x ,

$$E[MS(\hat{\alpha}_i)] = \sigma^2 + q\sigma_\alpha^2 \quad \text{and} \quad E[MS(\text{error})] = \sigma^2.$$

Therefore, to test the significance of $H_0 : \sigma_\alpha^2 = 0$, the test statistic is :

$$F = \frac{(T_{yy} - \hat{\beta}G_{xy} + \hat{\beta}E_{xy})/(p-1)}{(E_{yy} - \hat{\beta}E_{xy})/[p(q-1) - 1]}.$$

Example 6.1 : In an agricultural research station an experiment is conducted to study the productivity of 4 varieties of potato using nitrogen fertilizer. The agricultural plots for cultivation are found homogeneous in respect of fertility. The potato varieties are randomly allocated to different plots. But the amount of fertilizer used (x kg/plot) in different plots are not same. The production of potato (y kg) in different plots along with amount of fertilizer used are given below :

Plots	Potato varieties							
	1		2		3		4	
	y	x	y	x	y	x	y	x
1	45.2	1.2	55.0	1.5	30.5	1.0	40.0	1.5
2	46.4	1.0	54.0	1.3	35.2	1.5	44.2	1.2
3	44.0	1.0	50.0	1.2	32.4	1.3	40.0	1.3
4	50.0	1.5	50.0	1.0	38.0	2.0	42.2	1.2
5	48.5	1.2	54.2	1.4	40.2	2.0	41.2	1.0
Total y_i .	234.1		263.2		176.3		207.6	
x_i .	5.9		6.4		7.8		6.2	

- (i) Analyse the data and group the varieties of potato which are similar in productivity.
- (ii) Do you think that the impacts of concomitant variable are homogeneous for all varieties of potato.

Solution : (i) We have $p = 4$, $q = 5$, $G_y = 881.2$, $G_x = 26.3$.

$$C.T_y = \frac{G_y^2}{pq} = 38825.672, \quad C.T_x = \frac{G_x^2}{pq} = 34.5845.$$

$$G_{yy} = \sum \sum y_{ij}^2 - C.T_y = 39773.9 - 38825.672 = 948.228.$$

$$G_{xx} = \sum \sum x_{ij}^2 - C.T_x = 36.23 - 34.5845 = 1.6455.$$

$$C.T_{xy} = \frac{G_x G_y}{20} = 1158.778.$$

$$G_{xy} = \sum \sum x_{ij} y_{ij} - C.T_{xy} = 1155.12 - 1158.778 = -3.658.$$

$$T_{yy} = \frac{1}{q} \sum y_i^2 - C.T_y = \frac{198256.5}{5} - 38825.672 = 825.628.$$

$$T_{xx} = \frac{1}{q} \sum x_i^2 - C.T_x = \frac{175.05}{5} - 34.5845 = 0.4255.$$

$$T_{xy} = \frac{1}{q} \sum x_i y_i - C.T_{xy} = \frac{5727.93}{5} - 1158.778 = -13.192.$$

$$E_{yy} = G_{yy} - T_{yy} = 948.228 - 825.628 = 122.60.$$

$$E_{xx} = G_{xx} - T_{xx} = 1.6455 - 0.4255 = 1.22.$$

$$E_{xy} = G_{xy} - T_{xy} = -3.658 + 13.192 = 9.534.$$

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}} = \frac{9.534}{1.22} = 7.81, \quad \hat{\beta} = \frac{G_{xy}}{G_{xx}} = \frac{-3.658}{1.6455} = -2.223.$$

$$SS(\text{Potatoes}) = T_{yy} - \hat{\beta} G_{xy} + \hat{\beta} E_{xy} = 825.628 - 2.223 \times 3.658 + 7.81 \times 9.534 = 891.957.$$

$$SS(\text{error}) = E_{yy} - \hat{\beta} E_{xy} = 122.60 - 7.81 \times 9.534 = 48.139.$$

We need to test the significance of $H_0 : \beta = 0$. The test statistic is :

$$F = \frac{\hat{\beta} E_{xy}}{SS(\text{error})/15} = \frac{7.81 \times 9.534}{48.139/15} = 23.20.$$

Since $F = 23.20 > F_{0.05;1,15} = 4.54$, H_0 is rejected. The production of potato varies significantly with the variation in the amount of fertilizer. Hence, we need to estimate the effects of varieties of potato eliminating the effect of fertilizer.

The test statistic to test the significance of variety effect ($H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$) the test statistic is

$$F = \frac{SS(\text{Potatoes})/3}{MS(\text{error})} = \frac{891.957/3}{3.2093} = 92.64.$$

Since $F = 92.64 > F_{0.05;3,15} = 2.29$, H_0 is rejected. The varieties of potato are significantly different.

ANCOVA Table

Source of variation	$SS(x)$	$SS(y)$	$SP(xy)$	Regression coefficient	Adjusted SS	$SS(\hat{\alpha}_i)$ under H_0
Potatoes	0.4255	825.628	-13.192			891.953
Error	1.22	122.60	9.534	7.81	48.139	
Total						

The adjusted treatment means are :

$$\bar{y}_i \text{ (adjusted)} = \bar{y}_i - \hat{\beta}(\bar{x}_i - \bar{x}..)$$

$$\bar{y}_1 \text{ (adj)} = \bar{y}_1 - \hat{\beta}(\bar{x}_1 - \bar{x}..) = 46.82 - 7.81(1.18 - 1.315) = 47.87.$$

$$\bar{y}_2 \text{ (adj)} = \bar{y}_2 - \hat{\beta}(\bar{x}_2 - \bar{x}..) = 52.64 - 7.81(1.28 - 1.315) = 52.91$$

$$\bar{y}_3 \text{ (adj)} = \bar{y}_3 - \hat{\beta}(\bar{x}_3 - \bar{x}..) = 35.26 - 7.81(1.56 - 1.315) = 33.35$$

$$\bar{y}_4 \text{ (adj)} = \bar{y}_4 - \hat{\beta}(\bar{x}_4 - \bar{x}..) = 41.52 - 7.81(1.24 - 1.315) = 42.11.$$

To group these adjusted means we can perform Duncan's multiple range test, where the test statistic is :

$$D_k = d_{0.05, k, f} \sqrt{\frac{MS(\text{error})}{q} \left[1 + \frac{T_{xx}/p - 1}{E_{xx}} \right]}; \quad k = 2, 3, 4; \quad f = 15.$$

$$D_2 = 3.01 \times 0.845 = 2.55, \quad D_3 = 3.16 \times 0.846 = 2.67, \quad D_4 = 3.25 \times 0.846 = 2.75.$$

$$\bar{y}_2 - \bar{y}_3 = 19.56 > D_4, \text{ all means are significantly different.}$$

$$\bar{y}_2 - \bar{y}_4 = 10.80 > D_3, \text{ variety-2 and variety-4 are different.}$$

$$\bar{y}_1 - \bar{y}_3 = 14.52 > D_3, \text{ variety-1 and variety-3 are different.}$$

$$\bar{y}_4 - \bar{y}_3 = 8.76 > D_2, \text{ variety-3 and variety-4 are different.}$$

$$\bar{y}_1 - \bar{y}_4 = 5.76 > D_2, \text{ variety-1 and variety-4 are different.}$$

$$\bar{y}_2 - \bar{y}_1 = 5.04 > D_2, \text{ variety-1 and variety-2 are different.}$$

All varieties are significantly different.

(ii) We need to test the significance of the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$.

We know $\hat{\beta}_i = \frac{E_{xy_i}}{E_{xx_i}}$, where $E_{xx_i} = \sum_j x_{ij}^2 - \frac{x_i^2}{q}$

$$E_{xy_i} = \sum_j x_{ij}y_{ij} - \frac{x_i \cdot y_i}{q}, \quad i = 1, 2, 3, 4$$

$$E_{xy_1} = \sum x_{1j}y_{1j} - \frac{x_1 \cdot y_1}{q} = 277.84 - \frac{5.9 \times 234.1}{5} = 1.602.$$

$$E_{xy_2} = \sum x_{2j}y_{2j} - \frac{x_2 \cdot y_2}{q} = 338.58 - \frac{6.4 \times 263.2}{5} = 1.684.$$

$$E_{xy_3} = \sum x_{3j}y_{3j} - \frac{x_3 \cdot y_3}{q} = 281.82 - \frac{7.8 \times 176.3}{5} = 6.792.$$

$$E_{xy_4} = \sum x_{4j}y_{4j} - \frac{x_4 \cdot y_4}{q} = 256.88 - \frac{6.2 \times 207.6}{5} = -0.544.$$

$$E_{xx_1} = \sum x_{1j}^2 - \frac{x_1^2}{q} = 7.13 - \frac{(5.9)^2}{5} = 0.168, \quad \hat{\beta}_1 = \frac{E_{xy_1}}{E_{xx_1}} = 9.536.$$

$$E_{xx_2} = \sum x_{2j}^2 - \frac{x_2^2}{q} = 8.34 - \frac{(6.4)^2}{5} = 0.148, \quad \hat{\beta}_2 = \frac{E_{xy_2}}{E_{xx_2}} = 11.378.$$

$$E_{xx_3} = \sum x_{3j}^2 - \frac{x_3^2}{q} = 12.94 - \frac{(7.8)^2}{5} = 0.772, \quad \hat{\beta}_3 = \frac{E_{xy_3}}{E_{xx_3}} = 8.798.$$

$$E_{xx_4} = \sum x_{4j}^2 - \frac{x_4^2}{q} = 7.82 - \frac{(6.2)^2}{5} = 0.132, \hat{\beta}_4 = \frac{E_{xy_4}}{E_{xx_4}} = -4.121.$$

$$S_1 = E_{yy} - \sum_{i=1}^p \hat{\beta}_i E_{xy_i} = 122.60 - 116.628 = 5.972.$$

$$S_2 = E_{yy} - \hat{\beta} E_{xy} - S_1 = 122.60 - 7.81 \times 9.534 - 5.972 = 42.167.$$

$$F = \frac{S_2/p - 1}{S_1/[p(q-1) - p]} = \frac{42.167/3}{5.972/12} = 28.24.$$

Since $F > F_{0.05;3,12} = 3.49$, H_0 of homogeneous regression coefficients is rejected. Thus, different varieties of potato are influenced differently by fertilizer.

6.3 Covariance Analysis in Completely Randomized Design with Two Concomitant Variables

The model for this analysis is

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + \gamma(z_{ij} - \bar{z}_{..}) + e_{ij}, \quad (A)$$

$i = 1, 2, \dots, p; j = 1, 2, \dots, q.$

Here y_{ij} = result of i -th treatment in j -th replication, μ = general mean, α_i = effect of i -th treatment, x_{ij} and z_{ij} are the values of concomitant variable corresponding to y_{ij} , β = regression coefficient of y on x , γ = regression coefficient of y on z , e_{ij} = random error.

The normal equations to estimate the parameters in the model (A) are :

$$y_{..} = pq\hat{\mu} + q \sum \hat{\alpha}_i$$

$$y_{i.} = q\hat{\mu} + q\hat{\alpha}_i + q\hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) + q\hat{\gamma}(\bar{z}_{i.} - \bar{z}_{..})$$

$$\sum \sum y_{ij}(x_{ij} - \bar{x}_{..}) = q \sum \hat{\alpha}_i(\bar{x}_{i.} - \bar{x}_{..}) + \hat{\beta} \sum \sum (x_{ij} - \bar{x}_{..})^2 + \hat{\gamma} \sum \sum (x_{ij} - \bar{x}_{..})(\bar{z}_{i.} - \bar{z}_{..})$$

$$\sum \sum y_{ij}(z_{ij} - \bar{z}_{..}) = q \sum \hat{\alpha}_i(\bar{z}_{i.} - \bar{z}_{..}) + \hat{\beta} \sum \sum (x_{ij} - \bar{x}_{..})(z_{ij} - \bar{z}_{..}) + \hat{\gamma} \sum \sum (z_{ij} - \bar{z}_{..})^2.$$

$$\text{Let } G_{yy} = \sum \sum (y_{ij} - \bar{y}_{..})^2, G_{xx} = \sum \sum (x_{ij} - \bar{x}_{..})^2, G_{zy} = \sum \sum (x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..})$$

$$G_{xz} = \sum \sum (x_{ij} - \bar{x}_{..})(z_{ij} - \bar{z}_{..}), T_{yy} = q \sum (\bar{y}_{i.} - \bar{y}_{..})^2, T_{xx} = q \sum (\bar{x}_{i.} - \bar{x}_{..})^2$$

$$T_{zy} = q \sum (\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..}), T_{xz} = q \sum (\bar{x}_{i.} - \bar{x}_{..})(\bar{z}_{i.} - \bar{z}_{..})$$

$$G_{yz} = \sum \sum (y_{ij} - \bar{y}_{..})(z_{ij} - \bar{z}_{..}), T_{yz} = q \sum (\bar{y}_{i.} - \bar{y}_{..})(\bar{z}_{i.} - \bar{z}_{..})$$

$$E_{yy} = G_{yy} - T_{yy}, E_{xx} = G_{xx} - T_{xx}, E_{zy} = G_{zy} - T_{zy}, E_{yz} = G_{yz} - T_{yz}$$

$$E_{zz} = G_{zz} - T_{zz}.$$

Under the restriction $\sum \hat{\alpha}_i = 0$, we have

$$\hat{\mu} = \bar{y}_{..}, \hat{\alpha}_i = (\bar{y}_{i.} - \bar{y}_{..}) - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) - \hat{\gamma}(\bar{z}_{i.} - \bar{z}_{..})$$

$$E_{xy} = \hat{\beta}E_{xx} + \hat{\gamma}E_{xz}$$

$$E_{yz} = \hat{\beta}E_{zx} + \hat{\gamma}E_{zz}.$$

Solving equations shown above, we get:

$$\hat{\beta} = \frac{E_{zz}E_{xy} - E_{xz}E_{yz}}{E_{xx}E_{zz} - E_{xz}^2}, \hat{\gamma} = \frac{E_{xx}E_{yz} - E_{xz}E_{xy}}{E_{xx}E_{zz} - E_{xz}^2}.$$

$$\begin{aligned}
S_1 = SS(\text{estimates}) &= \hat{\mu}y_{..} + \sum \hat{\alpha}_i y_{i.} + \hat{\beta} \sum \sum y_{ij} (x_{ij} - \bar{x}_{..}) + \hat{\gamma} \sum \sum y_{ij} (z_{ij} - \bar{z}_{..}) \\
&= pq\bar{y}^2 + T_{yy} - \hat{\beta}T_{xy} - \hat{\gamma}T_{yz} + \hat{\beta}G_{xy} + \hat{\gamma}G_{yz} \\
&= pq\bar{y}^2 + T_{yy} + \hat{\beta}E_{xy} + \hat{\gamma}E_{yz}.
\end{aligned}$$

The d.f. of S_1 is $(p+2)$. The sum of squares due to error is :

$$S_2 = SS(\text{error}) = \sum \sum y_{ij}^2 - pq\bar{y}^2 - T_{yy} - \hat{\beta}E_{xy} - \hat{\gamma}E_{yz} = E_{yy} - \hat{\beta}E_{xy} - \hat{\gamma}T_{yz}.$$

The d.f. of S_2 is $p(q-1) - 2$.

The main objective of this analysis is to test the significance of

$$H_0 : \alpha_i = 0 \text{ against } H_A : \alpha_i \neq 0.$$

Under H_0 the model is :

$$y_{ij} = \mu + \beta(x_{ij} - \bar{x}_{..}) + \gamma(z_{ij} - \bar{z}_{..}). \quad (B)$$

The normal equations to estimate the parameters in the model (B) are :

$$y_{..} = pq\hat{\mu}, \quad G_{xy} = \hat{\beta}G_{xx} + \hat{\gamma}G_{xz}, \quad G_{yz} = \hat{\beta}G_{xz} + \hat{\gamma}G_{zz}.$$

On simplification, we get

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\beta} = \frac{G_{xy}G_{zz} - G_{xz}G_{yz}}{G_{xx}G_{zz} - G_{xz}^2}, \quad \hat{\gamma} = \frac{G_{yz}G_{xx} - G_{xy}G_{xz}}{G_{xx}G_{zz} - G_{xz}^2}.$$

The sum of squares of estimates is

$$\begin{aligned}
S_3 = SS(\text{estimate}) &= \hat{\mu}y_{..} + \sum \sum \hat{\beta}(x_{ij} - \bar{x}_{..})y_{ij} + \sum \sum \hat{\gamma}(z_{ij} - \bar{z}_{..})y_{ij} \\
&= pq\bar{y}^2 + \hat{\beta}G_{xy} + \hat{\gamma}G_{yz}.
\end{aligned}$$

The d.f. of S_3 is 3. The error sum of squares is

$$S_4 = SS(\text{error}) = \sum \sum y_{ij}^2 - pq\bar{y}^2 - \hat{\beta}G_{xy} - \hat{\gamma}G_{yz} = G_{yy} - \hat{\beta}G_{xy} - \hat{\gamma}G_{yz}.$$

The d.f. of S_4 is $(pq-3)$. Therefore, the test statistic for $H_0 : \alpha_i = 0$ is :

$$F = \frac{(S_4 - S_3)/p - 1}{S_4/(pq-3)}.$$

Here $S_4 - S_3 = SS(\hat{\alpha}_i)$ under $H_0 = T_{yy} + \hat{\beta}E_{xy} + \hat{\gamma}E_{yz} - \hat{\beta}G_{xy} - \hat{\gamma}G_{yz}$.

This F is distributed as variance ratio with $(p-1)$ and $(pq-3)$ d.f.

If the null hypothesis $H_0 : \alpha_i = 0$ is rejected, we need to compare the treatment effects in pairs. The null hypothesis for the comparison is :

$$H_0 : \alpha_i = \alpha_{i'} \quad (i \neq i' = 1, 2, \dots, p) \quad \text{or,} \quad H_0 : \alpha_i - \alpha_{i'} = 0.$$

The estimate of the contrast $\alpha_i - \alpha_{i'}$ ($i \neq i' = 1, 2, \dots, p$) is :

$$\hat{\alpha}_i - \hat{\alpha}_{i'} = (\bar{y}_{i.} - \bar{y}_{i'..}) - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) - \hat{\gamma}(\bar{z}_{i.} - \bar{z}_{..}).$$

The variance of the estimate of this contrast is :

$$\begin{aligned}
V(\hat{\alpha}_i - \hat{\alpha}_{i'}) &= \frac{2\sigma^2}{q} + (\bar{x}_{i.} - \bar{x}_{i'..})^2 V(\hat{\beta}) + (\bar{z}_{i.} - \bar{z}_{i'..})^2 V(\hat{\gamma}) \\
&\quad + 2(\bar{x}_{i.} - \bar{x}_{i'..})(\bar{z}_{i.} - \bar{z}_{i'..}) \text{Cov}(\hat{\beta}, \hat{\gamma}),
\end{aligned}$$

where $V(\hat{\beta}) = \frac{E_{zz}\sigma^2}{E_{xx}E_{zz} - E_{xz}^2}$, $V(\hat{\gamma}) = \frac{E_{xx}\sigma^2}{E_{xx}E_{zz} - E_{xz}^2}$, $Cov(\hat{\beta}, \hat{\gamma}) = \frac{E_{xz}\sigma^2}{E_{xx}E_{zz} - E_{xz}^2}$.

The estimate of σ^2 is $MS(\text{error}) = S_2/[p(q - 1) - 2]$.

Example 6.2 : In a dairy farm an experiment is conducted to study the impacts of 4 different types of fodder on milk production of cows. Each food is given to 5 different cows. But the amount of food (x kg) and lactation period (z) of cows are not same. The milk production (y kg) and the data on x and z are shown below :

Milk production of cows fed different fodder

Cows	Food-1			Food-2			Food-3			Food-4		
	y	x	z	y	x	z	y	x	z	y	x	z
1	28.5	10.5	2	26.5	14.2	2	24.2	12.2	1	18.4	9.2	1
2	19.6	9.6	3	32.6	16.4	3	26.4	11.6	2	22.7	10.4	2
3	30.2	12.2	1	30.2	15.2	2	35.6	15.0	3	22.0	11.6	3
4	26.4	11.4	2	24.6	12.6	1	30.4	14.2	2	26.2	12.0	3
5	28.7	11.8	1	25.2	12.8	1	28.6	11.8	2	17.2	9.0	1
Total y_i	133.4			139.1			145.2			106.5		
x_i	55.5			71.2			64.8			52.2		
z_i	9			9			10			10		

- (i) Analyse the data and comment.
- (ii) Is there any difference between food-3 and food-4?

Solution : (i) We have $p = 4$, $q = 5$, $G_y = 524.2$, $G_x = 243.7$.

$G_z = 38$, $C.T_y = \frac{G_y^2}{pq} = 13739.282$, $C.T_x = \frac{G_x^2}{pq} = 2969.4845$.

$C.T_z = \frac{G_z^2}{pq} = 72.2$, $C.T_{xy} = \frac{G_x G_y}{pq} = 6387.377$.

$C.T_{xz} = \frac{G_x G_z}{pq} = 463.03$, $C.T_{yz} = \frac{G_y G_z}{pq} = 995.98$.

$G_{yy} = \sum \sum y_{ij}^2 - C.T_y = 14158.56 - 13739.282 = 419.278$.

$G_{xx} = \sum \sum x_{ij}^2 - C.T_x = 3046.33 - 2969.4845 = 76.8455$.

$G_{zz} = \sum \sum z_{ij}^2 - C.T_z = 84 - 72.2 = 11.8$.

$G_{xy} = \sum \sum x_{ij} y_{ij} - C.T_{xy} = 6532.37 - 6387.377 = 144.993$.

$G_{xz} = \sum \sum x_{ij} z_{ij} - C.T_{xz} = 472.20 - 463.03 = 9.17$.

$G_{yz} = \sum \sum y_{ij} z_{ij} - C.T_{yz} = 1015.9 - 995.98 = 19.92$.

$T_{yy} = \frac{1}{q} \sum y_i^2 - C.T_y = \frac{69569.66}{5} - 13739.282 = 174.65$.

$T_{xx} = \frac{1}{q} \sum x_i^2 - C.T_x = \frac{15073.57}{5} - 2969.4845 = 45.2295$.

$$T_{zz} = \frac{1}{q} \sum z_i^2 - C.T_z = \frac{362}{5} - 72.2 = 0.2.$$

$$T_{xy} = \frac{1}{q} \sum x_i y_i - C.T_{xy} = \frac{32275.88}{5} - 6387.377 = 67.799.$$

$$T_{xz} = \frac{1}{q} \sum x_i z_i - C.T_{xz} = \frac{2310.3}{5} - 463.03 = -0.97.$$

$$T_{yz} = \frac{1}{q} \sum y_i z_i - C.T_{yz} = \frac{4969.5}{5} - 995.98 = -2.08.$$

$$E_{yy} = G_{yy} - T_{yy} = 244.628, E_{zz} = G_{zz} - T_{zz} = 31.616.$$

$$E_{xz} = G_{xz} - T_{xz} = 11.6.$$

$$E_{xy} = G_{xy} - T_{xy} = 77.194, E_{xz} = G_{xz} - T_{xz} = 10.14, E_{yz} = G_{yz} - T_{yz} = 22.0.$$

$$\hat{\beta} = \frac{E_{zz}E_{xy} - E_{xz}E_{yz}}{E_{zz}E_{zz} - E_{xz}^2} = \frac{11.6 \times 77.194 - 10.14 \times 22.0}{31.616 \times 11.6 - (10.14)^2} = \frac{672.3704}{263.926} = 2.547.$$

$$\hat{\gamma} = \frac{E_{zz}E_{yz} - E_{xz}E_{xy}}{E_{zz}E_{zz} - E_{xz}^2} = \frac{31.616 \times 22.0 - 10.14 \times 77.194}{31.616 \times 11.6 - (10.14)^2} = \frac{-87.1952}{263.926} = 0.330.$$

$$S_2 = E_{yy} - \hat{\beta}E_{xy} - \hat{\gamma}E_{yz} = 244.628 - 2.547 \times 77.194 + 0.33 \times 22.0 = 55.275.$$

$$\hat{\beta} = \frac{G_{zz}G_{xy} - G_{xz}G_{yz}}{G_{zz}G_{zz} - G_{xz}^2} = \frac{11.8 \times 144.993 - 9.17 \times 19.92}{76.8455 \times 11.8 - (9.17)^2} = \frac{1528.251}{822.688} = 1.858.$$

$$\hat{\gamma} = \frac{G_{zz}G_{yz} - G_{xz}G_{xy}}{G_{zz}G_{zz} - G_{xz}^2} = \frac{76.8455 \times 19.92 - 9.17 \times 144.993}{76.8455 \times 11.8 - (9.17)^2} = \frac{201.1765}{822.688} = 0.244.$$

$$S_5 = T_{yy} + \hat{\beta}E_{xy} + \hat{\gamma}E_{yz} - \hat{\beta}G_{xy} - \hat{\gamma}G_{yz} \\ = 174.65 + 2.547 \times 77.194 - 0.33 \times 22.0 - 1.858 \times 144.993 - 0.244 \times 19.92 = 89.746.$$

Therefore, to test the significance of food effect ($H_0 : \alpha_i = 0$) the test statistic is :

$$F = \frac{S_5/p - 1}{S_2/[p(q-1) - 2]} = \frac{89.746/3}{55.275/14} = 7.58.$$

Since $F > F_{0.05;3,14} = 3.34$, H_0 is rejected. The milk production changes with the change in food.

(ii) We need to test the significance of the hypothesis :

$$H_0 : \alpha_3 = \alpha_4 \quad \text{or,} \quad H_0 : \alpha_3 - \alpha_4 = 0.$$

The estimate of this contrast is : $\hat{\alpha}_3 - \hat{\alpha}_4$, where

$$\hat{\alpha}_3 - \hat{\alpha}_4 = \bar{y}_3 - \bar{y}_4 - \hat{\beta}(\bar{x}_3 - \bar{x}_4) - \hat{\gamma}(\bar{z}_3 - \bar{z}_4) \\ = (29.04 - 21.3) - 2.547(12.96 - 10.44) + 0.33(2.00 - 2.00) = 1.3216.$$

$$v(\hat{\beta}) = \frac{E_{zz}\hat{\sigma}^2}{E_{zz}E_{zz} - E_{xz}^2} = \frac{11.6 \times 3.9482}{263.926} = 0.1735, \quad \text{where } \hat{\sigma}^2 = \frac{S_2}{[p(q-1) - 2]} = 3.9482.$$

$$v(\hat{\gamma}) = \frac{E_{zz}\hat{\sigma}^2}{E_{zz}E_{zz} - E_{xz}^2} = \frac{31.616 \times 3.9482}{263.926} = 0.4729.$$

$$\text{Cov}(\hat{\beta}, \hat{\gamma}) = -\frac{E_{xx}\sigma^2}{E_{xx}E_{zz} - E_{zx}^2} = -\frac{1014 \times 3.9482}{263.926} = -0.1517.$$

$$\begin{aligned} \text{Now } v(\hat{\alpha}_3 - \hat{\alpha}_4) &= \frac{2\hat{\sigma}^2}{q} + (\bar{x}_3 - \bar{x}_4)^2 v(\hat{\beta}) + (\bar{z}_3 - \bar{z}_4)^2 v(\hat{\gamma}) \\ &\quad + 2(\bar{x}_3 - \bar{x}_4)(\bar{z}_3 - \bar{z}_4) \text{Cov}(\hat{\beta}\hat{\gamma}) \\ &= \frac{2 \times 3.9482}{5} + (12.96 - 10.44)^2 (0.1735) \quad [\because \bar{z}_3 = \bar{z}_4 = 2.00] \\ &= 2.6811. \end{aligned}$$

$$\text{Therefore, } t = \frac{\hat{\alpha}_3 - \hat{\alpha}_4}{\text{s.c}(\hat{\alpha}_3 - \hat{\alpha}_4)} = \frac{1.3216}{\sqrt{2.6811}} = 0.81.$$

Since $t < t_{0.05, 14} = 2.145$, H_0 is accepted. Food-3 does not differ significantly from food-4.

6.4 Covariance Analysis in Randomized Block Design with One Concomitant Variable

The model assumed for this analysis is

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma(x_{ij} - \bar{x}_{..}) + e_{ij}, \quad (\text{A})$$

$$i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q.$$

Here y_{ij} = the response of j -th treatment in i -th block, μ = general mean, α_i = effect of i -th block, β_j = effect of j -th treatment, x_{ij} = the value of concomitant variable corresponding to y_{ij} , γ = regression coefficient of y on x , e_{ij} = random error.

Assumption : $e_{ij} \sim \text{NID}(0, \sigma^2)$.

The normal equations to estimate the parameters in the model (A) are :

$$y_{..} = pq\hat{\mu} + q \sum \hat{\alpha}_i + p \sum \hat{\beta}_j$$

$$y_{i.} = q\hat{\mu} + q\hat{\alpha}_i + \sum \hat{\beta}_j + q\hat{\gamma}(\bar{x}_{i.} - \bar{x}_{..})$$

$$y_{.j} = p\hat{\mu} + \sum \hat{\alpha}_i + p\hat{\beta}_j + p\hat{\gamma}(\bar{x}_{.j} - \bar{x}_{..})$$

$$\sum \sum y_{ij}(x_{ij} - \bar{x}_{..}) = q \sum \hat{\alpha}_i(\bar{x}_{i.} - \bar{x}_{..}) + p \sum \hat{\beta}_j(\bar{x}_{.j} - \bar{x}_{..}) + \hat{\gamma} \sum \sum (x_{ij} - \bar{x}_{..})^2.$$

$$\text{Let } G_{yy} = \sum \sum (y_{ij} - \bar{y}_{..})^2, \quad G_{xx} = \sum \sum (x_{ij} - \bar{x}_{..})^2, \quad G_{xy} = \sum \sum (x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..})$$

$$T_{yy} = p \sum (\bar{y}_{.j} - \bar{y}_{..})^2, \quad T_{xx} = p \sum (\bar{x}_{i.} - \bar{x}_{..})^2, \quad T_{xy} = p \sum (\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{.j} - \bar{y}_{..})$$

$$B_{yy} = q \sum (\bar{y}_{i.} - \bar{y}_{..})^2, \quad B_{xx} = q \sum (\bar{x}_{.j} - \bar{x}_{..})^2, \quad B_{xy} = q \sum (\bar{x}_{.j} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..})$$

$$E_{yy} = G_{yy} - B_{yy} - T_{yy}, \quad E_{xx} = G_{xx} - B_{xx} - T_{xx}, \quad E_{xy} = G_{xy} - B_{xy} - T_{xy}.$$

Putting all values in normal equations and under restrictions, $\sum \hat{\alpha}_i = 0$, $\sum \hat{\beta}_j = 0$, we get

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\alpha}_i = (\bar{y}_{i.} - \bar{y}_{..}) - \hat{\gamma}(\bar{x}_{i.} - \bar{x}_{..}), \quad \hat{\beta}_j = (\bar{y}_{.j} - \bar{y}_{..}) - \hat{\gamma}(\bar{x}_{.j} - \bar{x}_{..})$$

$$\text{and } \hat{\gamma} = \frac{E_{xy}}{E_{xx}}.$$

The sum of squares due to estimates for the model is :

$$\begin{aligned} S_1 &= SS(\text{estimates}) \hat{\mu} y_{..} + \sum \hat{\alpha}_i y_{i.} + \sum \hat{\beta}_j y_{.j} + \hat{\gamma} \sum \sum y_{ij}(x_{ij} - \bar{x}_{..}) \\ &= pq\bar{y}_{..}^2 + T_{yy} + B_{yy} + \hat{\gamma} E_{xy}. \end{aligned}$$

This sum of squares has $(p + q)$ d.f. The sum of squares due to error is

$$S_2 = \sum \sum y_{ij}^2 - pq\bar{y}^2 - T_{yy} - B_{yy} - \hat{\gamma}E_{xy} = E_{yy} - \hat{\gamma}E_{xy}.$$

The d.f. of S_2 is $(pq - p - q)$.

The main objective of this analysis is to test the significance of

$$H_0 : \beta_j = 0 \text{ against } H_A : \beta_j \neq 0.$$

The model (A) under H_0 transforms to

$$y_{ij} = \mu + \alpha_i + \hat{\gamma}(x_{ij} - \bar{x}..) + e_{ij}. \quad (B)$$

The estimates of the parameters in the model (B) are :

$$\hat{\mu} = \bar{y}.., \hat{\alpha}_i = (\bar{y}_i. - \bar{y}..) - \hat{\gamma}(\bar{x}_i. - \bar{x}..) \text{ and } \hat{\gamma} = \frac{E_{xy} + T_{xy}}{E_{xx} + T_{xx}}.$$

The sum of squares due to estimates for the model is :

$$S_3 = pq\bar{y}^2 + B_{yy} + \hat{\gamma}(E_{xy} + T_{xy}).$$

The d.f. of S_3 is $(p + 1)$. The sum of squares of error is :

$$S_4 = \sum \sum y_{ij}^2 - pq\bar{y}^2 - B_{yy} - \hat{\gamma}(E_{xy} + T_{xy}) = E_{yy} + T_{yy} - \hat{\gamma}(E_{xy} + T_{xy}).$$

The d.f. of S_4 is $(pq - p - 1)$. Again, the sum of squares of treatment under H_0 is :

$$S_5 = SS(\text{treatment}) = S_4 - S_2 = T_{yy} + \hat{\gamma}E_{xy} + \hat{\gamma}(E_{xy} + T_{xy}).$$

This S_5 has $(q - 1)$ d.f. Hence, the test statistic to test the significance of $H_0 : \beta_j = 0$, is :

$$F = \frac{S_5/(q - 1)}{S_2/(pq - p - q)}.$$

This F is distributed as variance ratio with $(q - 1)$ and $(pq - p - q)$ d.f. Before performing the covariance analysis we need to test the significance of $H_0 : \gamma = 0$ against $H_A : \gamma \neq 0$. It is needed since the rejection of this hypothesis leads us to do the covariance analysis. If the regression coefficient γ is not significant, covariance analysis will not provide better information on treatment parameters. Now, under $H_0 : \gamma = 0$, the model takes the shape :

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}. \quad (C)$$

This is the usual model for the data collected from randomized block design. The sum of squares of error for the model (C) is $S_6 = E_{yy}$. This S_6 has $(p - 1)(q - 1)$ d.f. Hence, the sum of squares due to the estimate of γ is :

$$S_7 = S_6 - S_2 = \hat{\gamma}E_{xy}.$$

This S_7 has 1 d.f. Therefore, the test statistic to test the significance of $H_0 : \gamma = 0$ is :

$$F = \frac{S_7}{S_2/(pq - p - q)}.$$

This F is distributed as variance ratio with 1 and $(pq - p - q)$ d.f.

The significance of regression parameter γ can also be tested by t -test, where the test statistic is :

$$t = \frac{\hat{\gamma} - \gamma_0}{\sqrt{\frac{\hat{\sigma}^2}{E_{xx}}}}.$$

Here the null hypothesis is $H_0 = \gamma = \gamma_0$ (a specified value). This t has $(pq - p - q)$ d.f. Here

$$\hat{\sigma}^2 = S_2 / (pq - p - q).$$

The analysis provides adjusted treatment means, where the adjustment is done to eliminate the impact of concomitant variable. The adjusted treatment means are :

$$\bar{y}_{.j} \text{ (adjusted)} = \bar{y}_{.j} - \hat{\gamma}(\bar{x}_{.j} - \bar{x}_{..}); j = 1, 2, \dots, q$$

The variance of this adjusted mean is :

$$V[(\bar{y}_{.j}) \text{ adjusted}] = \sigma^2 \left[\frac{1}{p} + \frac{(\bar{x}_{.j} - \bar{x}_{..})^2}{E_{xx}} \right].$$

Also, we have

$$\sum_j d_j \hat{\beta}_j = \sum_j d_j [(\bar{y}_{.j} - \bar{y}_{..}) - \hat{\gamma}(\bar{x}_{.j} - \bar{x}_{..})], \sum_j d_j = 0.$$

The variance of this estimate of contrast $\sum d_j \hat{\beta}_j$ is :

$$V\left(\sum d_j \hat{\beta}_j\right) = \sigma^2 \left[\frac{\sum d_j^2}{p} + \frac{(\sum d_j \bar{x}_{.j})^2}{E_{xx}} \right].$$

To get the estimate of this variance σ^2 is to be replaced by $\hat{\sigma}^2 = S_2 / (pq - p - q)$. However, to compare the means of two treatments, we have average variance of the difference of any pair of treatment means. This variance is :

$$V(\bar{y}_{.j} - \bar{y}_{.s}) = \frac{2\sigma^2}{p} \left[1 + \frac{T_{xx} / (q - 1)}{E_{xx}} \right], j \neq s = 1, 2, \dots, q.$$

ANCOVA Table

Source of variation	$SS(x)$	$SS(y)$	$SP(xy)$	Regression coefficient	Adjusted SS	d.f.	SS under H_0
Blocks	B_{xx}	B_{yy}	B_{xy}			$p - 1$	
Treatment	T_{xx}	T_{yy}	T_{xy}	$\hat{\gamma} = \frac{E_{xy}}{E_{xx}}$		$q - 1$	$S_5 = S_4 - S_2$
Error	E_{xx}	E_{yy}	E_{xy}		$E_{yy} - \hat{\gamma}E_{xy}$	$pq - p - q$	
Treatment + Error	$E_{xx} + T_{xx}$	$E_{yy} + T_{yy}$	$E_{xy} + T_{xy}$				

Example 6.3 : In an agricultural research station an experiment is conducted to study the productivity of balsam apple under nitrogen fertilizer. Four levels of nitrogen as urea are used in the experiment. The levels are 30 kg/ha, 60 kg/ha, 90 kg/ha and 120 kg/ha. The levels of nitrogen are allocated to 4 different plots of a block. The design used is randomized block design having 5 blocks. The production of balsam apple in a plot of size 10' x 15' are recorded along with the number of plants per plot.

Production of balsam apple (y_{ij} kg) along with number of plants (x_{ij})

Blocks	Levels of nitrogen									
	N_1		N_2		N_3		N_4		Total	
	y	x	y	x	y	x	y	x	y_i	x_i
1	50.5	42	55.0	52	52.3	45	60.0	50	217.8	189
2	48.0	50	60.2	55	51.4	48	61.0	48	220.6	201
3	40.5	46	40.5	48	35.0	40	65.0	52	181.0	186
4	52.0	48	50.0	51	48.0	46	35.0	35	185.0	180
5	50.0	45	56.0	48	55.0	49	38.0	40	199.0	182
Total y_j	241.0		261.7		241.7		259.0		1003.4	
x_j	231		254		228		225		938	

- (i) Analyse the data and group the levels of nitrogen.
(ii) Justify the use of covariance analysis.
(iii) Estimate the efficiency of covariance analysis in randomized block design compared to analysis of randomized block design.

Solution : We have $p = 5$, $q = 4$, $G_y = 1003.4$, $G_x = 938$

$$C.T_y = \frac{G_y^2}{pq} = 50340.578, \quad C.T_x = \frac{G_x^2}{pq} = 43992.2$$

$$C.T_{xy} = \frac{G_x G_y}{pq} = 47059.46, \quad G_{yy} = \sum \sum y_{ij}^2 - C.T_y = 51770.04 - 50340.578 = 1429.462.$$

$$G_{xx} = \sum \sum x_{ij}^2 - C.T_x = 44426 - 43992.2 = 433.8.$$

$$G_{xy} = \sum \sum x_{ij} y_{ij} - C.T_{xy} = 47659.7 - 47059.46 = 600.24.$$

$$B_{yy} = \frac{1}{q} \sum y_i^2 - C.T_y = \frac{202688.2}{4} - 50340.578 = 331.472.$$

$$T_{yy} = \frac{1}{p} \sum y_j - C.T_y = \frac{252067.78}{5} - 50340.578 = 72.978.$$

$$B_{xx} = \frac{1}{q} \sum x_i^2 - C.T_x = \frac{176242}{4} - 43992.2 = 68.3.$$

$$T_{xx} = \frac{1}{p} \sum x_j^2 - C.T_x = \frac{220486}{5} - 43992.2 = 105.0.$$

$$B_{xy} = \frac{1}{q} \sum x_i y_i - C.T_{xy} = \frac{188688.8}{4} - 47059.46 = 112.74.$$

$$T_{xy} = \frac{1}{p} \sum x_j y_j - C.T_{xy} = \frac{235525.4}{5} - 47059.46 = 45.62.$$

$$E_{yy} = G_{yy} - B_{yy} - T_{yy} = 1429.462 - 331.472 - 72.978 = 1025.012.$$

$$E_{xx} = G_{xx} - B_{xx} - T_{xx} = 433.8 - 68.3 - 105.0 = 260.5.$$

$$E_{xy} = G_{xy} - B_{xy} - T_{xy} = 600.24 - 112.74 - 45.62 = 441.88.$$

$$\hat{\gamma} = \frac{E_{xy}}{E_{xx}} = \frac{441.88}{260.5} = 1.6963.$$

$$S_2 = E_{yy} - \hat{\gamma}E_{xy} = 1025.012 - 1.6963 \times 441.88 = 275.451$$

$$\hat{\gamma} = \frac{E_{xy} + T_{xy}}{E_{xx} + T_{xx}} = \frac{441.88 + 45.62}{260.5 + 105.0} = 1.3338.$$

$$S_4 = E_{yy} + T_{yy} - \hat{\gamma}(E_{xy} + T_{xy}) = 1025.012 + 72.978 - 1.3338(441.88 + 45.62) = 447.7625.$$

$$S_5 = SS(\text{Treatment}) \text{ adjusted} = S_4 - S_2 = 447.7625 - 275.451 = 172.3115.$$

Therefore, to test the significance of $H_0 : \beta_j = 0$, the test statistic is :

$$F = \frac{S_5/(q-1)}{S_2/(pq-p-q)} = \frac{172.3115/3}{275.451/11} = 2.29.$$

Since $F < F_{0.05;3,11} = 3.59$, H_0 is accepted. The levels of nitrogen belong to one group.

ANCOVA Table

Source of variation	SS(x)	SP(xy)	SS(y)	Adjusted	
				SS	d.f.
Blocks	68.3	112.74	331.472		
Treatment	105.0	45.62	72.978	172.3115	3
Error	260.5	441.88	1025.012	275.451	11

(ii) To justify the use of concomitant variable we need to test the significance of

$$H_0 : \gamma = 0 \text{ against } H_A : \gamma \neq 0.$$

The test statistic is :

$$F = \frac{\hat{\gamma}E_{xy}}{S_2/(pq-p-q)} = \frac{1.6963 \times 441.88}{275.451/11} = 29.93.$$

Since $F > F_{0.05;1,11} = 4.84$, H_0 is rejected. Regression is significant. Hence covariance analysis needed.

(iii) The average variance of the difference of two adjusted treatment means is

$$\begin{aligned} v(\bar{y}_{.j} - \bar{y}_{.j'}) &= \frac{2\hat{\sigma}^2}{p} \left[1 + \frac{T_{xx}/q - 1}{E_{xx}} \right], \hat{\sigma}^2 = MS(\text{error}) = 25.041 \\ &= \frac{25.041}{5} \left[1 + \frac{105.0/4 - 1}{260.5} \right] = 5.681. \end{aligned}$$

If the analysis is done without covariance analysis, then

$$\hat{\sigma}_R^2 = \frac{E_{yy}}{(p-1)(q-1)} = \frac{1025.012}{(5-1)(4-1)} = 85.418.$$

In such analysis,

$$v(\bar{y}_{.j} - \bar{y}_{.j'}) = \frac{2\hat{\sigma}_R^2}{p} = 34.167.$$

Hence, the relative efficiency of covariance analysis in randomized block design compared to simple randomized design is

$$\frac{v(\bar{y}_{.j} - \bar{y}_{.j'}) \text{ in case of RBD}}{v(\bar{y}_{.j} - \bar{y}_{.j'}) \text{ adjusted}} = \frac{34.167}{5.681} = 601.4\%.$$

6.5 Covariance Analysis in Randomized Block Design with Two Concomitant Variables

The model assumed for this analysis is

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma(x_{ij} - \bar{x}_{..}) + \delta(z_{ij} - \bar{z}_{..}) + e_{ij}, \quad (A)$$

$$i = 1, 2, \dots, p; j = 1, 2, \dots, q.$$

Here y_{ij} = result of j -th treatment in i -th block, α_i = effect of i -th block, β_j = effect of j -th treatment, x_{ij} = the value of one concomitant variable corresponding to y_{ij} , z_{ij} = the value of another concomitant variable corresponding to y_{ij} , γ = regression coefficient of y on x , δ = regression coefficient of y on z , e_{ij} = random error.

The normal equations to estimate the parameters in the model (A) are :

$$\begin{aligned} y_{..} &= pq\hat{\mu} + q \sum \hat{\alpha}_i + p \sum \hat{\beta}_j \\ y_{i.} &= q\hat{\mu} + q\hat{\alpha}_i + \sum \hat{\beta}_j + q\hat{\gamma}(\bar{x}_{i.} - \bar{x}_{..}) + q\hat{\delta}(\bar{z}_{i.} - \bar{z}_{..}) \\ y_{.j} &= p\hat{\mu} + \sum \hat{\alpha}_i + p\hat{\beta}_j + p\hat{\gamma}(\bar{x}_{.j} - \bar{x}_{..}) + p\hat{\delta}(\bar{z}_{.j} - \bar{z}_{..}) \\ \sum \sum y_{ij}(x_{ij} - \bar{x}_{..}) &= q \sum \hat{\alpha}_i(\bar{x}_{i.} - \bar{x}_{..}) + p \sum \hat{\beta}_j(\bar{x}_{.j} - \bar{x}_{..}) \\ &\quad + \hat{\gamma} \sum \sum (x_{ij} - \bar{x}_{..})^2 + \hat{\delta} \sum \sum (z_{ij} - \bar{z}_{..})(x_{ij} - \bar{x}_{..}) \\ \sum \sum y_{ij}(z_{ij} - \bar{z}_{..}) &= q \sum \hat{\alpha}_i(\bar{z}_{i.} - \bar{z}_{..}) + p \sum \hat{\beta}_j(\bar{z}_{.j} - \bar{z}_{..}) \\ &\quad + \hat{\gamma} \sum \sum (x_{ij} - \bar{x}_{..})(z_{ij} - \bar{z}_{..}) + \hat{\delta} \sum \sum (z_{ij} - \bar{z}_{..})^2. \end{aligned}$$

We can define G_{yy} , G_{xx} , G_{xy} , B_{yy} , B_{xx} , B_{xy} , T_{yy} , T_{xx} , T_{xy} , E_{yy} , E_{xx} and E_{xy} as these are defined in section (6.4).

Moreover, let

$$\begin{aligned} G_{zz} &= \sum \sum (z_{ij} - \bar{z}_{..})^2, G_{xz} = \sum \sum (x_{ij} - \bar{x}_{..})(z_{ij} - \bar{z}_{..}), \\ B_{zz} &= q \sum (\bar{z}_{i.} - \bar{z}_{..})^2, B_{xz} = q \sum (\bar{x}_{i.} - \bar{x}_{..})(\bar{z}_{i.} - \bar{z}_{..}), \\ T_{zz} &= p \sum (\bar{z}_{.j} - \bar{z}_{..})^2, T_{xz} = p \sum (\bar{x}_{.j} - \bar{x}_{..})(\bar{z}_{.j} - \bar{z}_{..}), \\ G_{yz} &= \sum \sum (y_{ij} - \bar{y}_{..})(z_{ij} - \bar{z}_{..}), B_{yz} = q \sum (\bar{y}_{i.} - \bar{y}_{..})(\bar{z}_{i.} - \bar{z}_{..}), \\ T_{yz} &= p \sum (\bar{y}_{.j} - \bar{y}_{..})(\bar{z}_{.j} - \bar{z}_{..}), E_{zz} = G_{zz} - B_{zz} - T_{zz}, \\ E_{xz} &= G_{xz} - B_{xz} - T_{xz}, E_{yz} = G_{yz} - B_{yz} - T_{yz}. \end{aligned}$$

Replacing these values in the normal equations and putting the restrictions $\sum \hat{\alpha}_i = \sum \hat{\beta}_j = 0$, we get, on simplification

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..}, \hat{\alpha}_i = (\bar{y}_{i.} - \bar{y}_{..}) - \hat{\gamma}(\bar{x}_{i.} - \bar{x}_{..}) - \hat{\delta}(\bar{z}_{i.} - \bar{z}_{..}) \\ \hat{\beta}_j &= (\bar{y}_{.j} - \bar{y}_{..}) - \hat{\gamma}(\bar{x}_{.j} - \bar{x}_{..}) - \hat{\delta}(\bar{z}_{.j} - \bar{z}_{..}) \\ \hat{\gamma} &= \frac{E_{zz}E_{xy} - E_{xz}E_{yz}}{E_{xx}E_{zz} - E_{xz}^2}, \hat{\delta} = \frac{E_{xx}E_{yz} - E_{xz}E_{xy}}{E_{xx}E_{zz} - E_{xz}^2}. \end{aligned}$$

The sum of squares due to estimates is :

$$\begin{aligned} S_1 &= \hat{\mu}y_{..} + \sum \hat{\alpha}_iy_{i.} + \sum \hat{\beta}_jy_{.j} + \hat{\gamma} \sum \sum (x_{ij} - \bar{x}_{..})y_{ij} + \hat{\delta} \sum \sum (z_{ij} - \bar{z}_{..})y_{ij} \\ &= pq\bar{y}_{..}^2 + T_{yy} + B_{yy} + \hat{\gamma}E_{xy} + \hat{\delta}E_{yz}. \end{aligned}$$

The d.f. of S_1 is $(p + q + 1)$. The sum of squares of error is :

$$S_2 = \sum \sum y_{ij}^2 - pq\bar{y}^2 - T_{yy} - B_{yy} - \hat{\gamma}E_{xy} - \hat{\delta}E_{zy} = E_{yy} - \hat{\gamma}E_{xy} - \hat{\delta}E_{zy}.$$

The d.f. of S_2 is $(pq - p - q + 1)$.

The main objective of this analysis is to test the significance of

$$H_0 : \beta_j = 0 \text{ against } H_A : \beta_j \neq 0.$$

Under this null hypothesis the model stands

$$y_{ij} = \mu + \alpha_i + \gamma(x_{ij} - \bar{x}_{..}) + \delta(z_{ij} - \bar{z}_{..}) + e_{ij}.$$

The estimates of the parameters in the model are :

$$\hat{\mu} = \bar{y}_{..}, \hat{\alpha}_i = (\bar{y}_{.i} - \bar{y}_{..}) - \hat{\gamma}(\bar{x}_{.i} - \bar{x}_{..}) - \hat{\delta}(\bar{z}_{.i} - \bar{z}_{..}).$$

$$\hat{\gamma} = \frac{E'_{zz}E'_{xy} - E'_{xz}E'_{yz}}{E'_{xx}E'_{zz} - E'^2_{xz}}, \hat{\delta} = \frac{E'_{xx}E'_{yz} - E'_{xz}E'_{xy}}{E'_{xx}E'_{zz} - E'^2_{xz}}.$$

Here $E'_{xx} = E_{xx} + T_{xx}$, $E'_{zz} = E_{zz} + T_{zz}$, $E'_{xz} = E_{xz} + T_{xz}$

$$E'_{xy} = E_{xy} + T_{xy}, E'_{yz} = E_{yz} + T_{yz}, E'_{yy} = E_{yy} + T_{yy}.$$

The sum of squares due to estimates in case of model is :

$$\begin{aligned} S_3 &= \hat{\mu}y_{..} + \sum \hat{\alpha}_i y_{i.} + \hat{\gamma} \sum \sum (x_{ij} - \bar{x}_{..}) y_{ij} + \hat{\delta} \sum \sum (z_{ij} - \bar{z}_{..}) y_{ij} \\ &= pq\bar{y}^2 + B_{yy} + \hat{\gamma}E'_{xy} + \hat{\delta}E'_{yz}. \end{aligned}$$

The d.f. of S_3 is $(p + 2)$. The sum of squares due to error is

$$S_4 = \sum \sum y_{ij}^2 - pq\bar{y}^2 - B_{yy} - \hat{\gamma}E'_{xy} - \hat{\delta}E'_{yz}.$$

The d.f. of S_4 is $(pq - p - 2)$. Hence, the sum of squares of treatment under H_0 is

$$S_5 = SS(\text{treatment}) \text{ adjusted} = S_4 - S_2 = T_{yy} + \hat{\gamma}E_{xy} + \hat{\delta}E_{yz} - \hat{\gamma}E'_{xy} - \hat{\delta}E'_{yz}.$$

The d.f. of S_5 is $(q - 1)$. Hence, the test statistic for the hypothesis is

$$F = \frac{S_5/q - 1}{S_2/(pq - p - q + 1)}.$$

The estimate of treatment contrast after eliminating the effect of concomitant variable is

$$\hat{\beta}_j - \hat{\beta}_{j'} = (\bar{y}_{.j} - \bar{y}_{.j'}) - \hat{\gamma}(\bar{x}_{.j} - \bar{x}_{.j'}) - \hat{\delta}(\bar{z}_{.j} - \bar{z}_{.j'}).$$

The variance of this estimated contrast is

$$\begin{aligned} V(\hat{\beta}_j - \hat{\beta}_{j'}) &= \frac{2\sigma^2}{p} + (\bar{x}_{.j} - \bar{x}_{.j'})^2 V(\hat{\gamma}) + (\bar{z}_{.j} - \bar{z}_{.j'})^2 V(\hat{\delta}) \\ &\quad + 2(\bar{x}_{.j} - \bar{x}_{.j'}) (\bar{z}_{.j} - \bar{z}_{.j'}) \text{Cov}(\hat{\gamma}, \hat{\delta}). \end{aligned}$$

$$\text{Here } V(\hat{\gamma}) = \frac{E_{zz}\sigma^2}{E_{xx}E_{zz} - E_{xz}^2}, V(\hat{\delta}) = \frac{E_{xx}\sigma^2}{E_{xx}E_{zz} - E_{xz}^2} \text{ and } \text{Cov}(\hat{\gamma}, \hat{\delta}) = -\frac{E_{xz}\sigma^2}{E_{xx}E_{zz} - E_{xz}^2}.$$

The estimate of σ^2 is $\hat{\sigma}^2 = S_2/(pq - p - q + 1)$.

Example 6.4 : In an agricultural research station an experiment is conducted to study the impact of 3 levels of nitrogen in producing sugarcane. The levels of nitrogen are $N_1 = 200$ kg/acre, $N_2 = 220$ kg/acre and $N_3 = 240$ kg/acre. The nitrogen fertilizer is applied in plots of 5 blocks. But the sizes of plots (x sq. ft) and hence, the number of plants in plots (z) are not same. The production of sugarcane in different plots (y kg) along with the values of x and z are shown below :

Blocks	Treatment									Total		
	N_1			N_2			N_3			x_i	y_i	z_i
	x	y	z	x	y	z	x	y	z			
1	300	85	28	400	125	30	350	105	28	1050	315	86
2	450	120	40	400	115	35	200	95	15	1050	330	90
3	400	105	32	300	88	25	300	110	26	1000	303	83
4	300	90	25	300	99	27	400	125	32	1000	314	84
5	350	100	30	300	94	25	400	120	30	1050	314	85
Total $x_{.j}$	1800			1700			1650			5150		
$y_{.j}$	500			521			555			1576		
$z_{.j}$	155			142			131			428		

(i) Analyse the data and group the levels of nitrogen.

(ii) Is there any justification of using concomitant variables?

Solution : (i) We have $p = 5$, $q = 3$, $G_x = 5150$, $G_y = 1576$.

$$C.T_x = \frac{G_x^2}{pq} = 1768166.667, \quad C.T_y = \frac{G_y^2}{pq} = 165585.0667.$$

$$C.T_{xy} = \frac{G_x G_y}{pq} = 541093.333, \quad G_z = 428, \quad C.T_z = \frac{G_z^2}{pq} = 12212.2667,$$

$$C.T_{xz} = \frac{G_x G_z}{pq} = 146946.667, \quad C.T_{yz} = \frac{G_y G_z}{pq} = 44968.533.$$

$$G_{yy} = \sum \sum y_{ij}^2 - C.T_y = 168156 - 165585.0667 = 2570.9333.$$

$$G_{xx} = \sum \sum x_{ij}^2 - C.T_x = 1827500 - 1768166.667 = 59333.333.$$

$$G_{zz} = \sum \sum z_{ij}^2 - C.T_z = 12646 - 12212.2667 = 433.7333.$$

$$G_{xy} = \sum \sum x_{ij} y_{ij} - C.T_{xy} = 550550 - 541093.333 = 9456.667.$$

$$G_{xz} = \sum \sum x_{ij} z_{ij} - C.T_{xz} = 151700 - 146946.667 = 4753.333.$$

$$G_{yz} = \sum \sum y_{ij} z_{ij} - C.T_{yz} = 45613 - 44968.533 = 644.467.$$

$$B_{yy} = \frac{1}{q} \sum y_i^2 - C.T_y = \frac{497126}{3} - 165585.0667 = 123.60.$$

$$B_{xx} = \frac{1}{q} \sum x_i^2 - C.T_x = \frac{5307500}{3} - 1768166.667 = 999.9996.$$

$$B_{zz} = \frac{1}{q} \sum z_i^2 - C.T_z = \frac{36666}{3} - 12212.2667 = 9.7333.$$

$$B_{xy} = \frac{1}{q} \sum x_i y_i - C.T_{xy} = \frac{1623950}{3} - 541093.333 = 223.3334.$$

$$B_{xz} = \frac{1}{q} \sum x_i z_i - C.T_{xz} = \frac{441050}{3} - 146946.667 = 69.9997.$$

$$B_{yz} = \frac{1}{q} \sum y_i z_i - C.T_{yz} = \frac{135005}{3} - 44968.533 = 33.1337.$$

$$T_{yy} = \frac{1}{p} \sum y_j^2 - C.T_y = \frac{829466}{5} - 165585.0667 = 308.1333$$

$$T_{zz} = \frac{1}{p} \sum z_j^2 - C.T_z = \frac{8852500}{5} - 1768166.667 = 2333.333$$

$$T_{zz} = \frac{1}{p} \sum z_j^2 - C.T_z = \frac{61350}{5} - 12212.2667 = 57.7333$$

$$T_{xy} = \frac{1}{p} \sum x_j y_j - C.T_{xy} = \frac{2701450}{5} - 541093.333 = -803.333$$

$$T_{xz} = \frac{1}{p} \sum x_j z_j - C.T_{xz} = \frac{736550}{5} - 146946.667 = 363.333$$

$$T_{yz} = \frac{1}{p} \sum y_j z_j - C.T_{yz} = \frac{224187}{5} - 44968.533 = -131.133$$

$$E_{yy} = G_{yy} - T_{yy} - B_{yy} = 2570.9333 - 308.1333 - 123.60 = 2139.2$$

$$E_{xx} = G_{xx} - B_{xx} - T_{xx} = 59333.333 - 999.9996 - 2333.333 = 56000.0004$$

$$E_{xz} = G_{xz} - B_{xz} - T_{xz} = 433.7333 - 9.7333 - 57.7333 = 366.2667$$

$$E_{xy} = G_{xy} - B_{xy} - T_{xy} = 9456.667 - 223.3334 + 803.333 = 10036.6666$$

$$E_{xz} = G_{xz} - B_{xz} - T_{xz} = 4753.333 - 69.9997 - 363.333 = 4320.0003$$

$$E_{yz} = G_{yz} - B_{yz} - T_{yz} = 644.467 - 33.1337 + 131.133 = 742.4663.$$

$$\hat{\gamma} = \frac{E_{xz}E_{xy} - E_{xz}E_{yz}}{E_{xx}E_{zz} - E_{xz}^2} = \frac{468642.1158}{1848532.754} = 0.253$$

$$\hat{\delta} = \frac{E_{xz}E_{yz} - E_{xz}E_{xy}}{E_{xx}E_{zz} - E_{xz}^2} = \frac{-1780289.626}{1848532.754} = -0.963.$$

$$S_2 = SS(\text{error}) = E_{yy} - \hat{\gamma}E_{xy} - \hat{\delta}E_{yz} \\ = 2139.2 - 0.253 \times 10036.6666 + 0.963 \times 742.4663 = 314.918$$

$$E'_{yy} = E_{yy} + T_{yy} = 2447.3333, E'_{xx} = E_{xx} + T_{xx} = 58333.3334$$

$$E'_{xz} = E_{xz} + T_{xz} = 424.0, E'_{xy} = E_{xy} + T_{xy} = 9233.3336$$

$$E'_{xz} = E_{xz} + T_{xz} = 4683.3333, E'_{yz} = E_{yz} + T_{yz} = 611.3333$$

$$\hat{\gamma} = \frac{E'_{xz}E'_{xy} - E'_{xz}E'_{yz}}{E'_{xx}E'_{zz} - E_{xz}^2} = \frac{1051857.25}{2799722.563} = 0.376$$

$$\hat{\delta} = \frac{E'_{xz}E'_{yz} - E'_{xz}E'_{xy}}{E'_{xx}E'_{zz} - E_{xz}^2} = \frac{-7581669.511}{2799722.563} = -2.708$$

$$S_4 = E'_{yy} - \hat{\gamma}E'_{xy} - \hat{\delta}E'_{yz}$$

$$= 2447.3333 - 0.376 \times 9233.3336 + 2.708 \times 611.3333 = 631.089$$

$$S_5 = SS(\text{Treatment}) \text{ adjusted} = S_4 - S_2 = 631.089 - 611.3333 = 19.7557.$$

Hence, to test the significance of $H_0 : \beta_j = 0$ against $H_A : \beta_j \neq 0$ the test statistic is :

$$F = \frac{S_5/q - 1}{S_2/(pq - p - q - 1)} = \frac{19.7557/2}{314.918/6} = 0.19.$$

Since $F < F_{0.05;2,6} = 5.14$, H_0 is accepted. There is no significant change in the production levels due to the changes in the levels of nitrogen. All levels of nitrogen belong to one group.

ANCOVA Table

Sources of variation	$SS(x)$	$SS(y)$	$SS(z)$	$SP(xy)$	$SP(xz)$	$SP(yz)$	SS adjusted
Block	999.9996	123.60	9.7333	223.3334	69.9997	33.1337	—
Treatment	2333.333	308.1333	57.7333	-803.333	363.333	-131.133	19.7557
Error	56000.0004	2139.2	366.2667	10036.6666	4320.0003	742.4663	314.918

The adjusted treatment means are shown in the table below :

Levels of nitrogen	$\bar{y}_{.j}$	$\bar{x}_{.j}$	$\bar{z}_{.j}$	$\hat{\gamma}(\bar{x}_{.j} - \bar{x}_{..})$	$\hat{\delta}(\bar{z}_{.j} - \bar{z}_{..})$	adjusted means = $(\bar{y}_{.j} - \hat{\gamma}(\bar{x}_{.j} - \bar{x}_{..}) - \hat{\delta}(\bar{z}_{.j} - \bar{z}_{..}))$
N_1	100.00	360.00	31.00	4.217	-2.379	98.162
N_2	104.20	340.00	28.40	-0.842	0.125	104.917
N_3	111.00	330.00	26.20	-3.372	2.244	111.128

To compare the adjusted means the estimated variance of the difference of two means is

$$V(\bar{y}_{.j} - \bar{y}_{.s}) \text{ adjusted} = \frac{2S^2}{P}, \text{ where}$$

$$S^2 = \hat{\sigma}^2 \left[1 + \frac{T_{xx}E_{zz} - 2T_{xz}E_{xz} + T_{zz}E_{xx}}{(q-1)(E_{xx}E_{zz} - E_{xz}^2)} \right], \quad \begin{aligned} \hat{\sigma}^2 &= \text{M.S. (error)} \\ &= S_2/(pq - p - q - 1) \\ &= 52.486 \end{aligned}$$

$$= 52.486 \left[1 + \frac{948489.6629}{(3-1)1848532.754} \right] = 65.95.$$

However, the means need not be compared since by F -test the means are found homogeneous.

(ii) To justify the use of concomitant variable we need to test the significance of

$$H_0 : \gamma = 0 \quad \text{and} \quad H_0 : \delta = 0.$$

If any of the hypothesis is rejected, the use of that corresponding concomitant variable is needed. We have

$$V(\hat{\gamma}) = \frac{E_{zz}\hat{\sigma}^2}{E_{xx}E_{zz} - E_{xz}^2} = \frac{366.2667 \times 52.486}{1848532.754} = 0.010399$$

$$s.c(\hat{\gamma}) = \sqrt{V(\hat{\gamma})} = 0.10198$$

$$V(\hat{\delta}) = \frac{E_{xx}\hat{\sigma}^2}{E_{xx}E_{zz} - E_{xz}^2} = \frac{56000.0004 \times 52.486}{1848532.754} = 1.590026$$

$$s.e(\hat{\delta}) = \sqrt{V(\hat{\delta})} = 1.26096.$$

Hence, we get

$$t = \frac{\hat{\gamma}}{s.c(\hat{\gamma})} = \frac{0.253}{0.10198} = 2.48 \quad \text{and} \quad t = \frac{\hat{\delta}}{s.e(\hat{\delta})} = \frac{-0.963}{1.26096} = -0.76.$$

Since $t_{0.05,6} = 2.447$, there is no evidence against $H_0 : \delta = 0$. So the concomitant variable z is not useful in the analysis. Again, $H_0 : \gamma = 0$ is rejected indicating the necessity of concomitant variable x .

6.6 Technique of Covariance Analysis in Analysing Data of Randomized Block Design with Missing Observations

We have discussed the technique of analysis of data with one and two missing observations in case of randomized block design. The technique is similar in case of data obtained from Latin square design, split-plot design, split-split-plot design, BIB design and other designs. But the technique becomes complicated if there are more than two missing observations. In such a situation technique of covariance analysis can be applied assuming one concomitant variable for one missing observation. Let us describe the technique to analyse the data of a randomized block design when there are two missing values. The method can be generalized if there are more than two missing values.

Let the observation of j -th treatment in i -th block (y_{ij}) and the observation of s -th treatment in k -th block (y_{ks}) be missing. Let us use one concomitant variable x for the missing observation of j -th treatment in i -th block and z for the missing observation of s -th treatment in k -th block. The x_{ij} be the value of concomitant variable x corresponding to y_{ij} , and z_{ks} be the value of concomitant variable z corresponding to y_{ks} . The values of x and z are either zero or 1. Let $x_{ij} = 1$ and other x_{ij} 's are zero. Similarly, $z_{ks} = 1$ and other z_{ij} 's are zero. The data set of the experiment can be shown as follows :

Block	Treatment												Total							
	1			2...			..	$j \dots$...	$\dots s$			$\dots q$			x_i	y_i	z_i
x	y	z	x	y	z	.	x	y	z	...	x	y	z	x	y	z				
1	0	y_{11}	0	0	y_{12}	0	...	0	y_{1j}	0	...	0	y_{1s}	0	0	y_{1q}	0	0	y_1	0
2	0	y_{21}	0	0	y_{22}	0	...	0	y_{2j}	0	...	0	y_{2s}	0	0	y_{2q}	0	0	y_2	0
⋮		
i	0	y_{i1}	0	0	y_{i2}	0	...	1	0	0	...	0	y_{is}	0	0	y_{iq}	0	1	y_i	0
⋮		
k	0	y_{k1}	0	0	y_{k2}	0	...	0	y_{kj}	0	...	0	0	1	0	y_{kq}	0	0	y_k	1
⋮		
p	0	y_{p1}	0	0	y_{p2}	0	...	0	y_{pj}	0	...	0	y_{ps}	0	0	y_{pq}	0	0	y_p	0
Total $x_{.j}$	0	0		0			1	0			0	0		0		0		1	$y_{.j}$	1
$y_{.j}$	$y_{.1}$		$y_{.2}$				$y_{.j}$	$y_{.s}$				$y_{.q}$								
$z_{.j}$	0		0				0					1		0						

The model for the above observations of y_{ij} is

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma(x_{ij} - \bar{x}_{..}) + \delta(z_{ij} - \bar{z}_{..}) + e_{ij}. \quad (X)$$

The parameters in the model have usual meanings. The analysis of the model (X) is to be performed in a similar way as it is done in analysing model (A) discussed in section 6.5. However, the sum of squares and sum of products related to the variables x and z have some definite values. These values are :

$$T_{xx} = \frac{1}{p} - \frac{1}{pq}, \quad B_{xx} = \frac{1}{q} - \frac{1}{pq}, \quad G_{xx} = 1 - \frac{1}{pq} \quad \left[\because \text{C.T}_x = \frac{1}{pq} \right]$$

$$T_{zz} = \frac{1}{p} - \frac{1}{pq}, \quad B_{zz} = \frac{1}{q} - \frac{1}{pq}, \quad G_{zz} = 1 - \frac{1}{pq} \quad \left[\because \text{C.T}_z = \frac{1}{pq} \right]$$

$$T_{xz} = -\frac{1}{pq}, \quad B_{xz} = -\frac{1}{pq}, \quad G_{xz} = -\frac{1}{pq}$$

$$T_{xy} = \frac{y_{.j}}{p} - \frac{y_{..}}{pq}, \quad B_{xy} = \frac{y_{i.}}{q} - \frac{y_{..}}{pq}, \quad G_{xy} = -\frac{y_{..}}{pq}$$

$$T_{zy} = \frac{y_{.s}}{p} - \frac{y_{..}}{pq}, \quad B_{zy} = \frac{y_{k.}}{q} - \frac{y_{..}}{pq}, \quad G_{zy} = -\frac{y_{..}}{pq}$$

$$E_{xx} = G_{xx} - B_{xx} - T_{xx} = 1 - \frac{1}{p} - \frac{1}{q} + \frac{1}{pq}, \quad E'_{xx} = E_{xx} + T_{xx} = 1 - \frac{1}{q}$$

$$E_{zz} = G_{zz} - B_{zz} - T_{zz} = 1 - \frac{1}{p} - \frac{1}{q} + \frac{1}{pq}, \quad E'_{zz} = E_{zz} + T_{zz} = 1 - \frac{1}{p}$$

$$E_{xz} = G_{xz} - B_{xz} - T_{xz} = \frac{1}{pq}, \quad E'_{xz} = E_{xz} + T_{xz} = 0$$

$$E_{xy} = G_{xy} - B_{xy} - T_{xy} = -\left(\frac{y_{.j}}{p} + \frac{y_{i.}}{q} - \frac{y_{..}}{pq} \right), \quad E'_{xy} = E_{xy} + T_{xy} = -\frac{y_{i.}}{q}$$

$$E_{zy} = G_{zy} - B_{zy} - T_{zy} = -\left(\frac{y_{k.}}{q} + \frac{y_{.s}}{p} - \frac{y_{..}}{pq} \right), \quad E'_{zy} = E_{zy} + T_{zy} = -\frac{y_{k.}}{p}$$

Thus, the estimates of γ and δ will be obtained in a similar way as it is done in the previous section. The other analytical steps are also similarly done. Here, we have

$$\hat{\beta}_j - \hat{\beta}_s = (\bar{y}_{.j} - \bar{y}_{.s}) - \hat{\gamma}(\bar{x}_{.j} - \bar{x}_{.s}) - \hat{\delta}(\bar{z}_{.j} - \bar{z}_{.s}) = (\bar{y}_{.j} - \bar{y}_{.s}) - \frac{1}{p}(\hat{\gamma} - \hat{\delta})$$

$$V(\hat{\beta}_j - \hat{\beta}_s) = \frac{2\sigma^2}{p} \left[1 + \frac{q}{pq - p - q} \right],$$

$$\hat{\beta}_j - \hat{\beta}_{j'} = (\bar{y}_{.j} - \bar{y}_{.j'}) - \frac{1}{p}\hat{\gamma}, \quad j \neq j' \neq s$$

$$V(\hat{\beta}_j - \hat{\beta}_{j'}) = \frac{\sigma^2}{p} \left[2 + \frac{q(p-1)(q-1)}{(p-1)^2(q-1)^2 - 1} \right].$$

Example 6.5 : An experiment is conducted to study the productivity of four different varieties of maize using five levels of nitrogen. All varieties of maize are cultivated using each level of nitrogen. Each level of nitrogen is used in four plots and in the plots 4 varieties of maize are randomly allocated. The design used is randomized block design. During experimentation

the production data of second variety of maize in the first block and the data of fourth variety in the third block are lost. The production data (y kg in plots of $15' \times 20'$) are shown below :

Blocks	Treatment												Total		
	M_1			M_2			M_3			M_4					
	x	y	z	x	y	z	x	y	z	x	y	z	x_i	y_i	z_i
1	0	20.0	0	1	0	0	0	28.0	0	0	28.5	0	1	76.5	0
2	0	20.2	0	0	29.2	0	0	28.5	0	0	27.0	0	0	104.9	0
3	0	20.8	0	0	30.0	0	0	26.0	0	0	0	1	0	76.8	1
4	0	20.9	0	0	31.5	0	0	27.5	0	0	27.0	0	0	106.9	0
5	0	20.0	0	0	29.5	0	0	28.0	0	0	27.5	0	0	105.0	0
Total $x_{.j}$	0			1			0			0			1		
$y_{.j}$		101.9			120.2			138.0			110.0			470.1	
$z_{.j}$			0			0			0			1			1

- (i) Analyse the data and justify the use of covariance technique in this analysis.
- (ii) Is there any difference between M_2 and M_4 ?
- (iii) Is there any difference between M_2 and M_1 ?

Solution : (i) We have $p = 5, q = 4$. Here $x = 1$ is used to indicate the missing value of second treatment and $z = 1$ is taken to indicate the missing value of fourth treatment.

$$C.T_y = \frac{G_y^2}{pq} = \frac{(470.1)^2}{20} = 11049.7005, G_{yy} = \sum \sum y_{ij}^2 - C.T_y = 1481.9295$$

$$T_{yy} = \frac{\sum y_{.j}^2}{p} - C.T_y = \frac{55975.65}{5} - 11049.7005 = 145.4295$$

$$B_{yy} = \frac{1}{q} \sum y_i^2 - C.T_y = \frac{45207.11}{4} - 11049.7005 = 252.077$$

$$E_{yy} = G_{yy} - B_{yy} - T_{yy} = 1084.423, E'_{yy} = E_{yy} + T_{yy} = 1229.8525$$

$$E_{xx} = 1 - \frac{1}{p} - \frac{1}{q} + \frac{1}{pq} = 0.60, E'_{xx} = 1 - \frac{1}{q} = 0.75, G_{xx} = 1 - \frac{1}{pq} = 0.95$$

$$E_{zz} = 1 - \frac{1}{p} - \frac{1}{q} + \frac{1}{pq} = 0.60, E'_{zz} = 1 - \frac{1}{q} = 0.75, G_{zz} = 1 - \frac{1}{pq} = 0.95$$

$$E_{xz} = \frac{1}{pq} = 0.05, E'_{xz} = 0$$

$$E_{xy} = - \left(\frac{y_i}{q} + \frac{y_{.j}}{p} - \frac{y_{.j}}{pq} \right) = -19.66, E'_{xy} = -\frac{y_i}{q} = -19.125$$

$$E_{zy} = - \left(\frac{y_k}{q} + \frac{y_{.s}}{p} - \frac{y_{.s}}{pq} \right) = -17.695, E'_{zy} = -\frac{y_k}{q} = -19.20$$

$$\hat{\gamma} = \frac{E_{zz}E_{xy} - E_{xz}E_{yz}}{E_{xx}E_{zz} - E_{xz}^2} = -30.52, \hat{\delta} = \frac{E_{xz}E_{zy} - E_{zz}E_{xy}}{E_{xx}E_{zz} - E_{xz}^2} = -26.95$$

$$\hat{\gamma} = \frac{E'_{zz}E'_{xy} - E'_{xz}E'_{yz}}{E'_{xx}E'_{zz} - E_{xz}^2} = -25.50, \hat{\delta} = \frac{E'_{xz}E'_{zy} - E'_{zz}E'_{xy}}{E'_{xx}E'_{zz} - E_{xz}^2} = -23.59$$

$$\begin{aligned}
 S_5 = SS(\text{treatment})_{\text{adjusted}} &= T_{yy} - \hat{\gamma}E'_{xy} - \hat{\delta}E'_{zy} + \hat{\gamma}E_{xy} + \hat{\delta}E_{zy} \\
 &= 145.4295 - 25.5 \times 19.125 - 23.59 \times 19.20 \\
 &\quad + 30.52 \times 19.66 + 26.95 \times 17.695 \\
 &= 281.72.
 \end{aligned}$$

$$\begin{aligned}
 S_2 = SS(\text{error})_{\text{adjusted}} &= E_{yy} - \hat{\gamma}E_{xy} - \hat{\delta}E_{zy} \\
 &= 1084.423 - 30.52 \times 19.66 - 26.95 \times 17.695 = 7.52.
 \end{aligned}$$

In a similar way the adjusted block sum of squares can be calculated, where this sum of squares is given by

$$SS(\text{Block})_{\text{adjusted}} = B_{yy} - \tilde{\gamma}E_{1xy} - \tilde{\delta}E_{1zy} + \tilde{\gamma}E_{xy} + \tilde{\delta}E_{zy}.$$

$$\text{Here } E_{1xy} = E_{xy} + B_{xy} = -19.66 + \left(\frac{y_i}{q} - \frac{y_{..}}{pq} \right) = -24.04$$

$$E_{1zy} = E_{zy} + B_{zy} = -17.695 + \left(\frac{y_k}{q} - \frac{y_{..}}{pq} \right) = -22.00$$

$$E_{1xx} = E_{xx} + B_{xx} = 0.60 + \left(\frac{1}{q} - \frac{1}{pq} \right) = 0.80$$

$$E_{1zz} = E_{zz} + B_{zz} = 0.60 + \left(\frac{1}{q} - \frac{1}{pq} \right) = 0.80$$

$$E_{1xz} = E_{xz} + B_{xz} = 0.05 - \frac{1}{pq} = 0.00$$

$$\tilde{\gamma} = \frac{E_{1zz}E_{1xy} - E_{1xz}E_{1yz}}{E_{1xx}E_{1zz} - E_{1xz}^2} = 30.05, \quad \tilde{\delta} = \frac{E_{1xx}E_{1yz} - E_{1xy}E_{1xz}}{E_{1xx}E_{1zz} - E_{1xz}^2} = -27.50$$

$$\begin{aligned}
 SS(\text{Block})_{\text{adjusted}} &= 252.077 - 30.05 \times 19.66 - 27.50 \times 22.00 \\
 &\quad + 30.52 \times 19.66 + 26.95 \times 17.695 \\
 &= 133.20.
 \end{aligned}$$

ANCOVA Table

Sources of variation	d.f.	Adjusted <i>SS</i>	<i>MS</i> (adjusted)	<i>F</i>	<i>F</i> _{0.05}
Block	4	133.20	33.30	44.28	3.48
Treatment	3	281.72	93.91	124.87	3.71
Error	10	7.52	0.752		
Total	17				

It is seen that the maize varieties differ significantly. The levels of nitrogen also vary significantly.

$$\text{We have } V(\hat{\gamma}) = \frac{E_{zz}\hat{\sigma}^2}{E_{xx}E_{zz} - E_{xz}^2}, \quad \text{where } \hat{\sigma}^2 = MS(\text{error}) = 0.752$$

$$\begin{aligned}
 &= \frac{0.60 \times 0.752}{0.60 \times 0.60 - (0.05)^2}, \quad V(\hat{\delta}) = \frac{E_{xx}\hat{\sigma}^2}{E_{xx}E_{zz} - E_{xz}^2} = \frac{0.60 \times 0.752}{0.60 \times 0.60 - (0.05)^2} \\
 &= 1.2621.
 \end{aligned}$$

$$\text{s.e}(\hat{\gamma}) = \sqrt{1.2621} = 1.1234, \quad \text{s.e}(\hat{\delta}) = 1.1234.$$

The covariance technique can be justified if the hypothesis $H_0 : \gamma = 0$ or, $H_0 : \delta = 0$ or both are rejected. The test statistic for the first hypothesis is

$$t = \frac{\hat{\gamma}}{s.e(\hat{\gamma})} = \frac{-30.52}{1.1234} = -27.17.$$

The test statistic for the second hypothesis is

$$t = \frac{\hat{\delta}}{s.e(\hat{\delta})} = \frac{-26.95}{1.1234} = -23.99.$$

Since $|t| > t_{0.05;10} = 2.228$, both the null hypothesis are rejected. The effects of concomitant variables used are significant. Hence, covariance analysis is fruitful in such analysis of data with missing values.

The analysis of data can also be performed estimating the missing values, where the estimate of second treatment is

$$x = \frac{(p-1)(q-1)(PB_1 + qT_2 - G) - (PB_3 + qT_4 - G)}{(p-1)^2(q-1)^2 - 1}.$$

Here B_1 = total of block-1 in which treatment M_2 is missing

B_3 = total of block-3 in which treatment M_4 is missing

T_2 = total of M_2 , T_4 = total of M_4 , G = grand total.

$$\begin{aligned} \therefore x &= \frac{(5-1)(4-1)(5 \times 76.5 + 4 \times 120.2 - 470.1) - (5 \times 76.8 + 4 \times 110.0 - 470.1)}{(5-1)^2(4-1)^2 - 1} \\ &= \frac{4718.4 - 353.9}{143} = 30.52. \end{aligned}$$

Similarly, the estimate of fourth treatment is

$$y = \frac{(p-1)(q-1)(pB_3 + qT_4 - G) - (pB_1 + qT_2 - G)}{(p-1)^2(q-1)^2 - 1} = 26.95.$$

Now, the corrected totals are

$$y_i : 107.02, 204.9, 103.75, 106.9, 105.0$$

$$y_j : 101.9, 150.72, 138.0, 136.95; G = 527.57.$$

$$C.T. = \frac{G^2}{pq} = \frac{(527.57)^2}{20} = 13916.5052.$$

$$SS(\text{Total}) = 272.8977, SS(\text{Block}) = 1.9855, SS(\text{Treatment}) = 263.381,$$

$$SS(\text{Error}) = 7.5312, \hat{\sigma}^2 = MS(\text{Error}) = 0.75312.$$

$$\begin{aligned} \text{Now, } V(\hat{\beta}_1 - \hat{\beta}_2) &= \frac{\hat{\sigma}^2}{p} \left[2 + \frac{q}{(p-1)(q-1)} \right] = \frac{0.75312}{5} \left[2 + \frac{4}{(5-1)(4-1)} \right] \\ &= 0.3514. \end{aligned}$$

Here $\hat{\beta}_1$ is the estimator of effect of M_1 and $\hat{\beta}_2$ is the estimator of effect of M_4 . The variance of the difference of these two estimators in case of covariance analysis is

$$\begin{aligned} V(\hat{\beta}_1 - \hat{\beta}_2) &= \frac{\hat{\sigma}^2}{p} \left[2 + \frac{q(p-1)(q-1)}{(p-1)^2(q-1)^2 - 1} \right] = \frac{0.752}{5} \left[2 + \frac{48}{143} \right] \\ &= 0.3513. \end{aligned}$$

Since in both the cases $V(\hat{\beta}_1 - \hat{\beta}_2)$ are almost same, the covariance analysis is justified. The technique is better, since it can be applied for analysis of data of randomized block design with several missing observations.

(ii) We need to test the significance of $H_0 : \beta_2 = \beta_4$.

The variance of the estimate of contrast $\hat{\beta}_2 - \hat{\beta}_4$ is

$$v(\hat{\beta}_2 - \hat{\beta}_4) = \frac{2\hat{\sigma}^2}{p} \left[1 + \frac{q}{pq - p - q} \right] = \frac{2 \times 0.752}{5} \left[1 + \frac{4}{20 - 5 - 4} \right] \\ = 0.4102.$$

$$\text{s.e}(\hat{\beta}_2 - \hat{\beta}_4) = \sqrt{0.4102} = 0.6404.$$

The adjusted estimate of $\beta_2 - \beta_4$ is

$$\hat{\beta}_2 - \hat{\beta}_4 = (\bar{y}_{.2} - \bar{y}_{.4}) - \frac{1}{p}(\hat{\gamma} - \hat{\delta}) = (24.04 - 22.00) - \frac{1}{5}(-30.52 + 26.95) = 2.754.$$

$$t = \frac{\hat{\beta}_2 - \hat{\beta}_4}{\text{s.e}(\hat{\beta}_2 - \hat{\beta}_4)} = \frac{2.754}{0.6404} = 4.30 > t_{0.05,10} = 2.228.$$

(iii) $H_0 : \beta_1 = \beta_2$, $V(\hat{\beta}_1 - \hat{\beta}_2) = 0.3513$, $\text{s.e}(\hat{\beta}_1 - \hat{\beta}_2) = 0.5927$

$$\hat{\beta}_1 - \hat{\beta}_2 = (\bar{y}_{.1} - \bar{y}_{.2}) - \frac{1}{p}\hat{\gamma} = 2.444$$

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{s.e}(\hat{\beta}_1 - \hat{\beta}_2)} = \frac{2.444}{0.5927} = 4.12 > t_{0.05,10} = 2.228.$$

Hence M_2 and M_4 differ significantly. The treatments M_1 and M_2 also differ significantly.

6.7 Covariance Analysis in Latin Square Design with One Concomitant Variable

The model assumed for this analysis is

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l + \delta(x_{ijl} - \bar{x}...) + e_{ijl}, \quad (\text{A})$$

$i = j = l = 1, 2, \dots, k$, where y_{ijl} = the result of l -th treatment in j -th column corresponding to i -th row when an experiment is conducted using $k \times k$ Latin square design, α_i, β_j and γ_l have their usual meanings, x_{ijl} = the value of the concomitant variable corresponding to y_{ijl} , δ = the regression coefficient of y on x and e_{ijl} = random error.

Assumption : e_{ijl} is normally and independently distributed with mean zero and variance σ^2 .

The normal equations to estimate the parameters in the model (A) are :

$$y_{...} = k^2\hat{\mu} + k \sum \hat{\alpha}_i + k \sum \hat{\beta}_j + k \sum \hat{\gamma}_l$$

$$y_{i..} = k\hat{\mu} + k\hat{\alpha}_i + \sum \hat{\beta}_j + \sum \hat{\gamma}_l + k\hat{\delta}(\bar{x}_{i..} - \bar{x}...)$$

$$y_{.j.} = k\hat{\mu} + \sum \hat{\alpha}_i + k\hat{\beta}_j + \sum \hat{\gamma}_l + k\hat{\delta}(\bar{x}_{.j.} - \bar{x}...)$$

$$y_{.l.} = k\hat{\mu} + \sum \hat{\alpha}_i + \sum \hat{\beta}_j + k\hat{\gamma}_l + k\hat{\delta}(\bar{x}_{.l.} - \bar{x}...).$$

$$\sum \sum \sum y_{ijl}(x_{ijl} - \bar{x}...) = k \sum \hat{\alpha}_i(\bar{x}_{i..} - \bar{x}...) + k \sum \hat{\beta}_j(\bar{x}_{.j.} - \bar{x}...) \\ + k\hat{\gamma}_l(\bar{x}_{.l.} - \bar{x}...) + \hat{\delta} \sum \sum \sum (x_{ijl} - \bar{x}...)^2.$$

$$\begin{aligned}
\text{Let } G_{yy} &= \sum \sum \sum (y_{ijl} - \bar{y}_{...})^2, \quad G_{xx} = \sum \sum \sum (x_{ijl} - \bar{x}_{...})^2, \quad R_{yy} = k \sum (\bar{y}_{i..} - \bar{y}_{...})^2 \\
G_{xy} &= \sum \sum \sum (y_{ijl} - \bar{y}_{...})(x_{ijl} - \bar{x}_{...}), \quad R_{xy} = k \sum (\bar{y}_{i..} - \bar{y}_{...})(\bar{x}_{i..} - \bar{x}_{...}) \\
R_{xx} &= k \sum (\bar{x}_{i..} - \bar{x}_{...})^2, \quad C_{yy} = k \sum (\bar{y}_{.j.} - \bar{y}_{...})^2, \quad C_{xx} = k \sum (\bar{x}_{.j.} - \bar{x}_{...})^2 \\
C_{xy} &= k \sum (\bar{y}_{.j.} - \bar{y}_{...})(\bar{x}_{.j.} - \bar{x}_{...}), \quad T_{yy} = k \sum (\bar{y}_{..l} - \bar{y}_{...})^2 \\
T_{xy} &= k \sum (\bar{x}_{..l} - \bar{x}_{...})(\bar{y}_{..l} - \bar{y}_{...}), \quad T_{xx} = k \sum (\bar{x}_{..l} - \bar{x}_{...})^2 \\
E_{xx} &= G_{xx} - R_{xx} - C_{xx} - T_{xx}, \quad E_{yy} = G_{yy} - R_{yy} - C_{yy} - T_{yy} \\
E_{xy} &= G_{xy} - R_{xy} - C_{xy} - T_{xy}, \quad E'_{xx} = E_{xx} + T_{xx}, \quad E'_{yy} = E_{yy} + T_{yy} \\
E'_{xy} &= E_{xy} + T_{xy}.
\end{aligned}$$

Now, under the restriction $\sum \hat{\alpha}_i = \sum \hat{\beta}_j = \sum \hat{\gamma}_l = 0$ and using the above notational values, we get

$$\begin{aligned}
\hat{\alpha}_i &= (\bar{y}_{i..} - \bar{y}_{...}) - \hat{\delta}(\bar{x}_{i..} - \bar{x}_{...}), \quad \hat{\beta}_j = (\bar{y}_{.j.} - \bar{y}_{...}) - \hat{\delta}(\bar{x}_{.j.} - \bar{x}_{...}) \\
\hat{\gamma}_l &= (\bar{y}_{..l} - \bar{y}_{...}) - \hat{\delta}(\bar{x}_{..l} - \bar{x}_{...}), \quad \hat{\mu} = \bar{y}_{...} \quad \text{and} \quad \hat{\delta} = \frac{E_{xy}}{E_{xx}}.
\end{aligned}$$

The sum of squares of the estimates is

$$S_1 = k^2 \bar{y}_{...}^2 + R_{yy} + C_{yy} + T_{yy} + \hat{\delta} E_{xy}.$$

This sum of squares has $(3k - 1)$ d.f. The adjusted sum of squares of error is

$$S_2 = E_{yy} - \hat{\delta} E_{xy}.$$

This S_2 has $(k - 1)(k - 2) - 1$ d.f.

Under the null hypothesis $H_0 : \gamma_l = 0$, the model stands

$$y_{ijl} = \mu + \alpha_i + \beta_j + \hat{\delta}(x_{ijl} - \bar{x}_{...}). \quad (\text{B})$$

The sum of squares of estimates for the model (B) is

$$S_3 = k^2 \bar{y}_{...}^2 + R_{yy} + C_{yy} + \hat{\delta} E'_{xy}, \quad \text{where} \quad \hat{\delta} = \frac{E'_{xy}}{E'_{xx}}.$$

This S_3 has $(2k)$ d.f. The sum of squares due to error in analysing the model (B) is

$$S_4 = E'_{yy} - \hat{\delta} E'_{xy}.$$

The d.f. of S_4 is $(k^2 - 2k)$. Hence, the sum of squares of treatment under H_0 is

$$S_5 = S_4 - S_2 = T_{yy} - \hat{\delta} E_{xy} - \hat{\delta} E'_{xy}.$$

This S_5 has $(k - 1)$ d.f. Hence, the test statistic to test the significance of treatment effect is

$$F = \frac{S_5 / (k - 1)}{S_2 / \{(k - 1)(k - 2) - 1\}}.$$

The covariance analysis is used profitably if $H_0 : \delta = 0$ is rejected. The test statistic to test the significance of this hypothesis is

$$F = \frac{\hat{\delta} E_{xy}}{S_2 / \{(k - 1)(k - 2) - 1\}}.$$

This F has 1 and $(k - 1)(k - 2) - 1$ d.f. To test the significance of the hypothesis t -test can also be applied, where the test statistic is

$$t = \frac{\hat{\delta}}{\text{s.e}(\hat{\delta})}$$

Here $\text{s.e}(\hat{\delta}) = \frac{\hat{\sigma}^2}{E_{xx}}$, $\hat{\sigma}^2 = S_2/\{(k - 1)(k - 2) - 1\}$.

The adjusted treatment mean is

$$\bar{y}_{..l}(\text{adjusted}) = \bar{y}_{..l} - \hat{\delta}(\bar{x}_{..l} - \bar{x}_{...})$$

Variance of this adjusted mean is

$$V(\bar{y}_{..l}) \text{ adjusted} = \sigma^2 \left[\frac{1}{k} + \frac{(\bar{x}_{..l} - \bar{x}_{...})^2}{E_{xx}} \right]$$

The variance of the estimated contrast $\sum d_l \hat{\eta}_l$ is

$$V(\sum d_l \hat{\eta}_l) = \sigma^2 \left[\frac{\sum d_l^2}{k} + \frac{(\sum d_l \bar{x}_{..l})^2}{E_{xx}} \right]$$

The average variance of $(\bar{y}_{..l} - \bar{y}_{..l'})$ adjusted is

$$V(\bar{y}_{..l} - \bar{y}_{..l'}) \text{ adjusted} = \frac{2\sigma^2}{k} \left[1 + \frac{T_{xx}/(k - 1)}{E_{xx}} \right]$$

This variance is used to compare the pairs of treatment.

Example 6.6 : To study the impact of dry food on milk production of cows an experiment is conducted using 16 cows, where cows are grouped according to lactation period and body weights. Cows are divided into 4 lactation periods and into 4 groups according to body weight. The cows are fed 4 types of dry food. During the experiment one day milk productions (y kg) are recorded. The experiment is conducted through a 4×4 Latin square design. The amounts of food per cow per day (x kg) are different.

Lactation period	Body weight								Total	
	w_1		w_2		w_3		w_4		$x_{i..}$	$y_{i..}$
	x	y	x	y	x	y	x	y		
L_1	A10.0	32.0	B12.5	35.6	C 8.0	30.5	D15.0	36.0	45.5	134.1
L_2	B11.5	34.0	A12.0	33.5	D10.0	34.0	C10.0	32.0	43.5	133.5
L_3	C12.0	35.5	D11.0	36.5	A12.0	33.5	B15.0	37.0	50.0	142.5
L_4	D 9.5	25.0	C 8.0	28.5	B10.0	35.0	A11.5	34.0	39.0	122.5
Total $y_{.j}$	126.5		134.1		133.0		139.0		532.6	
$x_{.j}$	43.0		43.5		40.0		51.5		178.0	

Analyse the data and comment on the performance of food.

Solution : $G_y = 532.6$, $G_x = 178.0$, $C.T_x = \frac{G_x^2}{k^2} = 1980.25$

$$C.T_y = \frac{G_y^2}{k^2} = 17728.92, G_{yy} = \sum \sum \sum y_{ijl}^2 - C.T_y = 148.94$$

$$G_{xx} = \sum \sum \sum x_{ijl}^2 - C.T_x = 61.75, G_{xy} = \sum \sum x_{ijl}y_{ijl} - C.T_{xy} = 67.825$$

$$C.T_{xy} = \frac{G_x G_y}{k^2} = 5925.175; R_{yy} = \frac{1}{k} \sum y_{i..}^2 - C.T_y = 50.47$$

$$R_{xx} = \frac{1}{k} \sum x_{i..}^2 - C.T_x = 15.625, R_{xy} = \frac{1}{k} \sum x_i y_i - C.T_{xy} = 27.65$$

$$y_{..l} : 133.0, 141.6, 126.5, 131.5$$

$$x_{..l} : 45.5, 49.0, 38.0, 45.5$$

$$C_{yy} = \frac{1}{k} \sum y_{.j}^2 - C.T_y = 19.845, C_{xy} = \frac{1}{k} \sum x_{.j} y_{.j} - C.T_{xy} = 12.66$$

$$C_{xx} = \frac{1}{k} \sum x_{.j}^2 - C.T_x = 18.125, T_{yy} = \frac{1}{k} \sum y_{.l}^2 - C.T_y = 29.595$$

$$T_{xx} = \frac{1}{k} \sum x_{.l}^2 - C.T_x = 16.125, T_{xy} = \frac{1}{k} \sum x_{.l} y_{.l} - C.T_{xy} = 19.862$$

$$E_{xx} = G_{xx} - R_{xx} - C_{xx} - T_{xx} = 11.875, E'_{xx} = E_{xx} + T_{xx} = 28.00$$

$$E_{yy} = G_{yy} - R_{yy} - C_{yy} - T_{yy} = 49.03, E'_{yy} = E_{yy} + T_{yy} = 78.625$$

$$E_{xy} = G_{xy} - R_{xy} - C_{xy} - T_{xy} = 7.653, E'_{xy} = E_{xy} + T_{xy} = 27.515$$

$$\hat{\delta} = \frac{E_{xy}}{E_{xx}} = 0.644, \hat{\delta} = \frac{E'_{xy}}{E'_{xx}} = 0.983$$

$$S_2 = E_{yy} - \hat{\delta} E_{xy} = 49.03 - 0.644 \times 7.653 = 44.101$$

$$S_4 = E'_{yy} - \hat{\delta} E'_{xy} = 78.625 - 0.983 \times 27.515 = 51.578.$$

Hence, the sum of squares due to treatment under $H_0 : \gamma_l = 0$ is

$$S_5 = S_4 - S_2 = 51.578 - 44.101 = 7.477.$$

The test statistic to test the significance of $H_0 : \gamma_l = 0$ is

$$F = \frac{S_5/(k-1)}{S_2/((k-1)(k-2)-1)} = \frac{7.477/3}{44.101/5} = 0.28.$$

Since $F < F_{0.05;3,5} = 5.41$, H_0 is accepted. The types of dry food do not differ significantly.

ANCOVA Table

Sources of variation	SS(x)	SS(y)	SP(xy)	adjusted		F
				SS	d.f.	
Rows	15.625	50.47	27.65	—	—	0.28
Columns	18.125	19.845	12.66	—	—	
Treatments	16.125	29.595	19.862	7.477	3	
Errors	11.875	49.03	7.653	44.101	5	

The adjusted treatment means are

$$\bar{y}_{..l} \text{ (adjusted)} = \bar{y}_{..l} - \hat{\delta}(\bar{x}_{..l} - \bar{x}_{...}) : 33.09, 34.67, 32.67, 32.71.$$

The average variance to compare the means in pairs is

$$\begin{aligned} v(\bar{y}_{..l} - \bar{y}_{..l'}) \text{ adjusted} &= \frac{2\hat{\sigma}^2}{k} \left[1 + \frac{T_{xx}/(k-1)}{E_{xx}} \right], \hat{\sigma}^2 = \frac{S_2}{(k-1)(k-2)-1} = 8.8202 \\ &= \frac{2 \times 8.8202}{4} \left[1 + \frac{16.125/3}{11.875} \right] = 6.41. \end{aligned}$$

Chapter 7

Variance Component Analysis

7.1 Introduction

The data collected from controlled experiment are represented by experimental design model. Model is of three types, viz., (a) Fixed effect model, (b) Mixed effect model, and (c) Random effect model. The analyses of random effect model and mixed effect model are known as variance component analysis.

Let us consider a model for the data of randomized block design, where the model is

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

$$i = 1, 2, \dots, p; j = 1, 2, \dots, q.$$

Here y_{ij} = the result of j -th treatment in i -th block, μ = general mean, α_i = effect of i -th block, β_j = effect of j -th treatment, e_{ij} = random error.

In the above experiment q treatments are randomly allocated to q plots of a block. The p blocks used in the experiment are assumed to be randomly selected from a population of blocks and q treatments are also assumed to be randomly selected from a population of treatments. For example, let us consider that an agricultural research station discovered 10 varieties of high yielding rice and the researcher in the station wants to verify the productivity of 5 varieties of rice. These 5 varieties can be selected randomly from 10 varieties. In such a case, the effect of the variety is random. Again, consider that, in the station there are 50 cropping areas. The selected 5 varieties of rice can be cultivated in plots of randomly selected 10 cropping areas. If each cropping area is considered a block, the block effect is random. In such a case, the model assumed for the analysis is random effect model. If the randomly selected varieties of rice are cultivated in all possible cropping areas, where an area is considered a block, the block effect is fixed and in that case the model for the data is mixed effect model. The general mean μ is always considered constant.

Since the block effect (α_i), treatment effect (β_j), except the random component (e_{ij}), are assumed random [e_{ij} is always random], they have their distribution. Let us assume that (i) $\alpha_i \sim \text{NID}(0, \sigma_\alpha^2)$, (ii) $\beta_j \sim \text{NID}(0, \sigma_\beta^2)$, (iii) $e_{ij} \sim \text{NID}(0, \sigma^2)$, and (iv) all random variables are mutually independent. Therefore, unlike the analysis of fixed effect model, the analysis of random effect model involves the estimation of variance components ($\sigma_\alpha^2, \sigma_\beta^2$ and σ^2) and tests the significance of these variance components. As block effect and treatment effect are random variable, we cannot estimate these effects or their contrast.

In analyzing fixed effect model [α_i and β_j are parameters], we usually test the significance of the hypothesis :

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_p \quad \text{or} \quad H_0 : \alpha_i = 0 \quad \text{for all } i = 1, 2, \dots, p$$

$$\text{and} \quad H_0 : \beta_1 = \beta_2 = \dots = \beta_q \quad \text{or} \quad H_0 : \beta_j = 0 \quad \text{for all } j = 1, 2, \dots, q.$$

First hypothesis indicates that the block effects are homogeneous and the second hypothesis indicates that the treatment effects are homogeneous. If block effects are same, the variance

of block effects is zero. Hence, when block effect is considered random variable, the equivalent hypothesis of similar block effects is $H_0 : \sigma_\alpha^2 = 0$ against $H_A : \sigma_\alpha^2 > 0$. In a similar way it can be mentioned that the homogeneity of treatment effects leads to enunciate the hypothesis $H_0 : \sigma_\beta^2 = 0$ against $H_A : \sigma_\beta^2 > 0$. Since analysis of random effect and mixed effect models involves the estimation and test of variance components, the analysis is known as variance component analysis.

7.2 Assumptions in Variance Component Analysis

Let the model for variance component analysis be

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl},$$

$i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, r$.

The assumptions to analyse the above model are :

Assumption-1 : The random variable $\alpha_i, \beta_j, (\alpha\beta)_{ij}$ and e_{ijl} are independently distributed each with mean zero and variances $\sigma_\alpha^2, \sigma_\beta^2, \sigma_{\alpha\beta}^2$ and σ^2 , respectively.

Assumption-2 : (i) $\alpha_i \sim \text{NID}(0, \sigma_\alpha^2)$, (ii) $\beta_j \sim \text{NID}(0, \sigma_\beta^2)$ (iii) $(\alpha\beta)_{ij} \sim \text{NID}(0, \sigma_{\alpha\beta}^2)$. (iv) $e_{ijl} \sim \text{NID}(0, \sigma^2)$, (v) all random variables are mutually independent.

Assumption-3 : (a) Let α_i be fixed effect and it is restricted that $\sum \alpha_i = 0$. But $\beta_j \sim \text{NID}(0, \sigma_\beta^2)$. Again, under the restriction $\sum_i (\alpha\beta)_{ij} = 0$. $(\alpha\beta)_{ij}$ is :

$$N\left(0, \frac{p-1}{p}\sigma_{\alpha\beta}^2\right); \text{Cov}[(\alpha\beta)_{ij}, (\alpha\beta)_{i'j}] = -\frac{1}{p}\sigma_{\alpha\beta}^2 \text{ (} i \neq i' \text{)}. \text{ Also, } \beta_j \text{ and } (\alpha\beta)_{ij} \text{ are independent.}$$

(b) Under the restriction $\sum \alpha_i = 0; \beta_j \sim \text{NID}(0, \sigma_\beta^2); (\alpha\beta)_{ij} \sim \text{NID}(0, \sigma_{\alpha\beta}^2)$. β_j and $(\alpha\beta)_{ij}$ are independent.

The first assumption is that the model is random effect model but the distributions of the random variables are not specified. Assumption-2 indicates that the model is random effect one and the random variables follow normal distribution. Assumption-3(a) indicates that the model is mixed effect model but one of the random variables is not independently distributed. The assumption-3(b) is also related to mixed effect model but all the random variables are independently distributed.

7.3 Method of Variance Component Analysis

It has already been mentioned that the variance component analysis involves the estimation and test of variance components. The estimation of variance components is done by (a) Method of least squares (Analysis of Variance Technique), (b) Method of maximum likelihood. The method depends on assumption of the random variable. For example, if assumption-1 discussed above is considered, the variance component analysis is to be performed using method of least squares. But method of maximum likelihood can be used if Assumption-2 is considered, since probability density function and hence, likelihood function for the variable can be found out. However, analysis of variance technique can be applied irrespective of the assumption. We shall only discuss the method of analysis of variance technique to estimate the variance component.

The analysis of variance technique leads us to find the $E(MS)$, where MS is the mean square of effects and/or interactions. The expected value of mean squares or expected value of functions of mean squares equals the variance component. Solution of the equations gives the estimate of variance component. Let us explain the method in analysing the model for completely randomized design.

The model assumed is

$$y_{ij} = \mu + \alpha_i + e_{ij}; \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q.$$

Assumption : (i) $\alpha_i \sim \text{NID}(0, \sigma_\alpha^2)$, (ii) $e_{ij} \sim \text{NID}(0, \sigma^2)$ (iii) α_i and e_{ij} are independently distributed.

According to analysis of variance technique the total sum of squares of the observation is partitioned as follows :

$$\sum_i^p \sum_j^q (y_{ij} - \bar{y}_{..})^2 = q \sum (\bar{y}_i - \bar{y}_{..})^2 + \sum \sum (y_{ij} - \bar{y}_i)^2 = S_1 + S_2.$$

Now we have $E(S_1) = Eq \sum (\bar{y}_i - \bar{y}_{..})^2 = (p - 1)[\sigma^2 + q\sigma_\alpha^2]$

$$E\left(\frac{S_1}{p - 1}\right) = \sigma^2 + q\sigma_\alpha^2.$$

Again, $E(S_2) = p(q - 1)\sigma^2$ or, $E\left[\frac{S_2}{p(q - 1)}\right] = \sigma^2.$

Here $s_2 = \frac{S_2}{p(q - 1)}$. Therefore, $\hat{\sigma}^2 = s_2 = MS$ (error).

Also we have $E(s_1 - s_2) = q\sigma_\alpha^2$, where $s_1 = S_1/(p - 1) = MS$ (treatment).

$$\therefore \hat{\sigma}_\alpha^2 = \frac{1}{q}(s_1 - s_2).$$

The estimates $\hat{\sigma}^2$ and $\hat{\sigma}_\alpha^2$ are unbiased and

$$V(\hat{\sigma}^2) = \frac{2\sigma^4}{p(q - 1)} \quad \text{and} \quad V(\hat{\sigma}_\alpha^2) = \frac{2}{q^2} \left[\frac{(\sigma^2 + q\sigma_\alpha^2)^2}{p - 1} + \frac{\sigma^4}{p(q - 1)} \right].$$

Under the assumption of α_i and e_{ij} we know that s_2 is distributed as $\chi^2\sigma^2$ with $p(q - 1)$ d.f. and s_1 is distributed as $\chi^2(\sigma^2 + q\sigma_\alpha^2)$ with $(p - 1)$ d.f. Hence, we get the above variances of $\hat{\sigma}^2$ and $\hat{\sigma}_\alpha^2$. The estimates $\hat{\sigma}^2$ and $\hat{\sigma}_\alpha^2$ are unbiased and their variances are minimum [Graybill (1961)]. Searle (1971a, 1971b) has shown that the unbiased estimates of these variances are :

$$v(\hat{\sigma}_\alpha^2) = \frac{2\hat{\sigma}^4}{p(q - 1) + 2} \quad \text{and} \quad v(\hat{\sigma}^2) = \frac{2}{q^2} \left[\frac{(\hat{\sigma}^2 + q\hat{\sigma}_\alpha^2)^2}{p + 1} + \frac{\hat{\sigma}^4}{p(q - 1) + 2} \right].$$

The main objective of this analysis is to test the significance of

$$H_0 : \sigma_\alpha^2 = 0 \quad \text{against} \quad H_A : \sigma_\alpha^2 > 0.$$

Since $E(s_1) = \sigma^2 + q\sigma_\alpha^2$ and $E(s_2) = \sigma^2$, we have under H_0 , $E(s_1) = E(s_2)$. Hence, the test statistic is

$$F = s_1/s_2.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	$E(MS)$	F
Treatment	$p - 1$	S_1	s_1	$\sigma^2 + q\sigma_\alpha^2$	s_1/s_2
Error	$p(q - 1)$	S_2	s_2	σ^2	
Total	$pq - 1$				

7.4 Variance Component Analysis in Two-Way Classification

Let A and B be two factors having levels p and q respectively. Consider that each level combination is replicated r times and the result corresponding to l -th replication of j -th level of B in presence of i -th level of A is y_{ijl} ($i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, r$). The model for y_{ijl} observation is

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl}, \tag{A}$$

where μ = general mean, α_i = effect of i -th level of A , β_j = effect of j -th level of B , $(\alpha\beta)_{ij}$ = interaction of i -th level of A with j -th level of B and e_{ijl} = random component.

Assumption : (i) $\alpha_i \sim \text{NID}(0, \sigma_\alpha^2)$, (ii) $\beta_j \sim \text{NID}(0, \sigma_\beta^2)$ (iii) $(\alpha\beta)_{ij} \sim \text{NID}(0, \sigma_{\alpha\beta}^2)$, (iv) $e_{ijl} \sim \text{NID}(0, \sigma^2)$, and (v) all random variables are mutually independent.

The total sum of squares of the observations is partitioned into component sum of squares as follows :

$$\begin{aligned} \sum_i^p \sum_j^q \sum_l^r (y_{ijl} - \bar{y} \dots)^2 &= qr \sum_i (\bar{y}_{i..} - \bar{y} \dots)^2 + pr \sum_j (\bar{y}_{.j.} - \bar{y} \dots)^2 \\ &\quad + r \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y} \dots)^2 + \sum_i \sum_j \sum_l (y_{ijl} - \bar{y}_{ij.})^2 \\ &= S_1 + S_2 + S_3 + S_4. \end{aligned}$$

Here $E(S_4) = E \sum_i \sum_j \sum_l (y_{ijl} - \bar{y}_{ij.})^2 = E \sum_i \sum_j \sum_l (e_{ijl} - \bar{e}_{ij.})^2$

$$\begin{aligned} &= \sum_j \sum_l \sum_i E(e_{ijl}^2) + r \sum_i \sum_j E(\bar{e}_{ij.}^2) - 2E \sum_i \sum_j \sum_l e_{ijl} \bar{e}_{ij.} \\ &= pqr\sigma^2 + pqr \frac{\sigma^2}{r} - 2r \sum_i \sum_j E\bar{e}_{ij.}^2 \\ &= pqr\sigma^2 + pq\sigma^2 - 2pq\sigma^2 = pq(r - 1)\sigma^2. \end{aligned}$$

Let $s_4 = \frac{S_4}{pq(r - 1)} = \text{M.S (error)}$. Then $E(s_4) = \sigma^2$.

In a similar way, we can show that

$$\begin{aligned} E(S_1) &= (\sigma^2 + r\sigma_{\alpha\beta}^2 + qr\sigma_\alpha^2)(p - 1) \\ E(S_2) &= (\sigma^2 + r\sigma_{\alpha\beta}^2 + pr\sigma_\beta^2)(q - 1) \\ E(S_3) &= (\sigma^2 + r\sigma_{\alpha\beta}^2)(p - 1)(q - 1). \end{aligned}$$

These analytical results are shown in the following analysis of variance table :

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{\text{d.f.}}$	$E(MS)$
A	$p - 1$	S_1	s_1	$\sigma^2 + r\sigma_{\alpha\beta}^2 + qr\sigma_\alpha^2$
B	$q - 1$	S_2	s_2	$\sigma^2 + r\sigma_{\alpha\beta}^2 + pr\sigma_\beta^2$
AB	$(p - 1)(q - 1)$	S_3	s_3	$\sigma^2 + r\sigma_{\alpha\beta}^2$
Error	$pq(r - 1)$	S_4	s_4	σ^2
Total	$pqr - 1$			

The objective of the analysis is to test the significance of the hypotheses :

- (i) $H_0 : \sigma_\alpha^2 = 0$, against $H_A : \sigma_\alpha^2 > 0$
- (ii) $H_0 : \sigma_\beta^2 = 0$, against $H_A : \sigma_\beta^2 > 0$
- (iii) $H_0 : \sigma_{\alpha\beta}^2 = 0$, against $H_A : \sigma_{\alpha\beta}^2 > 0$.

Further objective is to estimate these variance components.

The test statistic for hypothesis (iii) is $F_3 = s_3/s_4$. This F follows variance ratio distribution with $(p-1)(q-1)$ and $pq(r-1)$ d.f. If $F_3 \geq F_{0.05:(p-1)(q-1), pq(r-1)}$, H_0 is rejected. If H_0 (iii) is rejected, the test statistic for H_0 (ii) is $F_2 = s_2/s_3$. This F_2 follows variance ratio distribution with $(q-1)$ and $(p-1)(q-1)$ d.f. If H_0 (iii) is not rejected, s_2 is to be compared with the pooled value of s_3 and s_4 , where the pooled value is $(S_3 + S_4)/(pqr - p - q + 1)$. Thus, the test statistic is :

$$F_2 = \frac{s_2}{(S_3 + S_4)/(pqr - p - q + 1)}.$$

This F_2 follows variance ratio distribution with $(q-1)$ and $(pqr - p - q + 1)$ d.f. The test statistic for H_0 (i) is $F_1 = s_1/s_3$ when H_0 (iii) is rejected. If H_0 (iii) is not rejected, then

$$F_1 = \frac{s_1}{(S_3 + S_4)/(pqr - p - q + 1)}.$$

The conclusion at each step is drawn as usual.

It is observed that $E(s_4) = \sigma^2$. Hence, the estimator of σ^2 is s_4 . Again, it is observed that

$$E(s_3 - s_4) = r\sigma_{\alpha\beta}^2, \quad \therefore \hat{\sigma}_{\alpha\beta}^2 = \frac{1}{r}(s_3 - s_4).$$

$$\text{Similarly, } E(s_2 - s_3) = pr\sigma_\beta^2, \quad \therefore \hat{\sigma}_\beta^2 = \frac{1}{pr}(s_2 - s_3).$$

$$E(s_1 - s_3) = qr\sigma_\alpha^2, \quad \therefore \hat{\sigma}_\alpha^2 = \frac{1}{qr}(s_1 - s_3).$$

By assumption S_4 is distributed as $\chi^2\sigma^2$ with $pq(r-1)$ d.f. Hence, $V(S_4) = V(\chi^2\sigma^2) = \sigma^4V(\chi^2) = 2pq(r-1)\sigma^4$.

$$\text{Hence, } V(\hat{\sigma}^2) = V(s_4) = V\left[\frac{S_4}{pq(r-1)}\right] = \frac{2\sigma^4}{pq(r-1)}.$$

In a similar way, we find

$$V(\hat{\sigma}_\alpha^2) = \frac{2}{(qr)^2} \left[\frac{(\sigma^2 + r\sigma_{\alpha\beta}^2 + qr\sigma_\alpha^2)^2}{p-1} + \frac{(\sigma^2 + r\sigma_{\alpha\beta}^2)^2}{(p-1)(q-1)} \right]$$

$$V(\hat{\sigma}_\beta^2) = \frac{2}{(pr)^2} \left[\frac{(\sigma^2 + r\sigma_{\alpha\beta}^2 + pr\sigma_\beta^2)^2}{q-1} + \frac{(\sigma^2 + r\sigma_{\alpha\beta}^2)^2}{(p-1)(q-1)} \right]$$

$$V(\hat{\sigma}_{\alpha\beta}^2) = \frac{2}{r^2} \left[\frac{(\sigma^2 + r\sigma_{\alpha\beta}^2)^2}{(p-1)(q-1)} + \frac{\sigma^4}{pq(r-1)} \right].$$

The unbiased estimators of these variances are :

$$v(\hat{\sigma}^2) = \frac{2\hat{\sigma}^4}{pq(r-1) + 2}, \quad v(\hat{\sigma}_\alpha^2) = \frac{2}{(qr)^2} \left[\frac{(\hat{\sigma}^2 + r\hat{\sigma}_{\alpha\beta}^2 + qr\hat{\sigma}_\alpha^2)^2}{p+1} + \frac{(\hat{\sigma}^2 + r\hat{\sigma}_{\alpha\beta}^2)^2}{(p-1)(q-1) + 2} \right]$$

$$v(\hat{\sigma}_{\beta}^2) = \frac{2}{(pr)^2} \left[\frac{(\hat{\sigma}^2 + r\hat{\sigma}_{\alpha\beta}^2 + pr\hat{\sigma}_{\beta}^2)^2}{q+1} + \frac{(\hat{\sigma}^2 + r\hat{\sigma}_{\alpha\beta}^2)^2}{(p-1)(q-1)+2} \right]$$

$$v(\hat{\sigma}_{\alpha\beta}^2) = \frac{2}{r^2} \left[\frac{(\hat{\sigma}^2 + r\hat{\sigma}_{\alpha\beta}^2)^2}{(p-1)(q-1)+2} + \frac{\hat{\sigma}^4}{pq(r-1)+2} \right].$$

The above analysis is done assuming random effect model. Let us consider that the effect α_i is non-random and the restriction for α_i is $\sum \alpha_i = 0$. In that case, the model (A) is a mixed effect model. To analyse the model, α_i is to be estimated in a similar way as it is done in analysing fixed effect model. The sum of squares of different effects and interaction are found out as usual. However,

$$E[MS(\hat{\alpha}_i)] = E(s_1) = \sigma^2 + r\sigma_{\alpha\beta}^2 + \frac{qr}{p-1} \sum \alpha_i^2.$$

The other steps of the analysis remain same. However, the variance of the primary contrast of fixed effects is

$$\frac{2(MS \text{ used as denominator in testing the significance of fixed effect})}{\text{No. of observations in calculating means related to fixed effect}}.$$

7.5 Steps in Calculating $E(MS)$ for Variance Component Analysis

It is observed that the important steps in analysing mixed effect model or random effect model are to find out the $E(MS)$. Montgomery (1984) has proposed a simple method to find the value of $E(MS)$. The method is based on several steps. The steps are discussed below :

Step-1 : Let the model be

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl},$$

$i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, r.$

The model can be written as

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{(ij)l}.$$

Here i, j and l are suffixes. These suffixes can be classified as (a) alive, (b) dead and (c) absent. The suffix which is under bracket is considered as dead, otherwise the suffix is alive. In the model the suffixes i and j in $e_{(ij)l}$ are in bracket. These suffixes are dead. But i and j in $(\alpha\beta)_{ij}$ are alive. Again, l in $(\alpha\beta)_{ij}$ is absent.

Step-2 : The d.f. of any component in the model is the product of levels of dead suffixes and level minus one of alive suffixes present in a component. Thus, the d.f. of $(\alpha\beta)_{ij}$ is $(p-1)(q-1)$. Again, the d.f. of $e_{(ij)l}$ is $pq(r-1)$, since i and j in $e_{(ij)l}$ are dead suffixes but l is alive suffix.

Step-3 : In any model the components, except μ , are either fixed effect or random variable. Each random variable has a variance component and each fixed effect is denoted by a symbol. In finding $E(MS)$ we usually write variance component for the variance of the random variable and for fixed effect we write sum of squares of the effect divided by the respective d.f. as variance of that effect. For example, if α_i in a model ($i = 1, 2, \dots, p$) is fixed effect, then its variance is taken as $\sum \frac{\alpha_i^2}{(p-1)}$.

Step-4 : The coefficients of each variance component is multiplied by some numbers, where numbers are the levels of suffixes which are absent in a component. For example, the variance

component of $(\alpha\beta)_{ij}$ is multiplied by r , where r is the level of l and this l is absent in $(\alpha\beta)_{ij}$. The coefficients for different variance components are found out by preparing a table. The number of columns in the table is equal to the number of suffixes in the model. The number of rows in the table is equal to the number of components, except μ . Thus, there will be row for each component, except μ , in the model. The suffixes are written in a row outside the main table. There is another row outside the table to write the level of each suffix. Moreover, the suffixes are indicated by F or R if an effect is fixed or random, respectively. The body of the table is then filled up using the following principle :

(a) The value corresponding to a row and a column is 1 if the dead suffix in the row coincides with the suffix in column.

(b) The value corresponding to a row and a column is zero if the suffix in the row coincides with the suffix in column and if that suffix in column is for fixed effect F . But the value is 1 if the suffix in column is for random effect R .

(c) The elements in other rows and columns are filled up with the value of levels written in the column. All values in columns corresponding to error term are 1.

(d) To find $E(MS)$ of any component of a model, the elements in row corresponding to that component are filled up for all columns related to alive suffix of that component. The variance component related to a component in row is multiplied by the product of the visible element in that row. Therefore, each variance component has a coefficient. The expected value of $MS[E(MS)]$ of a component is the sum of variance components with corresponding components with corresponding coefficient. To take the sum of components we need to consider those variance components which are related to the suffix of the component for which $E(MS)$ is to be calculated. In calculating the product of visible numbers, the elements except the element corresponding to the suffix(es) of the component is considered. The $E(MS)$ for error is σ^2 .

Let us explain the method in calculating the $E(MS)$ of the model.

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{(ij)l}$$

Let us assume that the model is a fixed effect model. Here $i = 1, 2, \dots, p$; $j = 1, 2, \dots, q$; $l = 1, 2, \dots, r$.

Table to find $E(MS)$

Component	F	F	R	d.f.	$E(MS)$
	p	q	r		
	i	j	l		
α_i	0	q	r	$p - 1$	$\sigma^2 + \frac{qr}{p-1} \sum \alpha_i^2$
β_j	p	0	r	$q - 1$	$\sigma^2 + \frac{pr}{q-1} \sum \beta_j^2$
$(\alpha\beta)_{ij}$	0	0	r	$(p - 1)(q - 1)$	$\sigma^2 + \frac{r}{(p-1)(q-1)} \sum \sum (\alpha\beta)_{ij}^2$
$e_{(ij)l}$	1	1	1	$pq(r - 1)$	σ^2

Since l suffix is used for replication, it is always random. The d.f. is calculated, for example, for $(\alpha\beta)_{ij}$ as $(p - 1)(q - 1)$, since i and j both are alive suffix and the level of i is p , the level of j is q . So d.f. corresponding to alive suffixes ij is $(p - 1)(q - 1)$. The first element in the table is zero, since first row is related to suffix i and first column is related to suffix i with F . The suffixes j and l in second and third columns do not coincide with suffix i in first row and hence, the elements corresponding to second and third columns are q and r , respectively. Similar argument is true for other elements.

The $E(MS)$ for error is σ^2 . To calculate $E(MS)$ for component related to $(\alpha\beta)_{ij}$, we need to add variance components related to $(\alpha\beta)_{ij}$ and $e_{(ij)l}$, since the suffix ij is present in these two components. The coefficient related to variance component σ^2 related to $e_{(ij)l}$ is 1 since visible elements in the table is 1 corresponding to suffix l [except the suffixes i and j] in the columns. The coefficient related to variance component of $(\alpha\beta)_{ij}$ is r since the visible element in the row related to $(\alpha\beta)_{ij}$ is r , except the elements corresponding to i and j in the column. Therefore,

$$E[MS(\text{for } \alpha\beta_{ij})] = \sigma^2 + \frac{r}{(p-1)(q-1)} \sum \sum (\alpha\beta)_{ij}^2.$$

Here $(\alpha\beta)_{ij}$ is fixed effect and its variance component is

$$\frac{1}{(p-1)(q-1)} \sum \sum (\alpha\beta)_{ij}^2 \quad [\because \sum \sum (\alpha\beta)_{ij} = 0].$$

The $E(MS)$ corresponding to β_j will be the sum of variance components related to $e_{(ij)l}$, $(\alpha\beta)_{ij}$ and β_j , since the suffix j is present in all these components. The coefficient related to σ^2 is 1, since the product of elements in columns corresponding to i and l (except column corresponding to j) are 1 and 1. The coefficient related to variance component of $(\alpha\beta)_{ij}$ is zero, since the product of the elements in the row corresponding to $(\alpha\beta)_{ij}$ are 0 and r [except the element corresponding to column indicated by j]. Hence, the variance component related to $(\alpha\beta)_{ij}$ is not added in calculating $E(MS)$ related to β_j . The coefficient related to variance component of β_j is pr , since the visible numbers in the rows related to β_j are p and r (except the element corresponding to column indicated by j). Finally,

$$E[MS(\text{for } \beta_j)] = \sigma^2 + \frac{pr}{q-1} \sum \beta_j^2.$$

In a similar way, we get

$$E[MS(\text{for } \alpha_i)] = \sigma^2 + \frac{qr}{p-1} \sum \alpha_i^2.$$

Let us now consider the calculation of $E(MS)$ assuming random effect model. The values of $E(MS)$ are as given in the table below :

Table to find $E(MS)$

Component	R	R	R	d.f.	$E(MS)$
	i	j	l		
	p	q	r		
α_i	1	q	r	$p-1$	$\sigma^2 + r\sigma_{\alpha\beta}^2 + qr\sigma_{\alpha}^2$
β_j	p	1	r	$q-1$	$\sigma^2 + r\sigma_{\alpha\beta}^2 + pr\sigma_{\beta}^2$
$(\alpha\beta)_{ij}$	1	1	r	$(p-1)(q-1)$	$\sigma^2 + r\sigma_{\alpha\beta}^2$
$e_{(ij)l}$	1	1	1	$pq(r-1)$	σ^2

The argument to add various variance components is similar as it is made to calculate $E(MS)$ for fixed effect model. For example, let us consider the calculation of $E(MS)$ corresponding to component α_i . The suffix 'i' is present in α_i , $(\alpha\beta)_{ij}$ and $e_{(ij)l}$ and hence, to calculate $E(MS)$ for α_i we need to add the variance components σ_{α}^2 , $\sigma_{\alpha\beta}^2$ and σ^2 related to α_i , $(\alpha\beta)_{ij}$ and e_{ijl} , respectively. The coefficient of σ^2 is 1, since the product of visible numbers corresponding to row e_{ijl} are 1 and 1 (except the number corresponding to i in the column). The coefficient of

$\sigma_{\alpha\beta}^2$ is r , since the visible numbers corresponding to row $(\alpha\beta)_{ij}$ are 1 and r . The coefficient of σ_{α}^2 is qr , since the visible members corresponding to row α_i are q and r except the number corresponding to column head by i . Therefore, we have

$$E[MS(\text{for } \alpha_i)] = \sigma^2 + r\sigma_{\alpha\beta}^2 + qr\sigma_{\alpha}^2.$$

Similarly, we can find $E(MS)$ for all components assuming mixed effect model, where it is assumed that α_i is fixed with a restriction $\sum \alpha_i = 0$. It is further assumed that

$$\sum_i (\alpha\beta)_{ij} = 0 \text{ and } (\alpha\beta)_{ij} \sim N\left(0, \frac{p-1}{p}\sigma_{\alpha\beta}^2\right); \text{Cov}[(\alpha\beta)_{ij}, (\alpha\beta)_{i'j}] = -\frac{1}{p}\sigma_{\alpha\beta}^2 \text{ (} i \neq j\text{)}.$$

Also, it is assumed that β_j and $(\alpha\beta)_{ij}$ are independent. The $E(MS)$ of different components are shown below :

Table to find $E(MS)$

Component	R	R	R	d.f	$E(MS)$
	i	j	l		
	p	q	r		
α_i	0	q	r	$p-1$	$\sigma^2 + r\sigma_{\alpha\beta}^2 + \frac{qr}{p-1} \sum \alpha_i^2$
β_j	p	1	r	$q-1$	$\sigma^2 + pr\sigma_{\beta}^2$
$(\alpha\beta)_{ij}$	0	1	r	$(p-1)(q-1)$	$\sigma^2 + r\sigma_{\alpha\beta}^2$
$e_{(ij)l}$	1	1	1	$pq(r-1)$	σ^2

7.6 Demerits of Analysis of Variance Technique in Variance Component Analysis

In section 7.4, it is observed that $\hat{\sigma}_{\alpha\beta}^2 = \frac{1}{r}(s_3 - s_4)$. But this estimate may be negative, since it is not sure that s_3 is greater than s_4 . In practice, s_3 may be less than s_4 and hence, the estimate $\hat{\sigma}_{\alpha\beta}$ may be negative. This problem may arise for any estimate of variance component. But variance should not be negative. This is the problem of analysis of variance technique. However, the technique is widely used as a method of variance component analysis and any negative estimate is considered as insignificant.

Example 7.1 : The management of a poultry farm collected 10 varieties of dry concentrate from the market for its chicks. Initially the management planned to give 5 varieties of dry concentrate to the chicks and 5 varieties are randomly selected. The concentrates were given to chicks of 4 different ages. Each concentrate was continued up to the age 45 days of chicks and then the weights of chicks were recorded. The weights (in kg) are given below :

Weight of chicks (y_{ijl} , in kg)

Age of chick	Dry concentrate					Total $y_{i.}$
	D_1	D_2	D_3	D_4	D_5	
A_1	2.0,1.8,1.5,	2.0,2.0,1.8	1.8,1.5,1.5	2.0,2.1,2.2	1.5,1.4,1.2	26.3
A_2	1.5,1.5,1.4,	1.6,1.8,1.7,	1.6,1.5,1.5	1.8,2.0,1.9	1.1,1.2,1.0	23.1
A_3	1.4,1.4,1.2	1.3,1.4,1.5	1.0,1.0,1.0	1.4,1.5,1.4	1.0,1.1,1.0	18.6
A_4	1.0,1.0,0.8,	1.0,1.2,1.0	1.0,0.8,0.8	1.0,1.2,1.0	0.8,0.8,1.0	14.4
Total $y_{.j}$	16.5	18.3	15.0	19.5	13.1	82.4

(i) Analyse the data and comment on the use of dry concentrate.

(ii) Do you think that the use of dry concentrate is giving better result for the chicks of age A_1 compared to the chicks of age A_4 ?

Solution : (i) We have $p = 4, q = 5, r = 3$. Since dry concentrates are selected randomly, the effect of dry concentrate is random. But the effect of age may be assumed fixed. We have the total weights of 3 chicks in each group as follows :

The observations [y_{ij}]

Age	Dry concentrate					Mean $\bar{y}_{i..}$
	D_1	D_2	D_3	D_4	D_5	
A_1	5.3	5.8	4.8	6.3	4.1	1.75
A_2	4.4	5.1	4.6	5.7	3.3	1.54
A_3	4.0	4.2	3.0	4.3	3.1	1.24
A_4	2.8	3.2	2.6	3.2	2.6	0.96

$$C.T. = \frac{G^2}{pqr} = \frac{(82.4)^2}{4 \times 5 \times 3} = 113.1627, \text{ Total } SS = \sum \sum \sum y_{ijl}^2 - C.T. = 8.6373.$$

$$SS (\text{Concentrate}) = \frac{1}{pr} \sum y_{.j.}^2 - C.T. = \frac{138400}{4 \times 3} - 113.1627 = 2.1706.$$

$$SS (\text{Age}) = \frac{1}{qr} \sum y_{i..}^2 - C.T. = \frac{1778.62}{5 \times 3} - 113.1627 = 5.4120.$$

$$SS (\text{Age} \times \text{Concentrate}) = \frac{1}{r} \sum \sum y_{ij.}^2 - C.T. - SS (\text{Age}) - SS (\text{Concentrate})$$

$$= \frac{363.76}{3} - 113.1627 - 5.412 - 2.1706 = 0.508.$$

$$SS (\text{Error}) = \text{Total } SS - SS (\text{Age}) - SS (\text{Concentrate}) - SS (\text{Age} \times \text{Concentrate})$$

$$= 0.5467.$$

Anova Table

Sources of variation	d.f.	$MS = \frac{SS}{d.f.}$	$E(MS)$	F	$F_{0.05}$
Age	3	$s_1 = 1.803$	$\sigma^2 + r\sigma_{\alpha\beta}^2 + \frac{r}{p-1} \sum \alpha_i^2$	$F_1 = 42.93$	3.49
Concentrate	4	$s_2 = 0.543$	$\sigma^2 + pr\sigma_{\beta}^2$	$F_2 = 38.78$	2.61
Age \times concentrate	12	$s_3 = 0.042$	$\sigma^2 + r\sigma_{\alpha\beta}^2$	$F_3 = 3.00$	2.00
Error	40	$s_4 = 0.014$	σ^2		
Total	59				

The F -statistic for $H_0 : \sigma_{\alpha\beta}^2 = 0$ is $F_3 = s_3/s_4 = 3.00$ and the test statistic for $H_0 : \sigma_{\beta}^2 = 0$ is $F_2 = s_2/s_4 = 38.78$. It indicates that the concentrates differ significantly. Since $\sigma_{\alpha\beta}^2 > 0$, the test statistic for $H_0 : \alpha_i = 0$ is $F_1 = s_1/s_4 = 42.93$. The weights of chicks differ significantly.

Here $\hat{\sigma}_{\alpha\beta}^2 = \frac{1}{r}(s_3 - s_4) = 0.0093, \hat{\sigma}_{\beta}^2 = \frac{1}{pr}(s_2 - s_4) = 0.042, \hat{\sigma}^2 = 0.014.$

$$\begin{aligned} \text{Also, we have } v(\hat{\sigma}_{\alpha\beta}^2) &= \frac{2}{r^2} \left[\frac{(\hat{\sigma}^2 + r\hat{\sigma}_{\alpha\beta}^2)^2}{(p-1)(q-1)+2} + \frac{\hat{\sigma}^4}{pq(r-1)+2} \right] \\ &= \frac{2}{3^2} \left[\frac{(0.014 + 3 \times 0.0093)^2}{12+2} + \frac{(0.014)^2}{40+2} \right] = 0.00003. \end{aligned}$$

$$\begin{aligned} v(\hat{\sigma}_{\beta}^2) &= \frac{2}{(pr)^2} \left[\frac{(\hat{\sigma}^2 + pr\hat{\sigma}_{\beta}^2)^2}{q+1} + \frac{\hat{\sigma}^4}{pq(r-1)+2} \right] \\ &= \frac{2}{(4 \times 3)^2} \left[\frac{(0.014 + 4 \times 3 \times 0.042)^2}{5+1} + \frac{(0.014)^2}{40+2} \right] = 0.00062. \end{aligned}$$

$$v(\hat{\sigma}^2) = \frac{2\hat{\sigma}^4}{pq(r-1)+2} = \frac{2(0.014)^2}{40+2} = 0.0000093.$$

(ii) We need to test $H_0 : \alpha_1 = \alpha_4$ against $H_A : \alpha_1 \neq \alpha_4$.

$$\text{The test statistic is } t = \frac{\bar{y}_{1..} - \bar{y}_{4..}}{\sqrt{\frac{2(s_3)}{qr}}} = \frac{1.75 - 0.96}{\sqrt{\frac{2 \times 0.042}{15}}} = 10.56.$$

Here s_3 is used in the denominator since this mean square is used as denominator to test $H_0 : \alpha_i = 0$. Since $|t| > t_{0.05,12} = 2.179$, H_0 is rejected. The use of concentrate gives better result for the chicks of age A_1 .

7.7 Variance Component Analysis for Three-Way Classification

Let A, B and C be three factors having levels p, q and r . Consider that the result of l -th level of C corresponding to j -th level of B and i -th level of A is y_{ijl} . The model for this y_{ijl} ($i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, r$) observation is :

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{il} + (\beta\gamma)_{jl} + e_{ijl}.$$

The effects and interactions have their usual meanings.

Assumption : (a) (i) α_i is fixed effect with restriction $\sum \alpha_i = 0$.

(ii) $\beta_j \sim \text{NID}(0, \sigma_{\beta}^2)$, (iii) $\gamma_l \sim \text{NID}(0, \sigma_{\gamma}^2)$, (iv) $(\alpha\beta)_{ij} \sim N(0, \frac{p-1}{p}\sigma_{\alpha\beta}^2)$ (v) $(\alpha\gamma)_{il} \sim \text{NID}(0, \frac{p-1}{p}\sigma_{\alpha\gamma}^2)$, (vi) $(\beta\gamma)_{jl} \sim \text{NID}(0, \sigma_{\beta\gamma}^2)$, (vii) $e_{ijl} \sim \text{NID}(0, \sigma^2)$.

(viii) All the random variables are mutually independent.

$$(ix) \text{Cov}[(\alpha\beta)_{ij}, (\alpha\beta)_{i'j}] = -\frac{1}{p}\sigma_{\alpha\beta}^2, \quad i \neq i'.$$

$$(x) \text{Cov}[(\alpha\gamma)_{il}, (\alpha\gamma)_{i'l}] = -\frac{1}{p}\sigma_{\alpha\gamma}^2.$$

(b) (i) $\alpha_i \sim \text{NID}(0, \sigma_{\alpha}^2)$. Also the assumptions (ii) to (viii) are valid.

The total sum of squares of the observations is partitioned as follows :

$$\begin{aligned} \sum \sum \sum (y_{ijl} - \bar{y}_{...})^2 &= qr \sum (\bar{y}_{i..} - \bar{y}_{...})^2 + pr \sum (\bar{y}_{.j.} - \bar{y}_{...})^2 + pq \sum (\bar{y}_{..l} - \bar{y}_{...})^2 \\ &\quad + r \sum \sum (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + q \sum \sum (\bar{y}_{i.l} - \bar{y}_{i..} - \bar{y}_{..l} + \bar{y}_{...})^2 \\ &\quad \quad \quad + p \sum \sum (\bar{y}_{.jl} - \bar{y}_{.j.} - \bar{y}_{..l} + \bar{y}_{...})^2 \\ &\quad \quad \quad + \sum \sum \sum (y_{ijl} - \bar{y}_{ij.} - \bar{y}_{i.l} - \bar{y}_{.jl} + \bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{..l} - \bar{y}_{...})^2 \\ &= S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7. \end{aligned}$$

Now, we can find the $E(MS)$ of different components present in the model. The $E(MS)$ are shown in the following tables :

Table to find $E(MS)$ according to assumption (a)

Component	$F R R$ $i j l$ $p q r$	d.f.	$MS = \frac{SS}{d.f.}$	$E(MS)$
α_i	0 q r	$p - 1$	s_1	$\sigma^2 + r\sigma_{\alpha\beta}^2 + q\sigma_{\alpha\gamma}^2 + \frac{qr}{p-1} \sum \alpha_i^2$
β_j	p 1 r	$q - 1$	s_2	$\sigma^2 + p\sigma_{\beta\gamma}^2 + pr\sigma_{\beta}^2$
γ_l	p q 1	$r - 1$	s_3	$\sigma^2 + p\sigma_{\beta\gamma}^2 + pq\sigma_{\gamma}^2$
$(\alpha\beta)_{ij}$	0 1 r	$(p - 1)(q - 1)$	s_4	$\sigma^2 + r\sigma_{\alpha\beta}^2$
$(\alpha\gamma)_{il}$	0 q 1	$(p - 1)(r - 1)$	s_5	$\sigma^2 + q\sigma_{\alpha\gamma}^2$
$(\beta\gamma)_{jl}$	p 1 1	$(q - 1)(r - 1)$	s_6	$\sigma^2 + p\sigma_{\beta\gamma}^2$
e_{ijl}	1 1 1	$(p - 1)(q - 1)(r - 1)$	s_7	σ^2

Table to find $E(MS)$ according to assumption (b)

Component	$R R R$ $i j l$ $p q r$	d.f.	$MS = \frac{SS}{d.f.}$	$E(MS)$
α_i	1 q r	$(p - 1)$	s_1	$\sigma^2 + r\sigma_{\alpha\beta}^2 + q\sigma_{\alpha\gamma}^2 + qr\sigma_{\alpha}^2$
β_j	p 1 r	$(q - 1)$	s_2	$\sigma^2 + r\sigma_{\alpha\beta}^2 + p\sigma_{\beta\gamma}^2 + pr\sigma_{\beta}^2$
γ_l	p q 1	$(r - 1)$	s_3	$\sigma^2 + q\sigma_{\alpha\gamma}^2 + p\sigma_{\beta\gamma}^2 + pq\sigma_{\gamma}^2$
$(\alpha\beta)_{ij}$	1 1 r	$(p - 1)(q - 1)$	s_4	$\sigma^2 + r\sigma_{\alpha\beta}^2$
$(\alpha\gamma)_{il}$	1 q 1	$(p - 1)(r - 1)$	s_5	$\sigma^2 + q\sigma_{\alpha\gamma}^2$
$(\beta\gamma)_{jl}$	p 1 1	$(q - 1)(r - 1)$	s_6	$\sigma^2 + p\sigma_{\beta\gamma}^2$
e_{ijl}	1 1 1	$(p - 1)(q - 1)(r - 1)$	s_7	σ^2

Now, to test the significance of $H_0 : \sigma_{\beta\gamma}^2 = 0$, $H_0 : \sigma_{\alpha\gamma}^2 = 0$, $H_0 : \sigma_{\alpha\beta}^2 = 0$, the test statistics are respectively (under both assumptions)

$$F_6 = \frac{s_6}{s_7}, F_5 = \frac{s_5}{s_7}, F_4 = \frac{s_4}{s_7}$$

The test statistic for $H_0 : \sigma_{\gamma}^2 = 0$ [under assumption (a)],

$$F_3 = \frac{s_3}{s_6}, \text{ if } \sigma_{\beta\gamma}^2 > 0.$$

If $\sigma_{\beta\gamma}^2 = 0$, then $F_3 = \frac{s_3}{(S_6 + S_7)/p(q - 1)(r - 1)}$.

Under assumption (b) if $\sigma_{\beta\gamma}^2 = 0$ and $\sigma_{\alpha\gamma}^2 = 0$, then

$$F_3 = \frac{s_3}{(S_5 + S_6 + S_7)/(pq - 1)(r - 1)}$$

But, if $\sigma_{\beta\gamma}^2 = 0$ and $\sigma_{\alpha\gamma}^2 > 0$, then

$$F_3 = \frac{s_3}{s_5}.$$

Again, if $\sigma_{\beta\gamma}^2 > 0$ and $\sigma_{\alpha\gamma}^2 = 0$, then

$$F_3 = \frac{s_3}{s_6}.$$

However, the test statistic takes different form if both $\sigma_{\beta\gamma}^2 > 0$ and $\sigma_{\alpha\gamma}^2 > 0$. Under $H_0 : \sigma_{\gamma}^2 = 0$, we have

$$E(s_3 + s_7) = E(s_5 + s_6).$$

Therefore, $F_3 = \frac{s_3 + s_7}{s_5 + s_6}$ [Satterthwaite (1946)]

The d.f. of F_3 are $\frac{(s_3 + s_7)^2}{\frac{s_3^2}{r-1} + \frac{s_7^2}{(p-1)(q-1)(r-1)}}$ and $\frac{(s_5 + s_6)^2}{\frac{s_5^2}{(p-1)(r-1)} + \frac{s_6^2}{(q-1)(r-1)}}$.

The test statistic for $H_0 : \sigma_{\beta}^2 = 0$ [under assumption (a)] is

$$F_2 = \frac{s_2}{s_6}, \text{ when } \sigma_{\beta\gamma}^2 > 0.$$

If $\sigma_{\beta\gamma}^2 = 0$, then $F_2 = \frac{s_2}{(S_6 + S_7)/p(q-1)(r-1)}$

Under assumption (b) when both $\sigma_{\alpha\beta}^2 = 0$ and $\sigma_{\beta\gamma}^2 = 0$, then

$$F_2 = \frac{s_2}{(S_4 + S_6 + S_7)/(pr-1)(q-1)}.$$

Let $\sigma_{\alpha\beta}^2 = 0$ and $\sigma_{\beta\gamma}^2 > 0$, then $F_2 = \frac{s_2}{s_6}$.

Again, let $\sigma_{\alpha\beta}^2 > 0$ but $\sigma_{\beta\gamma}^2 = 0$, then $F_2 = \frac{s_2}{s_4}$.

If both $\sigma_{\alpha\beta}^2 > 0$ and $\sigma_{\beta\gamma}^2 > 0$, we have under $H_0 : \sigma_{\beta}^2 = 0$,

$$E(s_2 + s_7) = E(s_4 + s_6).$$

Therefore, $F_2 = \frac{s_2 + s_7}{s_4 + s_6}$.

The d.f. of F_2 are $\frac{(s_2 + s_7)^2}{\frac{s_2^2}{q-1} + \frac{s_7^2}{(p-1)(q-1)(r-1)}}$ and $\frac{(s_4 + s_6)^2}{\frac{s_4^2}{(p-1)(q-1)} + \frac{s_6^2}{(q-1)(r-1)}}$.

Finally, the test statistic for $H_0 : \alpha_i = 0$ or, $H_0 : \sigma_{\alpha}^2 = 0$ is

$$F_1 = \frac{s_1}{(S_4 + S_5 + S_7)/(p-1)qr}, \text{ when } \sigma_{\alpha\beta}^2 = 0 \text{ and } \sigma_{\alpha\gamma}^2 = 0.$$

But, if $\sigma_{\alpha\beta}^2 > 0$ and $\sigma_{\alpha\gamma}^2 > 0$, we have under $H_0 : \sigma_{\alpha}^2 = 0$,

$$E(s_1 + s_7) = E(s_4 + s_5).$$

Therefore, $F_1 = \frac{s_1 + s_7}{s_4 + s_5}$.

The d.f. of F_1 are $\frac{(s_1 + s_7)^2}{\frac{s_1^2}{p-1} + \frac{s_7^2}{(p-1)(q-1)(r-1)}}$ and $\frac{(s_4 + s_5)^2}{\frac{s_4^2}{(p-1)(q-1)} + \frac{s_5^2}{(p-1)(r-1)}}$.

Estimation of Variance Components

Assumption (a) : We have

$$\hat{\sigma}^2 = s_7, \hat{\sigma}_{\beta\gamma}^2 = \frac{1}{p}(s_6 - s_7), \hat{\sigma}_{\alpha\gamma}^2 = \frac{1}{q}(s_5 - s_7), \hat{\sigma}_{\alpha\beta}^2 = \frac{1}{r}(s_4 - s_7),$$

$$\hat{\sigma}_{\beta}^2 = \frac{1}{pr}(s_2 - s_6) \quad \text{and} \quad \hat{\sigma}_{\gamma}^2 = \frac{1}{pq}(s_3 - s_6).$$

The variances of these estimates are :

$$V(\hat{\sigma}^2) = \frac{2\sigma^4}{(p-1)(q-1)(r-1)}, \quad V(\hat{\sigma}_{\beta\gamma}^2) = \frac{2}{p^2} \left[\frac{(\sigma^2 + p\sigma_{\beta\gamma}^2)^2}{(q-1)(r-1)} + \frac{\sigma^4}{(p-1)(q-1)(r-1)} \right]$$

$$V(\hat{\sigma}_{\alpha\gamma}^2) = \frac{2}{q^2} \left[\frac{(\sigma^2 + q\sigma_{\alpha\gamma}^2)^2}{(p-1)(r-1)} + \frac{\sigma^4}{(p-1)(q-1)(r-1)} \right]$$

$$V(\hat{\sigma}_{\alpha\beta}^2) = \frac{2}{r^2} \left[\frac{(\sigma^2 + r\sigma_{\alpha\beta}^2)^2}{(p-1)(q-1)} + \frac{\sigma^4}{(p-1)(q-1)(r-1)} \right]$$

$$V(\hat{\sigma}_{\beta}^2) = \frac{2}{(pr)^2} \left[\frac{(\sigma^2 + p\sigma_{\beta\gamma}^2 + pr\sigma_{\beta}^2)^2}{q-1} + \frac{(\sigma^2 + p\sigma_{\beta\gamma}^2)^2}{(q-1)(r-1)} \right]$$

$$V(\hat{\sigma}_{\gamma}^2) = \frac{2}{(pq)^2} \left[\frac{(\sigma^2 + p\sigma_{\beta\gamma}^2 + pq\sigma_{\gamma}^2)^2}{r-1} + \frac{(\sigma^2 + p\sigma_{\beta\gamma}^2)^2}{(q-1)(r-1)} \right].$$

The unbiased estimates of these variances are :

$$v(\hat{\sigma}^2) = \frac{2s_7^2}{(p-1)(q-1)(r-1) + 2},$$

$$v(\hat{\sigma}_{\beta\gamma}^2) = \frac{2}{p^2} \left[\frac{s_6^2}{(q-1)(r-1) + 2} + \frac{s_7^2}{(p-1)(q-1)(r-1) + 2} \right]$$

$$v(\hat{\sigma}_{\alpha\gamma}^2) = \frac{2}{q^2} \left[\frac{s_5^2}{(p-1)(r-1) + 2} + \frac{s_7^2}{(p-1)(q-1)(r-1) + 2} \right]$$

$$v(\hat{\sigma}_{\alpha\beta}^2) = \frac{2}{r^2} \left[\frac{s_4^2}{(p-1)(q-1) + 2} + \frac{s_7^2}{(p-1)(q-1)(r-1) + 2} \right]$$

$$v(\hat{\sigma}_{\beta}^2) = \frac{2}{(pr)^2} \left[\frac{s_2^2}{q+1} + \frac{s_6^2}{(q-1)(r-1) + 2} \right]$$

$$v(\hat{\sigma}_{\gamma}^2) = \frac{2}{(pq)^2} \left[\frac{s_3^2}{r+1} + \frac{s_6^2}{(q-1)(r-1) + 2} \right].$$

Assumption (b) : The estimates, their variances and the estimates of variances for the components $\sigma_{\alpha\beta}^2, \sigma_{\alpha\gamma}^2, \sigma_{\beta\gamma}^2$ and σ^2 are similar as these are for assumption (a). However, we have

$$\hat{\sigma}_{\alpha}^2 = \frac{1}{qr} [s_1 + s_7 - s_4 - s_5], \quad \hat{\sigma}_{\beta}^2 = \frac{1}{pr} [s_2 + s_7 - s_4 - s_6]$$

$$\hat{\sigma}_{\gamma}^2 = \frac{1}{pq} [s_3 + s_7 - s_5 - s_6].$$

The variances of these estimates are :

$$\begin{aligned}
 V(\hat{\sigma}_\alpha^2) &= \frac{2}{(qr)^2} \left[\frac{(\sigma^2 + r\sigma_{\alpha\beta}^2 + q\sigma_{\alpha\gamma}^2 + qr\sigma_\alpha^2)^2}{p-1} + \frac{\sigma^4}{(p-1)(q-1)(r-1)} \right. \\
 &\quad \left. + \frac{(\sigma^2 + q\sigma_{\alpha\gamma}^2)^2}{(p-1)(r-1)} + \frac{(\sigma^2 + p\sigma_{\beta\gamma}^2)^2}{(q-1)(r-1)} \right] \\
 V(\hat{\sigma}_\beta^2) &= \frac{2}{(pr)^2} \left[\frac{(\sigma^2 + p\sigma_{\beta\gamma}^2 + r\sigma_{\alpha\beta}^2 + pr\sigma_\beta^2)^2}{q-1} + \frac{\sigma^4}{(p-1)(q-1)(r-1)} \right. \\
 &\quad \left. + \frac{(\sigma^2 + p\sigma_{\beta\gamma}^2)^2}{(q-1)(r-1)} + \frac{(\sigma^2 + r\sigma_{\alpha\beta}^2)^2}{(p-1)(r-1)} \right] \\
 V(\hat{\sigma}_\gamma^2) &= \frac{2}{(pq)^2} \left[\frac{(\sigma^2 + q\sigma_{\alpha\gamma}^2 + p\sigma_{\beta\gamma}^2 + pq\sigma_\gamma^2)^2}{r-1} + \frac{\sigma^4}{(p-1)(q-1)(r-1)} \right. \\
 &\quad \left. + \frac{(\sigma^2 + q\sigma_{\alpha\gamma}^2)^2}{(p-1)(r-1)} + \frac{(\sigma^2 + p\sigma_{\beta\gamma}^2)^2}{(q-1)(r-1)} \right].
 \end{aligned}$$

The estimates of these variances are :

$$\begin{aligned}
 v(\hat{\sigma}_\alpha^2) &= \frac{2}{(qr)^2} \left[\frac{s_1^2}{p+1} + \frac{s_7^2}{(p-1)(q-1)(r-1)+2} + \frac{s_4^2}{(p-1)(q-1)+2} + \frac{s_5^2}{(p-1)(r-1)+2} \right] \\
 v(\hat{\sigma}_\beta^2) &= \frac{2}{(pr)^2} \left[\frac{s_2^2}{q+1} + \frac{s_7^2}{(p-1)(q-1)(r-1)+2} + \frac{s_4^2}{(p-1)(q-1)+2} + \frac{s_6^2}{(q-1)(r-1)+2} \right] \\
 v(\hat{\sigma}_\gamma^2) &= \frac{2}{(pq)^2} \left[\frac{s_3^2}{r+1} + \frac{s_7^2}{(p-1)(q-1)(r-1)+2} + \frac{s_5^2}{(p-1)(r-1)+2} + \frac{s_6^2}{(q-1)(r-1)+2} \right]
 \end{aligned}$$

Chapter 8

Nested Classification

8.1 Introduction

The analysis of variance what we have so far discussed are cross-classification, where the levels of a factor are tested with all levels of another factor or other factors. As an example, we can mention the case of factorial experiment in which the levels of a factor are crossed with levels of one or more factors. If the factor is variety of an agricultural crop, it can be cultivated using different levels of fertilizer. However, there are some factor or factors the levels of which are similar for another factor but not same. For example, let us consider the study of fertility variation of couples of different social status living in rural and urban areas. The fertility behaviour of urban and rural couples is not expected to be similar. The couples of each area can be divided into different social status, viz., (i) low, (ii) medium, (iii) high. The social status of urban people and rural people may be similar but not same. Hence, we can assume that level of social status (B) are nested within the levels of residential area (A). In such a case many couples of each area and of each social status may be investigated to study the impact of these factors on their fertility. The method of data collection of such experiment is known as nested design.

We have mentioned two-stage nested design the data of which can be arranged as follows :

Factor: A	A_1				A_2				A_p				
Factor: B	B_1, B_2, \dots, B_q				B_1, B_2, \dots, B_q				B_1, B_2, \dots, B_q				
	y_{111}	y_{121}	...	y_{1q1}	y_{pq1}		
	y_{112}	y_{122}	...	y_{1q2}	y_{pq2}		
	\vdots	\vdots	\vdots	\vdots							\vdots		
	y_{11n}	y_{12n}	...	y_{1qn}	y_{pqn}		
Total	$y_{i.}$	$y_{11.}$	$y_{12.}$...	$y_{1q.}$					$y_{p1.}$	$y_{p2.}$...	$y_{pq.}$
	$y_{i.}$	$y_{1.}$				$y_{2.}$				$y_{p.}$			

Here there are n observations for each level of B nested within the levels of A . This is called balanced nested design. The number of observations for different levels of B within levels of A may be unequal. For example, let the number of observations of j th level of B within i th level of A be n_{ij} [$i = 1, 2, \dots, p; j = 1, 2, \dots, q$]. In such a case, the design is known as unbalanced nested design. Again, the levels of B may be different within different levels of A . Let q_i be the levels of B within the level of A_i . This is also a case of unbalanced nested design.

In some case, there may be three factors A, B and C such that the levels of C are nested within the levels of B and the levels of B are nested within the levels of A . Such a design is called *three-stage nested design*. In a similar way, multi-stage nested design can be formulated.

Again, it may happen that the levels of C are similar for all levels of B but not same and all levels of B are same for all levels of A . This is a case of cross and nested classification, where levels of C are nested within the levels of B and levels of B are crossed with levels of A .

As there are many factors (at least two factors) having many levels, the levels may be randomly selected for an experiment or all levels may be used for an experiment. Accordingly, the analysis of data of nested design can be performed assuming random effect and fixed effect models. The effects may be mixed also and the model may be mixed effect model. Let us now discuss some analyses of data collected through nested design.

8.2 Two-Stage Nested Classification

The model assumed for this analysis is

$$y_{ijl} = \mu + \alpha_i + \beta_{j(i)} + e_{ijl},$$

$$i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, r.$$

Here μ = general mean, α_i = effect of i th level of A , $\beta_{j(i)}$ = effect of j th level of B within i th level of A , e_{ijl} = random error, and y_{ijl} = the l th observation of j th level of B within i th level of A .

Assumption :

(a) α_i and $\beta_{j(i)}$ are fixed effects with restrictions $\sum_i \alpha_i = \sum_j \beta_{j(i)} = 0$.

(b) α_i is fixed effect with restriction $\sum \alpha_i = 0$ and $\beta_{j(i)}$ is random variable, where $\beta_{j(i)} \sim NID(0, \sigma_\beta^2)$.

(c) $\alpha_i \sim NID(0, \sigma_\alpha^2)$, $\beta_{j(i)} \sim NID(0, \sigma_\beta^2)$.

In every case, $e_{ijl} \sim NID(0, \sigma^2)$. Moreover, e_{ijl} are mutually independent of other random variables :

Analysis under Assumption (a) : The normal equations to estimate the parameters are :

$$y_{...} = pqr\hat{\mu} + qr \sum \hat{\alpha}_i + r \sum_i \sum_j \hat{\beta}_{j(i)}$$

$$y_{i..} = qr\hat{\mu} + qr\hat{\alpha}_i + r \sum_j \hat{\beta}_{j(i)}$$

$$y_{ij.} = r\hat{\mu} + r\hat{\alpha}_i + r\hat{\beta}_{j(i)}.$$

There are $(pq + p + 1)$ normal equations. All of them are not independent. First equation is the sum of p equations shown in second set and the second set of equations are obtained adding q equations shown in each set of last pq equations. Therefore, only last pq equations are independent. To get unique solutions of the normal equations, we need to put $(p + 1)$ restrictions. The restrictions are $\sum \hat{\alpha}_i = \sum_j \hat{\beta}_{j(i)} = 0$.

Under the restrictions, we have $\hat{\mu} = \bar{y}_{...}$, $\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$, $\hat{\beta}_{j(i)} = \bar{y}_{ij.} - \bar{y}_{i..}$.

The total sum of squares of observations is partitioned into component sum of squares and is shown below :

$$\begin{aligned} \sum \sum \sum (y_{ijl} - \bar{y}_{...})^2 &= qr \sum (\bar{y}_{i..} - \bar{y}_{...})^2 + r \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..})^2 + \sum \sum \sum (y_{ijl} - \bar{y}_{ij.})^2 \\ &= SS(A) + SS(B \text{ within } A) + SS(\text{Error}) = S_1 + S_2 + S_3. \end{aligned}$$

ANOVA Table

Sources of Variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	$E(MS)$	F
A	$p - 1$	S_1	s_1	$\sigma^2 + \frac{qr}{p-1} \sum \alpha_i^2$	$\frac{s_1}{s_3}$
B within A	$p(q - 1)$	S_2	s_2	$\sigma^2 + \frac{r}{p(q-1)} \sum_j \beta_{j(i)}^2$	$\frac{s_2}{s_3}$
Error	$pq(r - 1)$	S_3	s_3	σ^2	
Total	$pqr - 1$				

We need to test the significance of $H_0 : \alpha_i = 0$ and $H_0 : \beta_{j(i)} = 0$. The test statistics for these hypotheses are, respectively $F_1 = s_1/s_3$ and $F_2 = s_2/s_3$. At this stage, if it is needed to compare the pairwise levels of A, then the test statistic is

$$D_k = d_{0.05, k, pq(r-1)} \sqrt{\frac{s_3}{qr}}, \quad k = 2, 3, \dots, p.$$

Analysis under Assumptions (b) and (c) : We have $E(MS)$ under these two assumptions as follows :

$E(MS)$	Assumption	
	(b)	(c)
$E(s_1)$	$\sigma^2 + r\sigma_\beta^2 + \frac{qr}{p-1} \sum \alpha_i^2$	$\sigma^2 + r\sigma_\beta^2 + qr\sigma_\alpha^2$
$E(s_2)$	$\sigma^2 + r\sigma_\beta^2$	$\sigma^2 + r\sigma_\beta^2$
$E(s_3)$	σ^2	σ^2

The test statistic for $H_0 : \sigma_\beta^2 = 0$ is $F_2 = s_2/s_3$. If $\sigma_\beta^2 \geq 0$, the test statistic for $H_0 : \alpha_i = 0$ or $H_0 : \sigma_\alpha^2 = 0$ is $F_1 = s_1/s_2$, otherwise,

$$F_1 = \frac{s_1}{(S_2 + S_3)/p(qr - 1)}$$

We have $\hat{\sigma}^2 = s_3$, $\hat{\sigma}_\beta^2 = \frac{1}{r}(s_2 - s_3)$, $\hat{\sigma}_\alpha^2 = \frac{1}{qr}(s_1 - s_2)$, $V(\hat{\sigma}^2) = \frac{2\sigma^4}{pq(r - 1)}$

$$V(\hat{\sigma}_\beta^2) = \frac{2}{r^2} \left[\frac{(\sigma^2 + r\sigma_\beta^2)^2}{p(q - 1)} + \frac{\sigma^4}{pq(r - 1)} \right],$$

$$V(\hat{\sigma}_\alpha^2) = \frac{2}{(qr)^2} \left[\frac{(\sigma^2 + r\sigma_\beta^2 + qr\sigma_\alpha^2)^2}{p - 1} + \frac{(\sigma^2 + r\sigma_\beta^2)^2}{p(q - 1)} \right].$$

The estimates of these variances are :

$$v(\hat{\sigma}^2) = \frac{2s_3^2}{pq(r - 1) + 2}, \quad v(\hat{\sigma}_\beta^2) = \frac{2}{r^2} \left[\frac{s_2^2}{p(q - 1) + 2} + \frac{s_3^2}{pq(r - 1) + 2} \right].$$

$$v(\hat{\sigma}_\alpha^2) = \frac{2}{(qr)^2} \left[\frac{s_1^2}{p + 1} + \frac{s_2^2}{p(q - 1) + 2} \right].$$

Example 8.1 : The IQ of students of a university are investigated to differentiate the students of different departments within the faculties. The students are selected from four departments of each faculty. The included faculties are three.

The I.Q. of students (y_{ijl})

Faculty (A)	F_1				F_2				F_3			
	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
1	80	85	70	72	65	66	70	72	70	75	80	80
2	85	82	68	70	66	68	75	75	71	72	78	80
3	82	80	65	70	60	70	78	80	70	74	75	78
Total y_{ij}	247	247	203	212	191	204	223	227	211	221	233	238
Total $y_{i..}$	909				845				903			

- (i) Analyse the data and differentiate the faculties, if possible.
 (ii) Is there any difference between D_1 and D_3 in F_1 ?

Solution : (i) We have $p = 3, q = 4, r = 3, G = 2657$ C.T. $\therefore \frac{G^2}{pqr} = 196101.3611$.

$$SS(\text{Total}) = \sum \sum \sum y_{ijl}^2 - \text{C.T.} = 1327.6389.$$

$$SS(A) = \frac{1}{qr} \sum y_{i..}^2 - \text{C.T.} = \frac{2355715}{4 \times 3} - 196101.3611 = 208.2222.$$

$$\begin{aligned} SS(B \text{ within } A) &= \sum_i \left[\frac{\sum_j y_{ij}^2}{r} - \frac{y_{i..}^2}{qr} \right] \\ &= \left[\frac{208171}{3} - \frac{(909)^2}{4 \times 3} \right] + \left[\frac{179355}{3} - \frac{(845)^2}{4 \times 3} \right] + \left[\frac{204295}{3} - \frac{(903)^2}{4 \times 3} \right] \\ &= 533.5833 + 282.9167 + 147.5833 = 964.0833. \end{aligned}$$

$$\begin{aligned} SS(\text{Error}) &= SS(\text{Total}) - SS(A) - SS(B \text{ within } A) \\ &= 155.3334. \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{d.f.}$	F	$F_{0.05}$
A (faculty)	2	208.2222	$s_1 = 104.1111$	16.08	3.40
B within A	9	964.0833	$s_2 = 107.1204$	16.55	2.30
Error	24	155.3334	$s_3 = 6.4722$	—	—
Total	35				

It is observed that in terms of average I.Q. the students of different faculties are different. The averages are $\bar{F}_1 = 75.75, \bar{F}_2 = 70.42, \bar{F}_3 = 75.25$. To differentiate these means we can

perform Duncan's multiple range test, where the test statistic is

$$D_k = d_{0.05,k,pq(r-1)} \sqrt{\frac{s_3}{qr}}, \quad k = 2, 3.$$

$$D_2 = 2.92 \sqrt{\frac{6.4722}{12}} = 2.14, \quad D_3 = 3.07 \sqrt{\frac{6.4722}{12}} = 2.25.$$

Using the values of D_2 and D_3 it may be concluded that the students of F_1 and F_3 are significantly different than the students of F_2 in terms of average I.Q. The students of F_1 and F_3 are similar.

(ii) We need to test the significance of

$$H_0 : \beta_{1(1)} = \beta_{3(1)} \quad \text{against} \quad H_A : \beta_{1(1)} \neq \beta_{3(1)}.$$

The test statistic is

$$t = \frac{\bar{y}_{11.} - \bar{y}_{13.}}{\sqrt{\frac{2s_3}{r}}} = \frac{82.33 - 67.67}{\sqrt{\frac{2 \times 6.4722}{3}}} = 7.06.$$

Since $|t| > t_{0.05,24} = 2.064$, H_0 is rejected. The average I.Q. of students of D_1 and D_3 within F_1 are significantly different.

The above analysis is performed assuming fixed effect model. If it is assumed that the faculties and the departments within a faculty are randomly selected, then we need to test the significance of $H_0 : \sigma_\beta^2 = 0$ and $H_0 : \sigma_\alpha^2 = 0$. The test statistic for $H_0 : \sigma_\beta^2 = 0$ is $F_2 = 16.55$ and it indicates that $\sigma_\beta^2 > 0$, i.e., the variation in average I.Q. of students in different departments within faculties are significantly different.

Since $\sigma_\beta^2 > 0$, the test statistic for $H_0 : \sigma_\alpha^2 = 0$ is $F_1 = s_1/s_2 = 0.97$. This test statistic does not provide evidence of differential I.Q. of students of different faculties.

The estimates of variance components and the estimates of variances of estimates of variance components are given below :

$$\hat{\sigma}_\beta^2 = \frac{1}{r}(s_2 - s_3) = 33.55, \quad \hat{\sigma}_\alpha^2 = \frac{1}{qr}(s_1 - s_2) = -0.25.$$

The negative variance indicates the insignificance of this component. The phenomenon is already been observed by F -test.

$$v(\hat{\sigma}_\beta^2) = \frac{2}{r^2} \left[\frac{s_2^2}{p(q-1)+2} + \frac{s_3^2}{pq(r-1)+2} \right] = \frac{2}{9} \left[\frac{(107.1204)^2}{9+2} + \frac{(6.4722)^2}{24+2} \right]$$

$$= 232.17$$

$$v(\hat{\sigma}_\alpha^2) = \frac{2}{(qr)^2} \left[\frac{s_1^2}{p+1} + \frac{s_2^2}{p(q-1)+2} \right] = \frac{2}{144} \left[\frac{(104.1111)^2}{3+1} + \frac{(107.1204)^2}{9+2} \right]$$

$$= 52.12.$$

8.3 Three-Stage Nested Classification

Let there be three factors A , B and C having levels p , q and r respectively. The levels of C are such that these are similar for each level of B but not same and the levels of B are nested within the levels of A . Let y_{ijkl} be the k -th observation of l -th level of C within j -th level of B and i -th level of A ($i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, r; k = 1, 2, \dots, m$). The model for this y_{ijkl} observation is

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{l(ij)} + e_{ijkl}. \tag{A}$$

where μ = general mean, α_i = effect of i -th level of A , $B_{j(i)}$ = effect of j -th levels of B within i -th level of A , $\gamma_{l(ij)}$ = effect of l -th level of C within i -th and j -th level of A and B , e_{ijkl} is a random error.

Assumption : (a) (i) all effects are fixed effects with restrictions

$$\sum \alpha_i = \sum_j \beta_{j(i)} = \sum_l \gamma_{l(ij)} = 0.$$

(b) (i) $\alpha_i \sim \text{NID}(0, \sigma_\alpha^2)$, (ii) $\beta_{j(i)} \sim \text{NID}(0, \sigma_\beta^2)$ (iii) $\gamma_{l(ij)} \sim \text{NID}(0, \sigma_\gamma^2)$.

(c) (i) α_i is fixed effect with restriction $\sum \alpha_i = 0$,

(ii) $\beta_{j(i)} \sim \text{NID}(0, \sigma_\beta^2)$, (iii) $\gamma_{l(ij)} \sim \text{NID}(0, \sigma_\gamma^2)$.

Moreover, in every case $e_{ijkl} \sim \text{NID}(0, \sigma^2)$ and all random variables are mutually independent.

Analysis under assumption (a) : The normal equations to estimate the parameters are :

$$y_{\dots} = pqr m \hat{\mu} + qrm \sum \hat{\alpha}_i + rm \sum \sum \hat{\beta}_{j(i)} + m \sum \sum \sum \hat{\gamma}_{l(ij)}$$

$$y_{i\dots} = qrm \hat{\mu} + qrm \hat{\alpha}_i + rm \sum_j \hat{\beta}_{j(i)} + m \sum_j \sum_l \hat{\gamma}_{l(ij)}$$

$$y_{ij\dots} = rm \hat{\mu} + rm \hat{\alpha}_i + rm \hat{\beta}_{j(i)} + m \sum_l \hat{\gamma}_{l(ij)}$$

$$y_{ijkl} = m(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_{j(i)} + \hat{\gamma}_{l(ij)}).$$

There are $(pqr + pq + p + 1)$ normal equations. Among these, only pqr equations in the last set are independent. Hence, to get the unique solution of these equations we need to put $pq + p + 1$ restrictions. The restrictions are

$$\sum \hat{\alpha}_i = \sum_j \hat{\beta}_{j(i)} = \sum_l \hat{\gamma}_{l(ij)} = 0.$$

Under the restrictions the estimates are

$$\hat{\mu} = \bar{y}_{\dots}, \hat{\alpha}_i = \bar{y}_{i\dots} - \bar{y}_{\dots}, \hat{\beta}_{j(i)} = \bar{y}_{ij\dots} - \bar{y}_{i\dots}, \hat{\gamma}_{l(ij)} = \bar{y}_{ijkl} - \bar{y}_{ij\dots}$$

The total sum of squares of observations is partitioned as follows :

$$\begin{aligned} \sum \sum \sum \sum (y_{ijkl} - \bar{y}_{\dots})^2 &= qrm \sum (\bar{y}_{i\dots} - \bar{y}_{\dots})^2 + rm \sum \sum (\bar{y}_{ij\dots} - \bar{y}_{i\dots})^2 \\ &\quad + m \sum (\bar{y}_{ijkl} - \bar{y}_{ij\dots})^2 + \sum \sum \sum \sum (y_{ijkl} - \bar{y}_{ijkl})^2 \\ &= SS(A) + SS(B \text{ within } A) + SS(C \text{ within } A \text{ and } B) + SS(\text{Error}) \\ &= S_1 + S_2 + S_3 + S_4. \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	E(MS)	F
A	$p - 1$	S_1	s_1	$\sigma^2 + \frac{qrm}{p-1} \sum \alpha_i^2$	s_1/s_4
B within A	$p(q - 1)$	S_2	s_2	$\sigma^2 + \frac{rm}{p(q-1)} \sum \sum \beta_{j(i)}^2$	s_2/s_4
C within B and A	$pq(r - 1)$	S_3	s_3	$\sigma^2 + \frac{m}{pq(r-1)} \sum \sum \sum \gamma_{l(ij)}^2$	s_3/s_4
Error	$pqr(m - 1)$	S_4	s_4	σ^2	
Total	$pqrm - 1$				

The main objective of the analysis is to test the significance of $H_0 : \alpha_i = 0$, $H_0 : \beta_{j(i)} = 0$ and $H_0 : \gamma_{l(ij)} = 0$. The test statistics for these hypotheses are respectively $F_1 = s_1/s_4$, $F_2 = s_2/s_4$ and $F_3 = s_3/s_4$. The multiple comparison is also done as usual.

Analysis under assumption (b) and (c) : The expected mean square of different components are shown below :

Component	E(MS) under		MS
	Assumption (b)	Assumption (c)	
α_i	$\sigma^2 + m\sigma_\gamma^2 + mr\sigma_\beta^2 + \frac{qrm}{p-1} \sum \alpha_i^2$	$\sigma^2 + m\sigma_\gamma^2 + rm\sigma_\beta^2 + qrm\sigma_\alpha^2$	s_1
$\beta_{j(i)}$	$\sigma^2 + m\sigma_\gamma^2 + rm\sigma_\beta^2$	$\sigma^2 + m\sigma_\gamma^2 + rm\sigma_\beta^2$	s_2
$\gamma_{l(ij)}$	$\sigma^2 + m\sigma_\gamma^2$	$\sigma^2 + m\sigma_\gamma^2$	s_3
e_{ijkl}	σ^2	σ^2	s_4

The objective of these analyses are to test the significance of (i) $H_0 : \sigma_\gamma^2 = 0$, (ii) $H_0 : \sigma_\beta^2 = 0$, (iii) $H_0 : \sigma_\alpha^2 = 0$ or $\alpha_i = 0$. The test statistics are discussed below :

The test statistic for $H_0 : \sigma_\gamma^2$ is $F_3 = s_3/s_4$. If it is observed that $\sigma_\gamma^2 > 0$, then the test statistic for $H_0 : \sigma_\beta^2 = 0$ is $F_2 = s_2/s_3$, otherwise,

$$F_2 = \frac{s_2}{(S_3 + S_4)/pq(rm - 1)}$$

The test statistic for $H_0 : \sigma_\alpha^2 = 0$ or $H_0 : \alpha_i = 0$ is $F_1 = s_1/s_2$, provided $\sigma_\beta^2 > 0$. If $\sigma_\beta^2 = 0$ but $\sigma_\gamma^2 > 0$, then F_1 is to be computed as follows :

$$F_1 = \frac{s_1}{(S_2 + S_3)/p(qr - 1)}$$

If both $\sigma_\gamma^2 = 0$ and $\sigma_\beta^2 = 0$, then $F_1 = \frac{s_1}{(S_2 + S_3 + S_4)/p(qrm - 1)}$.

The estimates of variance components are :

$$\hat{\sigma}^2 = s_4, \sigma_\gamma^2 = \frac{1}{m}(s_3 - s_4), \hat{\sigma}_\beta^2 = \frac{1}{rm}(s_2 - s_3) \text{ and } \hat{\sigma}_\alpha^2 = \frac{1}{qrm}(s_1 - s_2).$$

The variances of these estimates are :

$$V(\hat{\sigma}^2) = \frac{2\sigma^4}{pqr(m - 1)}, V(\hat{\sigma}_\gamma^2) = \frac{2}{m^2} \left[\frac{(\sigma^2 + m\sigma_\gamma^2)^2}{pq(r - 1)} + \frac{\sigma^4}{pqr(m - 1)} \right]$$

$$V(\hat{\sigma}_\beta^2) = \frac{2}{(rm)^2} \left[\frac{(\sigma^2 + m\sigma_\gamma^2 + rm\sigma_\beta^2)^2}{p(q - 1)} + \frac{(\sigma^2 + m\sigma_\gamma^2)^2}{pq(r - 1)} \right]$$

$$V(\hat{\sigma}_\alpha^2) = \frac{2}{(qrm)^2} \left[\frac{(\sigma^2 + m\sigma_\gamma^2 + rm\sigma_\beta^2 + qrm\sigma_\alpha^2)^2}{p - 1} + \frac{(\sigma^2 + m\sigma_\gamma^2 + rm\sigma_\beta^2)^2}{p(q - 1)} \right]$$

The estimates of these variances are :

$$v(\hat{\sigma}^2) = \frac{2s_4^2}{pqr(m - 1) + 2}, v(\hat{\sigma}_\gamma^2) = \frac{2}{m^2} \left[\frac{s_3^2}{pq(r - 1) + 2} + \frac{s_4^2}{pqr(m - 1) + 2} \right]$$

$$v(\hat{\sigma}_\beta^2) = \frac{2}{(rm)^2} \left[\frac{s_2^2}{p(q - 1) + 2} + \frac{s_3^2}{p(q - 1) + 2} \right], v(\hat{\sigma}_\alpha^2) = \frac{2}{(qrm)^2} \left[\frac{s_1^2}{p + 1} + \frac{s_2^2}{p(q - 1) + 2} \right]$$

Example 8.2 : An experiment is conducted to study the differential in productivity of local variety (LV) of rice and high yielding variety of rice (HYV). Each variety has two crops, viz., *Aus* and *Aman*. The rice varieties are cultivated in three different seasons in different agricultural plots. The production per plot (y_{ijkl} kg) in different replications are recorded for analysis. Analyse the data and comment on the differentiability of rice variety and crop variety within a variety.

The production of rice (y_{ijkl} kg)

Replication	Season (A)											
	S_1				S_2				S_3			
	Variety of rice (B)											
	LV		HYV		LV		HYV		LV		HYV	
	Crop type (C)											
	<i>Aus</i>	<i>Aman</i>	<i>Aus</i>	<i>Aman</i>	<i>Aus</i>	<i>Aman</i>	<i>Aus</i>	<i>Aman</i>	<i>Aus</i>	<i>Aman</i>	<i>Aus</i>	<i>Aman</i>
1	8.5	10.2	15.6	20.8	9.5	12.5	18.6	25.6	8.0	10.4	15.0	21.6
2	9.0	10.4	16.0	20.0	9.0	14.6	18.0	25.0	8.0	11.0	15.0	22.4
3	9.2	10.6	16.0	20.5	9.8	13.8	18.5	25.8	8.4	11.2	14.8	22.0
4	9.3	10.4	16.2	20.6	9.4	12.8	18.2	25.5	8.3	11.5	15.3	21.8
Total y_{ijl}	36.0	41.6	63.8	81.9	37.7	53.7	73.3	101.9	32.7	44.1	60.1	87.8
$y_{ij..}$	77.6		145.7		91.4		175.2		76.8		147.9	
$y_{i...}$	223.3				266.6				224.7			

Solution : We have $p = 3$, $q = 2$, $r = 2$, $m = 4$, $G = 714.6$

$$C.T. = \frac{G^2}{pqrm} = \frac{(714.6)^2}{48} = 10638.6075,$$

$$SS(\text{Total}) = \sum \sum \sum \sum y_{ijkl}^2 - C.T. = 1417.5725.$$

$$SS(A) = \frac{1}{qrm} \sum y_{i...}^2 - C.T. = \frac{171428.54}{16} - 10638.6075 = 75.6762.$$

$$\begin{aligned} SS(B \text{ within } A) &= \sum_i \left[\frac{\sum_j y_{ij..}^2}{rm} - \frac{y_{i...}^2}{qrm} \right] \\ &= \left[\frac{27250.25}{8} - \frac{(223.3)^2}{16} \right] + \left[\frac{39049.00}{8} - \frac{(266.6)^2}{16} \right] \\ &\quad + \left[\frac{27772.65}{8} - \frac{(224.7)^2}{16} \right] \\ &= 281.4712 + 438.9025 + 315.9606 = 1036.3243. \end{aligned}$$

$$\begin{aligned} SS(C \text{ within } A \text{ and } B) &= \sum_i \sum_j \left[\frac{\sum_l y_{ijl.}^2}{m} - \frac{y_{ij..}^2}{rm} \right] \\ &= [756.64 - 752.72] + [2694.5125 - 2653.5612] \end{aligned}$$

$$\begin{aligned}
 &+ [1076.245 - 1044.245] + [3939.125 - 3836.88] \\
 &+ [753.525 - 737.2] + [2830.2125 - 2734.3012] = 246.4013. \\
 SS(\text{Error}) &= SS(\text{Total}) - SS(A) - SS(B \text{ within } A) - SS(C \text{ within } A \text{ and } B) = 59.1707.
 \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F	F _{0.05}
A	2	75.6762	37.8381	23.02	3.27
B within A	3	1036.3243	345.4414	210.17	2.87
C with A and B	6	246.4013	41.0668	24.98	2.37
Error	36	59.1707	1.6436	—	
Total	47				

It is observed that the crop variety within a variety differs significantly ($\because F_3 > F_{0.05;6,36}$). Again, the variety of rice within seasons also differs significantly.

The above analysis has been performed assuming fixed effect model. Let us assume that the model is a random effect model. Then to test the significance of $H_0 : \sigma_\gamma^2 = 0$, the test statistic is $F_3 = 24.98$ and it is observed that $\sigma_\gamma^2 > 0$. The test statistic for $H_0 : \sigma_\beta^2 = 0$ is $F_2 = s_2/s_3 = 8.41$. This test indicates that $\sigma_\beta^2 > 0$. Now, the test statistic for $H_0 : \sigma_\alpha^2 = 0$ is $F_1 = s_1/s_2 = 0.11$. This indicates that the seasonal variation in production is insignificant.

The estimates of variance components and the estimates of variance of the estimates of variance components are given below :

$$\hat{\sigma}^2 = 1.6436, \hat{\sigma}_\beta = 38.05, \hat{\sigma}_\gamma^2 = 9.86, \hat{\sigma}_\alpha^2 = -19.23.$$

The negative variance indicates insignificant effect of season.

$$v(\hat{\sigma}^2) = \frac{2s_4^2}{pq(m-1)r+2} = \frac{2(1.6436)^2}{36+2} = 0.142.$$

$$\begin{aligned}
 v(\hat{\sigma}_\gamma^2) &= \frac{2}{m^2} \left[\frac{s_3^2}{pq(r-1)+2} + \frac{s_4^2}{pqr(m-1)+2} \right] = \frac{2}{16} \left[\frac{(41.0668)^2}{6+2} + \frac{(1.6436)^2}{36+2} \right] \\
 &= 26.36.
 \end{aligned}$$

$$\begin{aligned}
 v(\hat{\sigma}_\beta^2) &= \frac{2}{(rm)^2} \left[\frac{s_2^2}{p(q-1)+2} + \frac{s_3^2}{pq(r-1)+2} \right] = \frac{2}{64} \left[\frac{(345.4414)^2}{3+2} + \frac{(41.0668)^2}{6+2} \right] \\
 &= 752.40.
 \end{aligned}$$

$$\begin{aligned}
 v(\hat{\sigma}_\alpha^2) &= \frac{2}{(qrm)^2} \left[\frac{s_1^2}{p+1} + \frac{s_2^2}{p(q-1)+2} \right] = \frac{2}{256} \left[\frac{(37.8381)^2}{3+1} + \frac{(345.4414)^2}{3+2} \right] \\
 &= 189.25.
 \end{aligned}$$

8.4 Unbalanced Three-Stage Nested Classification

The model for this analysis is

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{l(ij)} + e_{ijkl}, \tag{A}$$

$i = 1, 2, \dots, p; j = 1, 2, \dots, q_i, l = 1, 2, \dots, r_{ij}; k = 1, 2, \dots, n_{ijl}.$

The parameters in the model have their usual meanings. The assumptions to analyse the data are assumed to be similar to the assumptions discussed in the previous section.

$$\text{Let } n = \sum_i \sum_j \sum_l n_{ijl}, \quad N_{i..} = \sum_j \sum_l n_{ijl}, \quad N_{ij.} = \sum_l n_{ijl}, \quad R = \sum_i \sum_j r_{ij}, \quad R_i = \sum_j r_{ij},$$

$$Q = \sum_i q_i.$$

The normal equations to estimate the parameters in the model (A) are :

$$\begin{aligned} y_{i...} &= n\hat{\mu} + \sum N_{i..}\hat{\alpha}_i + \sum \sum N_{ij.}\hat{\beta}_{j(i)} + \sum \sum \sum n_{ijl}\hat{\gamma}_{l(ij)} \\ y_{i..} &= N_{i..}\hat{\mu} + N_{i..}\hat{\alpha}_i + \sum_j N_{ij.}\hat{\beta}_{j(i)} + \sum_j \sum_l n_{ijl}\hat{\gamma}_{l(ij)} \\ y_{ij.} &= N_{ij.}\hat{\mu} + N_{ij.}\hat{\alpha}_i + N_{ij.}\hat{\beta}_{j(i)} + \sum_l n_{ijl}\hat{\gamma}_{l(ij)} \\ y_{ijl} &= n_{ijl}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_{j(i)} + \hat{\gamma}_{l(ij)}). \end{aligned}$$

To get the unique solution of these normal equations we need to put the following restrictions :

$$\sum_i N_{i..}\hat{\alpha}_i = \sum_j N_{ij.}\hat{\beta}_{j(i)} = \sum_l n_{ijl}\hat{\gamma}_{l(ij)} = 0.$$

Under the restrictions the estimates are :

$$\begin{aligned} \hat{\mu} &= \bar{y}_{i...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{i...}, \quad \hat{\beta}_{j(i)} = \bar{y}_{ij.} - \bar{y}_{i...} \\ \hat{\gamma}_{l(ij)} &= \bar{y}_{ijl} - \bar{y}_{ij.}. \quad \text{Here the means are :} \\ \bar{y}_{i...} &= \frac{1}{N_{i..}} \sum_j N_{ij.}\bar{y}_{ij.}, \quad \bar{y}_{ij.} = \frac{1}{N_{ij.}} \sum_l \bar{y}_{ijl}, \quad \bar{y}_{ijl} = \frac{1}{n_{ijl}} \sum_k y_{ijlk}, \\ \bar{y}_{i...} &= \frac{1}{n} \sum_i N_{i..}\bar{y}_{i...}. \end{aligned}$$

The sum of squares due to estimate is

$$\begin{aligned} SS(\text{estimates}) &= \hat{\mu}y_{i...} + \sum \hat{\alpha}_i y_{i..} + \sum \sum \hat{\beta}_{j(i)} y_{ij.} + \sum \sum \sum n_{ijl} y_{ijl} \\ &= \sum_i \sum_j \sum_l \frac{1}{n_{ijl}} y_{ijl}^2. \end{aligned}$$

The d.f. of this sum of squares is R .

Therefore, the sum of squares due to error is

$$S_4 = \sum_i \sum_j \sum_l \sum_k y_{ijlk}^2 - \sum_i \sum_j \sum_l \frac{1}{n_{ijl}} y_{ijl}^2.$$

The d.f. of S_4 is $(n - R)$.

The main objective of this analysis is to test the following hypotheses :

(i) $H_0 : \alpha_i = 0$, (ii) $H_0 : \beta_{j(i)} = 0$, (iii) $H_0 : \gamma_{l(ij)} = 0$.

Under null hypothesis (iii), the model becomes

$$y_{ijlk} = \mu + \alpha_i + \beta_{j(i)} + e_{ijlk}. \quad (\text{B})$$

$i = 1, 2, \dots, p$; $j = 1, 2, \dots, q_i$; $l = 1, 2, \dots, r_{ij}$; $k = 1, 2, \dots, n_{ijl}$.

The sum of squares due to estimates following the model (B) is

$$S_5 = SS(\text{Estimates}) = \hat{\mu}y_{i\dots} + \sum \hat{\alpha}_i y_{i\dots} + \sum \sum \hat{\beta}_{j(i)} y_{ij\dots}$$

$$= \sum_i \sum_j \frac{1}{N_{ij}} y_{ij\dots}^2$$

The error sum of squares for the model is

$$S_6 = \sum \sum \sum y_{ijkl}^2 - \sum_i \sum_j \frac{1}{N_{ij}} y_{ij\dots}^2$$

The d.f. of S_6 is $n - Q$. Therefore, $SS(\hat{\gamma}_{l(ij)})$ under null hypothesis (iii) is

$$S_3 = S_6 - S_4 = \sum_i \sum_j \left[\sum_l \frac{y_{ijl\dots}^2}{n_{ijl}} - \frac{y_{ij\dots}^2}{N_{ij}} \right]$$

The d.f. of S_3 is $R - Q$. Hence, the test statistic for $H_0 : \gamma_{l(ij)} = 0$ is

$$F_3 = \frac{(S_6 - S_4)/(R - Q)}{S_4/(n - R)}$$

Under null hypothesis (ii), the model (B) becomes

$$y_{ijkl} = \mu + \alpha_i + e_{ijkl} \tag{C}$$

The sum of squares due to estimates following mode (C) is

$$S_7 = SS(\text{Estimates}) = \hat{\mu}y_{i\dots} + \sum \hat{\alpha}_i y_{i\dots} = \sum \frac{1}{N_{i\dots}} y_{i\dots}^2$$

The d.f. of S_7 is p . The error sum of squares of this model is

$$S_8 = SS(\text{Error}) = \sum \sum \sum \sum y_{ijkl}^2 - \sum \frac{1}{N_{i\dots}} y_{i\dots}^2$$

The d.f. of S_8 is $(n - p)$. Hence, the $SS(\hat{\beta}_{j(i)})$ is

$$S_2 = S_8 - S_6 = \sum_i \left[\sum \frac{1}{N_{ij}} y_{ij\dots}^2 - \frac{y_{i\dots}^2}{N_{i\dots}} \right]$$

The d.f. of S_2 is $(Q - p)$. Hence, the test statistic for $H_0 : \beta_{j(i)} = 0$ is

$$F_2 = \frac{(S_8 - S_6)/Q - p}{S_4/(n - R)}$$

Following the model (C), we have

$$S_1 = SS(\hat{\alpha}_i) = \sum \frac{1}{N_{i\dots}} y_{i\dots}^2 - \frac{y_{i\dots}^2}{n}$$

This S_1 has $(p - 1)$ d.f. Hence, the test statistic for $H_0 : \alpha_i = 0$ is

$$F_1 = \frac{S_1/(p - 1)}{S_4/(n - R)}$$

Let us now write the analysis of variance table for mixed effect and random effect model.

ANOVA Table

Sources of variation	d.f	SS	MS = $\frac{SS}{d.f}$	E(MS)	
				Under Assumption (b)	Under Assumption (c)
A	$p-1$	S_1	s_1	$\sigma^2 + C_3\sigma_\gamma^2 + C_4\sigma_\beta^2 + \frac{1}{p-1} \sum_i N_{i..} \alpha_i^2$	$\sigma^2 + C_3\sigma_\gamma^2 + C_4\sigma_\beta^2 + C_5\sigma_\alpha^2$
B within A	$Q-p$	S_2	s_2	$\sigma^2 + C_1\sigma_\gamma^2 + C_2\sigma_\beta^2$	$\sigma^2 + C_1\sigma_\gamma^2 + C_2\sigma_\beta^2$
C within B and A	$R-Q$	S_3	s_3	$\sigma^2 + C_0\sigma_\gamma^2$	$\sigma^2 + C_0\sigma_\gamma^2$
Error	$n-R$	S_4	s_4	σ^2	σ^2
Total	$n-1$				

$$\text{Here } C_0 = \frac{1}{R-Q} \left[n - \sum_i \sum_j \left(\frac{\sum_l n_{ijl}^2}{N_{ij.}} \right) \right], \quad C_2 = \frac{1}{Q-p} \left[n - \sum_j \left(\frac{\sum_i N_{ij.}^2}{N_{i..}} \right) \right]$$

$$C_1 = \frac{1}{Q-p} \left[\sum_i \left\{ \sum_j \left(\frac{\sum_l n_{ijl}^2}{N_{ij.}} \right) \right\} - \sum_i \left(\frac{\sum_j \sum_l n_{ijl}^2}{N_{i..}} \right) \right]$$

$$C_3 = \frac{1}{p-1} \left[\sum_i \left(\frac{\sum_j \sum_l n_{ijl}^2}{N_{i..}} \right) - \frac{1}{n} \sum \sum \sum n_{ijl}^2 \right]$$

$$C_4 = \frac{1}{p-1} \left[\sum_i \left(\frac{\sum_j N_{ij.}}{N_{i..}} \right) - \frac{1}{n} \sum \sum N_{ij.}^2 \right], \quad C_5 = \frac{1}{p-1} \left[n - \frac{\sum N_{i..}^2}{n} \right].$$

The test statistic to test the significance of $H_0 : \sigma_\gamma^2 = 0$ is $F_3 = s_3/s_4$. If $\sigma_\gamma^2 = 0$ is noted down, then the test statistic for $H_0 : \sigma_\beta^2 = 0$ is

$$F_2 = \frac{s_2}{(S_3 + S_4)/(n-Q)}$$

However, if $\sigma_\gamma^2 > 0$, F_2 does not provide exact result. An approximate test statistic can be found out as follows : Under $H_0 : \sigma_\beta^2 = 0$, we have

$$E[C_0s_2 + C_1s_4] = E[C_1s_3 + C_0s_4].$$

Hence, the test statistic is

$$F_4 = \frac{C_0s_2 + C_1s_4}{C_1s_3 + C_0s_4}.$$

The d.f. of this F_4 are

$$\frac{(C_0s_2 + C_1s_4)^2}{\frac{(C_0s_2)^2}{Q-p} + \frac{(C_1s_4)^2}{n-R}} \quad \text{and} \quad \frac{(C_1s_3 + C_0s_4)^2}{\frac{(C_1s_3)^2}{R-Q} + \frac{(C_0s_4)^2}{n-R}}.$$

The F -statistic for $H_0 : \sigma_\alpha^2 = 0$ can be found out in a similar way. Here, under $\sigma_\beta^2 = 0$, we have

$$E[C_1s_1 + C_3s_4] = E[C_3s_2 + C_1s_4].$$

We have $\hat{\sigma}^2 = s_4$, $\hat{\sigma}_\gamma^2 = \frac{1}{C_0}(s_3 - s_4)$,

$\hat{\sigma}_\alpha^2$ (approx) = $\frac{1}{C_5}(s_1 - s_2)$

$\hat{\sigma}_\beta^2 = \frac{1}{C_0 C_2} [C_0 s_2 + C_1 s_4 - C_1 s_3 - C_0 s_4]$.

If $\sigma_\beta^2 > 0$, the exact estimate of σ_α^2 is not available.

The estimates of variance of the above estimates

$v(\hat{\sigma}^2) = \frac{2s_4^2}{n - R + 2}$, $v(\hat{\sigma}_\gamma^2) = \frac{2}{C_0^2} \left[\frac{s_3^2}{R - Q + 2} + \frac{s_4^2}{n - R + 2} \right]$,

$v(\hat{\sigma}_\beta^2) = \frac{2}{(C_0 C_2)^2} \left[\frac{(C_0 s_4)^2}{n - R + 2} + \frac{(C_1 s_3)^2}{R - Q + 2} + \frac{(C_0 s_2)^2}{Q - p + 2} + \frac{(C_1 s_4)^2}{n - R + 2} \right]$.

The estimate $\hat{\sigma}_\alpha^2$ is exact, if $C_1 = C_3$ and $C_2 = C_4$.

Under this condition the estimates of variance of $\hat{\sigma}_\alpha^2$ is

$v(\hat{\sigma}_\alpha^2) = \frac{2}{C_5^2} \left[\frac{s_1^2}{p + 1} + \frac{s_2^2}{Q - p + 2} \right]$.

Example 8.3 : A social scientist investigated the fertility behaviour of some urban and rural couples. The couples were classified into three social classes, viz., Low (*L*), Medium (*M*) and High (*H*). The couples of each social status have different levels of education. The couples were divided into three groups, viz., illiterate (*I*), educated (*E*) and higher educated (*HE*) in respect of female's education. The number of ever born children of each couple is recorded for analysis. Investigate the variability in fertility behavior according to social factors.

Number of ever born children (y_{ijkl}) per couple

Replication	Area (A)												
	Urban						Rural						
	Social status (B)												
	L		M			H			L		M		
	Level of education												
	I	E	I	E	HE	I	E	HE	I	E	I	E	HE
	4	3	5	2	2	5	4	2	6	4	6	4	2
	5	3	4	3	2	6	4	2	7	3	5	3	2
	6	5		3	2	4	3		4	3	6	3	
	3			4	3				3		4		
	4				2				5				
									6				
Total y_{ijl}	22	11	9	12	11	15	11	4	31	10	21	10	4
$y_{ij..}$	33		32			30			41		35		
$y_{i...}$	95						76						

Here $n = 45$; $N_{11} = 8$, $N_{12} = 11$, $N_{13} = 8$, $N_{21} = 9$, $N_{22} = 9$; $N_{1..} = 27$.

$N_{2..} = 18$; $q_1 = 3$, $q_2 = 2$; $r_{11} = 2$, $r_{12} = 3$, $r_{13} = 3$, $r_{21} = 2$, $r_{22} = 3$; $Q = 5$;

$R = 13$; $C_0 = 3.21$, $C_1 = 3.82$, $C_2 = 8.93$, $C_3 = 3.96$, $C_4 = 9.09$.

$C_5 = 21.6$; $G = 171$, C.T. = $\frac{G^2}{n} = 649.8$, $SS(\text{Total}) = 85.2$.

$$SS(A) = \sum_i \frac{y_{i..}^2}{N_{i..}} - \text{C.T.} = 5.35,$$

$$SS(B \text{ within } A) = \sum_l \left[\sum_j \frac{y_{l.j.}^2}{N_{l.j.}} - \frac{y_{l..}^2}{N_{l..}} \right] = (341.72 - 334.26) + (322.89 - 320.89) \\ = 9.46.$$

$$SS(C \text{ within } A \text{ and } B) = \sum_i \sum_j \left[\sum_l \frac{y_{ijl.}^2}{n_{ijl.}} - \frac{y_{ij.}^2}{N_{ij.}} \right] \\ = (137.133 - 136.125) + (100.7 - 93.091) + (123.335 - 112.8) \\ + (193.5 - 186.778) + (151.583 - 136.111) \\ = 41.344.$$

$$SS(\text{Error}) = SS(\text{Total}) - SS(A) - SS(B \text{ within } A) - SS(C \text{ within } A \text{ and } B) = 29.046.$$

ANOVA Table

Sources of variation	d.f.	SS	$MS = \frac{SS}{\text{d.f.}}$	$E(MS)$ for random effect model	F (fixed effect)
A	1	5.35	5.35	$\sigma^2 + 3.96\sigma_\gamma^2 + 9.09\sigma_\beta^2 + 21.6\sigma_\alpha^2$	5.89
B within A	3	9.46	3.153	$\sigma^2 + 3.82\sigma_\gamma^2 + 8.93\sigma_\beta^2$	3.47
C within A and B	8	41.344	5.168	$\sigma^2 + 3.21\sigma_\gamma^2$	5.69
Error	32	29.046	0.908	σ^2	-
Total	44				

The fertility levels in rural and urban areas are significantly different [$F_1 = 5.89 > F_{0.05;1,32} = 4.16$]. The fertility levels according to social status within an area are also significantly different [$F_2 = 3.47 > F_{0.05;3,32} = 2.91$]. The fertility levels vary significantly by levels of education within social status and area [$F_3 = 5.69 > F_{0.05;8,32} = 2.26$].

Let us now discuss the analysis assuming random effect model. The test statistic to test the significance of $H_0 : \sigma_\gamma^2 = 0$ is $F_3 = s_1/s_3 = 5.69$. It indicates that $\sigma_\gamma^2 > 0$. Since $\sigma_\gamma^2 > 0$, the test statistic for $H_0 : \sigma_\beta^2 = 0$ is

$$F_4 = \frac{C_0s_2 + C_1s_4}{C_1s_3 + C_0s_4} = \frac{3.21 \times 3.153 + 3.82 \times 0.908}{3.82 \times 5.168 + 3.21 \times 0.908} = \frac{13.590}{22.656} = 0.60.$$

The d.f. of F_4 are

$$\frac{(C_0s_2 + C_1s_4)^2}{\frac{(C_0s_2)^2}{Q_P} + \frac{(C_1s_4)^2}{n-R}} = \frac{(13.590)^2}{34.522} = 5, \quad \frac{(C_1s_3 + C_0s_4)^2}{\frac{(C_1s_3)^2}{R-Q} + \frac{(C_0s_4)^2}{n-R}} = \frac{(22.656)^2}{48.983} = 10.$$

Thus, $H_0 : \sigma_\beta^2 = 0$ is true.

Since $\sigma_\beta^2 = 0$, the test statistic for $H_0 : \sigma_\alpha^2 = 0$ is

$$F = \frac{C_1 s_1 + C_3 s_4}{C_3 s_2 + C_1 s_4} = \frac{3.82 \times 5.35 + 3.96 \times 0.908}{3.96 \times 3.153 + 3.82 \times 0.908} = \frac{24.318}{15.954} = 1.52.$$

The d.f. of F are :

$$\frac{(C_1 s_1 + C_3 s_4)^2}{\frac{(C_1 s_1)^2}{p-1} + \frac{(C_3 s_4)^2}{n-R}} = \frac{(24.318)^2}{418.075} = 1, \quad \frac{(C_3 s_2 + C_1 s_4)^2}{\frac{(C_3 s_2)^2}{Q-p} + \frac{(C_1 s_4)^2}{n-R}} = \frac{(15.954)^2}{52.342} = 5.$$

This also indicates that $\sigma_\alpha^2 = 0$.

The estimates of variance components are :

$$\hat{\sigma}^2 = s_4 = 0.908, \quad \hat{\sigma}_\gamma^2 = \frac{1}{C_0} (s_3 - s_4) = 1.33.$$

$$\hat{\sigma}_\beta^2 = \frac{1}{C_0 C_2} [C_0 s_2 + C_1 s_4 - C_1 s_3 - C_0 s_4] = -0.32 \text{ (insignificant).}$$

We have $C_1 \approx 4$, $C_3 = 4$, $C_2 \approx 9$ and $C_4 \approx 9$.

Therefore, $\hat{\sigma}_\alpha^2 = \frac{1}{C_5} (s_1 - s_2) = 0.10$.

$$v(\hat{\sigma}^2) = \frac{2s_4^2}{n-R+2} = 0.048.$$

$$v(\hat{\sigma}_\beta^2) = \frac{2}{(C_0 C_2)^2} \left[\frac{(C_0 s_2)^2}{Q-p+2} + \frac{(C_1 s_4)^2}{n-R+2} + \frac{(C_1 s_3)^2}{R-Q+2} + \frac{(C_0 s_4)^2}{n-R+2} \right]$$

$$= \frac{2 \times 59.961}{821.699} = 0.146.$$

$$v(\hat{\sigma}_\alpha^2) = \frac{2}{C_5^2} \left[\frac{s_1^2}{p+1} + \frac{s_2^2}{Q-p+2} \right] = 0.049.$$

8.5 Nested and Cross Classification

Let there be three factors A, B and C having levels p, q and r respectively. The q levels of B are such that these can be used for any level of B . For example, if the levels of A are doses of fertilizer and the levels of B are varieties of a crop, then each level of B can be cultivated using any or all levels of A . Here all levels of B are same for all levels of A . The levels of factors A and B are crossed. Again, let us consider that the r levels of C are such that these are similar for any level of B but not same. For example, let the levels of C be the seed variety of a crop variety. The seed varieties are similar for different crop varieties but not same. Here the levels of C are nested within the levels of B . Thus, we have nested and get cross classification in performing experiment with the levels of A, B and C .

Let y_{ijkl} be the experimental result of k -th replication of l -th level of C nested within the j -th level of B and in presence of i -th level of A ($i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, r; k = 1, 2, \dots, m$). The model for this observation is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{l(j)} + (\alpha\gamma)_{il(j)} + e_{ijkl}. \tag{A}$$

Here μ = general mean, α_i = effect of i -th level of A , β_j = effect of j -th level of B , $(\alpha\beta)_{ij}$ = interaction of j -th level of B with i -th level of A , $\gamma_{l(j)}$ = effect of l -th level of C within j -th level

of B , $(\alpha\gamma)_{il(j)}$ = interaction of i -th level of A and l -th level of C within j -th level of B , e_{ijkl} = random error. Let us assume that the model (A) is a fixed effect model with restrictions :

$$\sum \alpha_i = \sum \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = \sum_l \gamma_{l(j)} = \sum_i (\alpha\gamma)_{il(j)} = \sum_l (\alpha\gamma)_{il(j)} = 0.$$

The main assumption for the analysis is that the errors are normally and independently distributed.

The normal equations to estimate the parameters in the model (A) are :

$$y_{...} = pqr m \hat{\mu} + qrm \sum \hat{\alpha}_i + pr m \sum \hat{\beta}_j + rm \sum \sum (\alpha\hat{\beta})_{ij} + pm \sum \sum \hat{\gamma}_{l(j)} + m \sum_i \sum_j \sum_l (\alpha\hat{\gamma})_{il(j)}.$$

$$y_{i...} = qrm \hat{\mu} + qrm \hat{\alpha}_i + rm \sum \hat{\beta}_j + rm \sum (\alpha\hat{\beta})_{ij} + m \sum \sum \hat{\gamma}_{l(j)} + m \sum \sum (\alpha\hat{\gamma})_{il(j)}.$$

$$y_{.j..} = pr m \hat{\mu} + rm \sum \hat{\alpha}_i + pr m \hat{\beta}_j + rm \sum_i (\alpha\hat{\beta})_{ij} + mp \sum_l \hat{\gamma}_{l(j)} + m \sum_i \sum_l (\alpha\hat{\gamma})_{il(j)}.$$

$$y_{ij.} = rm \hat{\mu} + rm \hat{\alpha}_i + rm \hat{\beta}_j + rm (\alpha\hat{\beta})_{ij} + m \sum_l \hat{\gamma}_{l(j)} + m \sum_l (\alpha\hat{\gamma})_{il(j)}.$$

$$y_{.jl.} = pm \hat{\mu} + m \sum \hat{\alpha}_i + pm \hat{\beta}_j + m \sum_i (\alpha\hat{\beta})_{ij} + pm \hat{\gamma}_{l(j)} + m \sum_i (\alpha\hat{\gamma})_{il(j)}.$$

$$y_{ijl.} = m(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + (\alpha\hat{\beta})_{ij} + \hat{\gamma}_{l(j)} + (\alpha\hat{\gamma})_{il(j)}).$$

There are $(pqr + pq + qr + p + q + 1)$ normal equations. But except the pqr equations in the last set all other equations depend on last set of equations. Hence, to get the unique solution of these equations, we need to put the following restrictions :

$$\sum \hat{\alpha}_i = \sum \hat{\beta}_j = \sum_i (\widehat{\alpha\beta})_{ij} = \sum_j (\widehat{\alpha\beta})_{ij} = \sum_l \hat{\gamma}_{l(j)} = \sum_i (\widehat{\alpha\gamma})_{il(j)} = \sum_l (\widehat{\alpha\gamma})_{il(j)} = 0.$$

Under the restrictions, and on simplification, we get

$$\hat{\mu} = \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i...} - \bar{y}_{...}, \quad \hat{\beta}_j = \bar{y}_{.j..} - \bar{y}_{...},$$

$$(\alpha\hat{\beta})_{ij} = \bar{y}_{ij.} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{...}, \quad \hat{\gamma}_{l(j)} = \bar{y}_{.jl.} - \bar{y}_{.j..}$$

$$(\widehat{\alpha\gamma})_{il(j)} = (\bar{y}_{ijl.} - \bar{y}_{ij.} - \bar{y}_{.jl.} + \bar{y}_{.j..}).$$

The total sum of squares of observations is partitioned and, we get the following component sum of squares :

$$\begin{aligned} \sum \sum \sum \sum (y_{ijkl} - \bar{y}_{...})^2 &= qrm \sum (\bar{y}_{i...} - \bar{y}_{...})^2 + pr m \sum (\bar{y}_{.j..} - \bar{y}_{...})^2 \\ &\quad + rm \sum \sum (\bar{y}_{ij.} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{...})^2 + pm \sum \sum (\bar{y}_{.jl.} - \bar{y}_{.j..})^2 \\ &\quad + m \sum \sum \sum (\bar{y}_{ijl.} - \bar{y}_{ij.} - \bar{y}_{.jl.} + \bar{y}_{.j..})^2 + \sum \sum \sum \sum (\bar{y}_{ijkl} - \bar{y}_{ijl.})^2 \end{aligned}$$

$$SS(\text{Total}) = SS(A) + SS(B) + SS(AB) + SS(C \text{ within } B)$$

$$+ SS(C \text{ within } B \text{ in presence of } A) + SS(\text{Error})$$

$$= S_1 + S_2 + S_3 + S_4 + S_5 + S_6.$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f}$	E(MS)
A	$p - 1$	S_1	s_1	$\sigma^2 + \frac{qr m}{p-1} \sum \alpha_i^2$
B	$q - 1$	S_2	s_2	$\sigma^2 + \frac{pr m}{q-1} \sum \beta_j^2$
AB	$(p - 1)(q - 1)$	S_3	s_3	$\sigma^2 + \frac{r m}{(p-1)(q-1)} \sum \sum (\alpha\beta)_{ij}^2$
C within B	$q(r - 1)$	S_4	s_4	$\sigma^2 + \frac{p r m}{q(r-1)} \sum \sum \gamma_{l(j)}^2$
C within B in presence of A	$(p - 1)(r - 1)q$	S_5	s_5	$\sigma^2 + \frac{r m}{(p-1)(r-1)q} \sum \sum \sum (\alpha\gamma)_{il(j)}^2$
Error	$pqr(m - 1)$	S_6	s_6	σ^2
Total	$pqrm - 1$			

The main objective of the analysis is to test the significance of the hypotheses $H_0 : \alpha_i = 0$, $H_0 : \beta_j = 0$, $H_0 : (\alpha\beta)_{ij} = 0$, $H_0 : \gamma_{l(j)} = 0$ and $H_0 : (\alpha\gamma)_{il(j)} = 0$. The test statistics for these hypotheses are, respectively $F_1 = s_1/s_6$, $F_2 = s_2/s_6$, $F_3 = s_3/s_6$, $F_4 = s_4/s_6$, $F_5 = s_5/s_6$. The other analysis is similarly done as it is done in analysing fixed effect model.

If it is assumed that the model is a random effect model, then we need to find the $E(MS)$, where mean squares are $s_h^2 (h = 1, 2, \dots, 6)$. The $E(MS)$ values are shown below :

Component	RRRR i j l k p q r m	E(MS)	MS
α_i	1 q r m	$\sigma^2 + m\sigma_{\alpha\gamma}^2 + r m\sigma_{\alpha\beta}^2 + q r m\sigma_{\alpha}^2$	s_1
β_j	p 1 r m	$\sigma^2 + m\sigma_{\alpha\gamma}^2 + r m\sigma_{\alpha\beta}^2 + p r m\sigma_{\beta}^2 + p m\sigma_{\gamma}^2$	s_2
$(\alpha\beta)_{ij}$	1 1 r m	$\sigma^2 + m\sigma_{\alpha\gamma}^2 + r m\sigma_{\alpha\beta}^2$	s_3
$\gamma_{l(j)}$	p 1 1 m	$\sigma^2 + m\sigma_{\alpha\gamma}^2 + p m\sigma_{\gamma}^2$	s_4
$(\alpha\gamma)_{il(j)}$	1 1 1 m	$\sigma^2 + m\sigma_{\alpha\gamma}^2$	s_5
$e_{ijkl(k)}$	1 1 1 1	σ^2	s_6

The above table is prepared on the basis of assumptions :

- (i) $\alpha_i \sim \text{NID}(0, \sigma_{\alpha}^2)$, (ii) $\beta_j \sim \text{NID}(0, \sigma_{\beta}^2)$, (iii) $(\alpha\beta)_{ij} \sim \text{NID}(0, \sigma_{\alpha\beta}^2)$, (iv) $\gamma_{l(j)} \sim \text{NID}(0, \sigma_{\gamma}^2)$, (v) $(\alpha\gamma)_{il(j)} \sim \text{NID}(0, \sigma_{\alpha\gamma}^2)$, (vi) $e_{ijkl} \sim \text{NID}(0, \sigma^2)$, (vii) all random variables are mutually independent.

The test statistics for the significance of different variance components are to be found out as usual.

Chapter 9

Group of Experiments

9.1 Introduction

In agriculture, industry, medical science, biological science, psychology and in sociological experiment the treatments are repeated over places and/or seasons to select the best group of treatments suitable for all places and/or seasons. Due to changes in fertility levels in different places or due to changes in weather and atmosphere or due to experimental conditions or due to changes in method of cultivation, the treatments behave differently in different places and/or seasons. Therefore, from a single experiment conducted in only one place or only in one season or only in one experimental condition no decision can be made to recommend a treatment or a group of treatments as best. The repetition of the experiment over places and/or seasons is needed to select a best group of treatments suitable for all places and/or for all seasons. This involves the analysis of group of experiments. Here repeated experiment over places and/or seasons is known as group of experiments.

The analysis of group of experiments is not complicated if the experimental materials and weather conditions remain same for all experiments. Usual method of least squares can be applied to analyse the data except that two new parameters, one for place or seasonal effect and one for the interaction of places \times treatments interaction, are introduced in the model along with usual parameters in the model for a design. However, the experimental materials or experimental conditions may not be homogeneous for all experiments. The heterogeneous experimental conditions make the analysis complicated. The sources of complicated analysis are (i) experiments of unequal sizes, (ii) unequal precisions of the experiments, and (iii) instability in the behaviour of treatments over places and/or seasons.

Besides the above mentioned problems, same experiment may be conducted in different methods by different researchers. Thus, the precision of the experiments may not be same. Therefore, the combined analysis for all the data becomes complicated. The usual test statistic is affected due to heterogeneity in the experimental conditions. The sources of inaccuracy in the usual analysis of variance test are :

- (i) heterogeneity in error variances due to unstable behaviour of treatments,
- (ii) heterogeneity in error variances due to variability in blocks,
- (iii) non-normality of error variances,
- (iv) heterogeneity in weather conditions,
- (v) heterogeneity in methods of experimentation.

It has already been mentioned that usual analysis of variance technique is applied if the error variances of all experiments are found homogeneous. The analysis becomes slightly complicated if the place \times treatments interaction is found heterogeneous. The analysis is performed assuming mixed effect model, where places \times treatments interaction and effect of places are assumed random factors.

The analysis becomes complicated if the error variances are heterogeneous. However, approximate analysis of the groups of experiments is suggested by many statisticians. Important works in these fields are due to Cochran (1937, 1954), Yates and Cochran (1938), Gomes and Guimaraes (1958), Jones et al (1960), Calinski (1966), Afonja (1968), Tyagi et al. (1970), Raheja and Tyagi (1974). Bhuyan (1982) has presented three different methods which are superior than the methods available for the data. An exact test has been proposed by Bhuyan for equality of treatments effects, where treatments are stable over all places. Bhuyan also proposed a method to select stable treatments for places and/or seasons using all the information of the experiments. Here analysis of groups of experiments for a group of randomized block designs are presented.

9.2 Analysis of Groups of Randomized Block Designs

Let there be v treatments randomly allocated to b blocks of a randomized block design. Consider that the experiment is repeated in p randomly selected places. Let y_{hij} be the result of j th treatment in i th block of h th place ($h = 1, 2, \dots, p; i = 1, 2, \dots, b; j = 1, 2, \dots, v$). The model for y_{hij} observations is

$$y_{hij} = \mu + \alpha_h + \beta_{hi} + \gamma_j + (\alpha\gamma)_{hj} + e_{hij}.$$

Here μ = genral mean, α_h = effect of h th place, β_{hi} = effect of i th block within h th place, γ_j = effect of j th treatment, $(\alpha\gamma)_{hj}$ = interaction of j th treatment with h th place and e_{hij} = random error.

Assumptions for Analysis : (i) $e_{hij} \sim NID(0, \sigma^2)$, (ii) $\alpha_h \sim NID(0, \sigma_\alpha^2)$, (iii) $\beta_{hi} \sim NID(0, \sigma_\beta^2)$, (iv) $(\alpha\gamma)_{hj} \sim NID(0, \sigma_{\alpha\gamma}^2)$, (v) all random variables are mutually independent.

Restriction : $\sum \gamma_j = 0$

Since error variances are assumed homogeneous, usual analysis of variance technique can be applied for the analysis. The total sum of squares of observations can be partitioned as follows :

$$\begin{aligned} \sum \sum \sum (y_{hij} - \bar{y}_{...})^2 &= bv \sum (\bar{y}_{h..} - \bar{y}_{...})^2 + v \sum \sum (\bar{y}_{hi.} - \bar{y}_{h..})^2 \\ &+ pb \sum (\bar{y}_{...j} - \bar{y}_{...})^2 + b \sum \sum (\bar{y}_{h.j} - \bar{y}_{h..} - \bar{y}_{...j} + \bar{y}_{...})^2 \\ &+ \sum \sum \sum (y_{hij} - \bar{y}_{hi.} - \bar{y}_{h.j} + \bar{y}_{h..})^2. \end{aligned}$$

The sum of squares in right-hand and left-hand sides of the equations are distributed as $\chi^2 \sigma^2$ with $(pbv - 1)$, $(p - 1)$, $p(b - 1)$, $(v - 1)$, $(p - 1)(v - 1)$ and $p(b - 1)(v - 1)$ d.f., respectively.

ANOVA Table

Sources of Variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	E(MS)
Places	$p - 1$	S_1	s_1	$\sigma^2 + b\sigma_{\alpha\gamma}^2 + bv\sigma_\alpha^2 + v\sigma_\beta^2$
Blocks within places	$p(b - 1)$	S_2	s_2	$\sigma^2 + v\sigma_\beta^2$
Treatments	$v - 1$	S_3	s_3	$\sigma^2 + b\sigma_{\alpha\gamma}^2 + \frac{pb}{v-1} \sum \gamma_j^2$
Places \times treatments	$(p - 1)(v - 1)$	S_4	s_4	$\sigma^2 + b\sigma_{\alpha\gamma}^2$
Error	$p(b - 1)(v - 1)$	S_5	s_5	σ^2
Total	$pbv - 1$			

The main objective of the analysis is to test the significance of

$$H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_v \quad (a)$$

against H_A : at least one of the equality does not hold good. Before performing the test for the above hypothesis, we need to test the significance of $H_0 : \sigma_{\alpha\gamma}^2 = 0$ against $H_A : \sigma_{\alpha\gamma}^2 > 0$. The test statistic for this hypothesis is

$$F_1 = \frac{s_4}{s_5}$$

This F_1 is distributed as variance ratio with $(p-1)(v-1)$ and $p(b-1)(v-1)$ d.f. If this hypothesis is rejected, then the test statistic for the hypothesis (a) is

$$F_2 = \frac{s_3}{s_4}$$

This F_2 is distributed as variance ratio with $(v-1)$ and $(p-1)(v-1)$ d.f. However, if $\sigma_{\alpha\gamma}^2 = 0$, the test statistic for (a) is

$$F_3 = \frac{s_3(pb-1)(v-1)}{(p-1)(v-1)s_4 + p(b-1)(v-1)s_5}$$

9.3 Analysis when Error Variances are Heterogeneous

Let us assume that $e_{hij} \sim N(0, \sigma_h^2)$; $h = 1, 2, \dots, p$. Then the total sum of squares is found out by weighted analysis, where weights are reciprocal of error variances. Let $W_h = 1/\sigma_h^2$. Then

$$\begin{aligned} \sum \sum \sum W_h (y_{hij} - \bar{y}_{...})^2 &= bv \sum W_h (\bar{y}_{h..} - \bar{y}_{...})^2 + v \sum \sum W_h (\bar{y}_{hi.} - \bar{y}_{h..})^2 \\ &+ Wb \sum (\bar{y}_{..j} - \bar{y}_{...})^2 + b \sum \sum W_h (\bar{y}_{h.j} - \bar{y}_{h..} - \bar{y}_{..j} + \bar{y}_{...})^2 \\ &+ \sum \sum \sum W_h (\bar{y}_{hij} - \bar{y}_{hi.} - \bar{y}_{h.j} + \bar{y}_{h..})^2 \end{aligned}$$

Here $\bar{y}_{...} = \frac{\sum W_h y_{h..}}{Wbv}$, $\bar{y}_{..j} = \frac{1}{W} \sum W_h \bar{y}_{h.j}$, $W = \sum W_h$, $V(\bar{y}_{..j} - \bar{y}_{...}) = \frac{2}{Wb}$, $j \neq s = 1, 2, \dots, v$.

ANOVA Table

Sources of Variation	d.f.	$MS = \frac{SS}{d.f.}$	$E(MS)$
Places	$p-1$	s_1	$1 + \frac{W}{p-1} \left(1 - \frac{\sum W_h^2}{W^2}\right) [b\sigma_{\alpha\gamma}^2 + v\sigma_\beta^2 + bv\sigma_\alpha^2]$
Blocks within places	$p(b-1)$	s_2	$1 + \frac{Wv}{p} \sigma_\beta^2$
Treatments	$v-1$	s_3	$1 + \frac{b \sum W_h^2}{W} \sigma_{\alpha\gamma}^2 + \frac{Wb}{v-1} \sum \gamma_j^2$
Places \times treatments	$(p-1)(v-1)$	s_4	$1 + \frac{Wb}{p-1} \left(1 - \frac{\sum W_h^2}{W^2}\right) \sigma_{\alpha\gamma}^2$
Error	$p(b-1)(v-1)$	s_5	1
Total	$pbv-1$		

Here also the test statistic for $H_0 : \sigma_{\alpha\gamma}^2 = 0$ is $F = s_4/s_5$ and if this H_0 is accepted, then F -statistic will be similar to F_3 mentioned above. But, if $\sigma_{\alpha\gamma}^2 \neq 0$, the usual analysis of variance

F -statistic is not available, even if W_h is known. Here

$$k_1 E(s_3) = k_2 E(s_4) + k_3 E(s_5),$$

where $k_1 = W^2 - \sum W_h^2$, $k_2 = (p-1) \sum W_h^2$, $k_3 = W^2 - p \sum W_h^2$.

Therefore, $F_4 = \frac{k_3 s_3}{k_2 s_4 + k_3 s_5}$ to test the significance of hypothesis (a).

If W_h is known, this latter F -statistic is distributed approximately as variance ratio with $(v-1)$ and $\frac{(k_2 s_4 + k_3 s_5)^2}{\frac{(k_2 s_4)^2}{(p-1)(v-1)} + \frac{(k_3 s_5)^2}{p(b-1)(v-1)}}$ d.f.

If W_h is not known, the approximation is not satisfactory.

In practice, W_h is not known and it is estimated from h th experiment, where $\hat{W}_h = \frac{1}{\hat{\sigma}_h^2}$ ($h = 1, 2, \dots, p$). Here

$$\hat{\sigma}_h^2 = \sum_i \sum_j \frac{(y_{hij} - \bar{y}_{hi0} - \bar{y}_{h0j} - \bar{y}_{h00})^2}{(b-1)(v-1)} = MS \text{ (Error) from } h\text{th experiment.}$$

It has already been mentioned that the distribution of F_4 is approximate even if W_h is known. If W_h is not known and it is replaced by its estimate, the distribution of F_4 may not be approximate to variance ratio distribution. Therefore, the conclusion regarding treatment contrast may not be appropriate when weighted analysis is performed using estimated error variance as weight. However, if the d.f. of the individual experiment, $(b-1)(v-1)$, is large enough, then F_4 statistic will be nearer to variance ratio distribution.

Due to the weighted analysis, the estimate of treatment effect and F -statistic become biased. However, the bias of order $p/(b-1)(v-1)$ can be removed using Meier's (1953) theorem [Bhuyan (1982)]. The adjusted estimated-treatment effect is

$$\hat{\gamma}_j \text{ (adjusted)} = \frac{\sum W_h}{W} (\bar{y}_{h..j} - \bar{y}_{h..}) - \frac{2}{(b-1)(v-1)} \left[\sum \hat{\omega}_h (1 - \hat{\omega}_h) (\bar{y}_{h..j} - \bar{y}_{h..} - \hat{y}_{..j} + \bar{y}_{...}) \right].$$

Here $\hat{W}_h = \frac{W_h}{W}$, $\hat{y}_{..j} = \frac{\sum W_h \bar{y}_{h..j}}{W}$, $\hat{y}_{...} = \frac{\sum W_h \bar{y}_{h..}}{W}$.

The adjusted variance of this estimated effect and the adjusted variance of the estimated treatment difference are given by

$$V(\hat{\gamma}_j) \text{ (adjusted)} = \frac{v-1}{Wbv} \left[1 + \frac{2}{(b-1)(v-1)} \sum \hat{\omega}_h (1 - \hat{\omega}_h) \right]$$

$$V(\hat{\gamma}_j - \hat{\gamma}_s) \text{ (adjusted)} = \frac{2}{Wb} \left[1 + \frac{2}{(b-1)(v-1)} \sum \hat{\omega}_h (1 - \hat{\omega}_h) \right]$$

respectively.

The adjusted F -statistic to test the significance of $H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_v$ assuming fixed effect model is given by

$$F \text{ (adjusted)} = \frac{\hat{\omega}_v \sum_j^v (\hat{y}_{..j} - \hat{y}_{...})^2}{\frac{\sum \sum \sum \sum W_h (y_{hij} - \bar{y}_{hi.} - \bar{y}_{h.j} + \bar{y}_{h..})^2}{p(b-1)(v-1)}}$$

$$-\frac{2}{(b-1)(v-1)^2} \sum_h [2b(1-\hat{\omega}_h)W_h \sum_j \hat{y}_{..j}(\bar{y}_{h..j} - \bar{y}_{h..}) - b\hat{W}_h(1-\hat{\omega}_h) \sum_j (\hat{y}_{..j} - \hat{y}_{...})^2 + b\hat{W}_h\hat{\omega}_h \sum_j (\bar{y}_{h..j} - \bar{y}_{h..})^2]$$

Bhuyan and Miah (1989) have discussed that, if $(b-1)(v-1)$ is large, the adjusted F -statistic provides satisfactory result.

We have already observed that the weighted analysis of groups of experiments, when error variances are heterogeneous, becomes complicated and the usual weighted analysis does not provide satisfactory F -statistic (unless adjusted assuming fixed effect model) to test the significance of

$$H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_v.$$

The analysis becomes complicated due to the presence of places \times treatments interaction term in the model. However, if the treatments are stable over all places/seasons, the places \times treatments interaction term can be dropped from the model. Bhuyan (1982) has suggested to select the stable group of treatments. The stability of treatment effects is tested on the basis of work of James (1951) and Welch (1951).

Theorem : Let t_{hj} be the estimated treatment effect of j -th treatment in h -th place ($h = 1, 2, \dots, p; j = 1, 2, \dots, v$). Consider that $t_{1j}, t_{2j}, \dots, t_{pj}$ are normally and independently distributed with means $T_{1j}, T_{2j}, \dots, T_{pj}$ and variances $D_{1j}, D_{2j}, \dots, D_{pj}$, respectively. Also consider that $d_{1j}, d_{2j}, \dots, d_{pj}$ are the estimates of $D_{1j}, D_{2j}, \dots, D_{pj}$, with d.f. f_1, f_2, \dots, f_p , respectively. Here d_{hj} is distributed as χ^2 and it is independent of t_{hj} . Thus, to test the significance of

$$H_0 : T_{1j} = T_{2j} = \dots = T_{pj}, \quad (b)$$

the test statistic is

$$F = \frac{\sum W_{hj}(t_{hj} - t_j)^2 / (p-1)}{1 + \frac{2(p-2)}{p^2-1} \sum f_h \left(1 - \frac{W_{hj}}{W_j}\right)^2}.$$

This F is distributed as variance ratio to the order $\frac{1}{f_h}$. The d.f. of F is

$$(p-1) \text{ and } \left[\frac{3}{p^2-1} \sum \frac{1}{f_h} \left(1 - \frac{W_{hj}}{W_j}\right)^2 \right]^{-1}.$$

Here $W_{hj} = \frac{1}{d_{hj}}$, $W_j = \sum_h W_{hj}$, $t_j = \sum_h W_{hj}t_{hj}/W_j$,

$$d_{hj} = \frac{v-1}{bv} \hat{\sigma}_h^2, \quad \hat{\sigma}_h^2 = \frac{1}{(b-1)(v-1)} \sum \sum (y_{hij} - \bar{y}_{hi} - \bar{y}_{h..j} + \bar{y}_{h..})^2.$$

The proof of this theorem has been discussed by Welch (1951). James (1951) has proved that, if f_h is large, the statistic $\sum_h W_{hj}(t_{hj} - t_j)^2$ is distributed as χ^2 with $(p-1)$ d.f. and his statistic can be used to test the hypothesis (b). For small f_h the statistic is to be compared with

$$\chi^2 \left[1 + \frac{3\chi^2 + (p+1)}{2(p^2-1)} \sum \frac{1}{f_h} \left(1 - \frac{W_{hj}}{W_j}\right)^2 \right].$$

Here χ^2 is the table value of χ^2 distribution at $100\alpha\%$ level of significance with $(p-1)$ d.f.

The hypothesis (b) can be tested for all v treatments. The rejection of the hypothesis leads as to conclude that j -th treatment is not stable over all places. The unstable treatment is not suitable to recommend for all places/seasons and hence, such treatments can be dropped from combined analysis.

The combined analysis can be performed for stable treatments. Let there be $q \leq v$ stable treatments for all places, where the estimate of j -th treatment effect in h -th place be

$$t_{hj} = \bar{y}_{h..j} - \bar{y}_{h..}; \quad h = 1, 2, \dots, p; \quad j = 1, 2, \dots, q \leq v.$$

The variance of t_{hj} is $V(t_{hj}) = \frac{(v-1)}{bv} \sigma_h^2$ and its estimator is $v(t_{nj}) = \frac{v-1}{bv} \hat{\sigma}_h^2$. The combined treatment effect is estimated by

$$t_j = \frac{\sum W_{hj} t_{hj}}{W_j}.$$

The variance of t_j is

$$V(t_j) = \frac{1}{W_j} \left[1 + \frac{4}{W_j^2} \sum_h \frac{1}{f'_h} W_{hj} (W_j - W_{hj}) \right],$$

where $f'_h = f_h - \frac{4(p-2)}{p-1}$. The variance formula has been derived on the basis of work of Cochran and Carroll (1953) with an adjustment suggested by Meier (1953).

The analysis with stable group of treatments enable us to drop places \times treatments interaction from the model and hence, the problem in weighted analysis due to the interaction term is avoided. Moreover, due to the use of stable group of treatments for combined analysis the heterogeneity of error variances may be removed. This was observed by Bhuyan and Saha (1984). Besides this advantage, we can perform analysis with data of individual experiment. The problem is to test the significance of

$$H_0 : T_1 = T_2 = \dots = T_q, \quad q = v. \quad (c)$$

The hypothesis (c) can be tested separately with the data of p places. Let F_h ($h = 1, 2, \dots, p$) be the F -statistic for hypothesis (c) in h -th place and let P_h be the p -value of F_h -statistic. The P_h values observed in all p places can be combined to draw a single conclusion for the hypothesis (c). According to Fisher (1941),

$$Z = -2 \sum_{h=1}^p \ln P_h$$

is distributed as chi-square with $2p$ d.f. This Z -statistic is used to draw the conclusion regarding hypothesis (c). Similar Z -statistic can also be found out to test the significance of

$$H_0 : T_j = T_s \text{ against } H_A : T_j \neq T_s, \quad j \neq s = 1, 2, \dots, q.$$

We have discussed combined analysis with stable group of treatments. In practice, the treatments or most of the treatments may not be stable. In such a situation we need alternative procedure for the combined analysis. Our problem is to test the significance of

$$H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_v. \quad (d)$$

The hypothesis can alternatively formulated by

$$H_0 : A\theta_1 = A\theta_2 = \dots = A\theta_p,$$

$$\text{where } A\theta_h = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & \cdots & \cdots & -1 \end{bmatrix}_{(v-1) \times v} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_v \end{bmatrix}.$$

Let $A\hat{\theta}_h$ be the estimate of $A\theta_h$, where

$$A\hat{\theta}_h = T_h = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & \cdots & \cdots & -1 \end{bmatrix}_{(v-1) \times v} \begin{bmatrix} t_{h1} \\ t_{h2} \\ \vdots \\ t_{hv} \end{bmatrix}.$$

Here $t_{hj} = \bar{y}_{h.j} - \bar{y}_{h..}$, $V(t_{hj} - t_{hs}) = \frac{2\sigma_h^2}{b}$, $h = 1, 2, \dots, p$.

$T_h \sim \text{IN}(A\theta_h, W_h\sigma_h^2)$, $W_h = AC_h^{-1}A'$, W_h is a $(v-1) \times (v-1)$ non-singular matrix and it is similar for any generalized inverse of C_h . Then, if σ_h^2 is known, we can use the statistic :

$$t = \sum_{h=1}^p (T_h - T)' W_h^{-1} (T_h - T) / \sigma_h^2.$$

This t is distributed as χ^2 with $(p-1)(v-1)$ d.f. Here

$$T = \left[\sum_{h=1}^p (W_h\sigma_h^2)^{-1} \right]^{-1} \sum_{h=1}^p W_h^{-1} T_h / \sigma_h^2$$

is the combined estimator of treatment contrast from all places.

The proof regarding the distribution of ' t ' has been discussed by James (1954). According to him, if σ_h^2 is not known, its estimator $\hat{\sigma}_h^2$ can be used to find ' t ' statistic, where the statistic is

$$\hat{t} = \sum_{h=1}^p (T_h - \hat{T}) W_h^{-1} (T_h - \hat{T}) / \hat{\sigma}_h^2.$$

Here $\hat{T} = \left[\sum_{h=1}^p (W_h\hat{\sigma}_h^2)^{-1} \right]^{-1} \sum_{h=1}^p W_h^{-1} T_h / \hat{\sigma}_h^2$.

James has shown that, if f_h is large enough \hat{t} is distributed as χ^2 with $(p-1)(v-1)$ d.f. For smaller values of f_h ($h = 1, 2, \dots, p$) \hat{t} is to be compared with

$$k = \chi^2 \left[1 + \frac{3\chi^2 + \{(p-1)(v-1) + 2\}}{2\{(p-1)(v-1)\}\{(p-1)(v-1) + 2\}} \sum_h \frac{1}{f_h} \left\{ 1 - \frac{2\hat{\sigma}_h^2}{b} \right\}^{-1} \sum_h \left(\frac{2\hat{\sigma}_h^2}{b} \right)^{-1} \right]^2.$$

Here χ^2 is the tabulated value of χ^2 at $100\alpha\%$ level of significance with $(p-1)(v-1)$ d.f. Here

$$v(t_{hj} - t_{hs}) = \frac{2\hat{\sigma}_h}{b}, \quad j \neq s.$$

The above method of analysis is suitable to test the significance of treatment effects from combined analysis of any experiment conducted through any design. Let us explain the method of analysis for the analysis of three experiments conducted through BIB design having

parameters $b = v = 7, r = k = 3$ and $\lambda = 1$. The treatment effects estimated from experiments of 3 places are shown below :

Places	Estimates of treatment effects (t_{hj})						
	1	2	3	4	5	6	7
1	-2.1543	1.9971	-0.2743	3.1129	-3.9741	-1.2371	2.7200
2	-1.1829	-3.3743	5.0729	-5.1686	-7.8357	8.8229	3.6657
3	-0.4590	-0.8360	0.3760	0.3290	-0.3060	0.3860	-0.1010

The other analytical results are :

Places	$\hat{\sigma}_h^2$	MS(treatment)	F
1	5.05311	16.75948	3.32
2	3.14572	84.74342	26.94
3	0.24240	0.54850	2.27

The error variances $\sigma_h^2 (h = 1, 2, 3)$ are found heterogeneous by Bartlett's (1937) χ^2 test.

The estimates of T_h are :

$$T_1 = \begin{bmatrix} -4.1514 \\ -1.8800 \\ -5.2672 \\ 1.8171 \\ -0.9172 \\ -4.8743 \end{bmatrix}, T_2 = \begin{bmatrix} 2.7914 \\ -6.2557 \\ 3.9857 \\ 6.6528 \\ -10.0057 \\ -4.8486 \end{bmatrix}, T_3 = \begin{bmatrix} 0.3770 \\ -0.8350 \\ -0.7880 \\ -0.7650 \\ -0.8450 \\ -0.3580 \end{bmatrix}$$

The inverse of estimated variance-covariance matrices of T_h are :

$$v(T_1) = (W_1 \hat{\sigma}_1^2)^{-1} = \begin{bmatrix} 0.39580 & -0.06596 & \dots & -0.06596 \\ -0.06596 & 0.39580 & \dots & -0.06596 \\ \dots & \dots & \dots & \dots \\ -0.06596 & \dots & \dots & 0.39580 \end{bmatrix}_{6 \times 6}$$

$$v(T_2) = (W_2 \hat{\sigma}_2^2)^{-1} = \begin{bmatrix} 0.63579 & -0.10596 & \dots & -0.10596 \\ \dots & \dots & \dots & \dots \\ -0.10596 & -0.10596 & \dots & 0.63579 \end{bmatrix}_{6 \times 6}$$

$$v(T_3) = (W_3 \hat{\sigma}_3^2)^{-1} = \begin{bmatrix} 8.25128 & -1.37521 & \dots & -1.37521 \\ -1.37521 & 8.25128 & \dots & -1.37521 \\ \dots & \dots & \dots & \dots \\ -1.37521 & \dots & \dots & 8.25128 \end{bmatrix}_{6 \times 6}$$

Therefore, $\hat{T}' = [0.1232 \quad -1.4359 \quad -0.7992 \quad -0.4647 \quad -1.1661 \quad -1.1333]$.

The test statistic $\hat{t} = 161.04$. It is approximately distributed as χ^2 since $f_1 = f_2 = f_3 = 8$. Thus, \hat{t} is compared with $k = 42.51$ (value of χ^2 at 5% level of significance with 12 d.f.). Since $\hat{t} > k$, H_0 is rejected. The treatment contrasts are significantly different. Here

$$v(t_{hj} - t_{hs}) = \frac{2\hat{\sigma}_h^2}{rE}, j \neq s = 1, 2, \dots, v.$$

If the hypothesis $H_0 : A\theta_1 = A\theta_2 = \dots = A\theta_p$ is rejected, we need to test the significance of $H_0 : A\theta_h = \delta$ against $H_A : A\theta_h \neq \delta$, where $\delta = 0$ indicates that any contrast $[\gamma_j - \gamma_s, j \neq s]$ is insignificant. The test statistic using estimated error variances is

$$\hat{t}_1 = \hat{T}' \sum \left(\frac{W_h^{-1}}{\hat{\sigma}_h^2} \right) \hat{T}$$

This \hat{t}_1 is also approximately distributed as χ^2 with $(v - 1)$ d.f., provided f_h 's are large. For smaller values of f_h the calculated \hat{t}_1 is to be compared with

$$k_1 = \chi^2 \left[1 + \frac{3\chi^2 + (v + 1)}{2(v^2 - 1)} \sum \frac{1}{f_h} \left\{ \left(1 - \frac{2\hat{\sigma}_h^2}{b} \right)^{-1} \left(\sum_h \left(\frac{2\hat{\sigma}_h^2}{b} \right)^{-1} \right) \right\}^{-1} \right],$$

where χ^2 is the table value of chi-square at 100 α % level of significance with $(v - 1)$ d.f.

The above hypothesis indicates that all $(v - 1)$ contrasts are insignificant. In practice, we need to investigate the insignificance of any particular contrast $\gamma_j - \gamma_s, j \neq s = 1, 2, \dots, v$. For this the test statistic is

$$\hat{t}_2 = \sum_h \left(\frac{2\hat{\sigma}_h^2}{b} \right)^{-1} \hat{T}_{1h}^2 - \left\{ \sum \left(\frac{2\hat{\sigma}_h^2}{b} \right)^{-1} T_{1h} \right\}^2 \left\{ \sum \left(\frac{2\hat{\sigma}_h^2}{b} \right)^{-1} \right\}^{-1}$$

For large value of f_h, \hat{t}_2 is approximately distributed as χ^2 with $(p - 1)$ d.f. For smaller values of $f_h (h = 1, 2, \dots, p), \hat{t}_2$ is to be compared with

$$k_2 = \chi^2 \left[1 + \frac{3\chi^2 + (p + 1)}{2(p^2 - 1)} \sum_h \frac{1}{f_h} \left(1 - \frac{2\hat{\sigma}_h^2}{b} \right)^{-1} \left\{ \sum \left(\frac{2\hat{\sigma}_h^2}{b} \right)^{-1} \right\} \right],$$

where χ^2 is the table value of χ^2 at 100 α % level of significance with $(p - 1)$ d.f. Here

$$\hat{T}_{1h} = t_{hj} - t_{hs}, j \neq s = 1, 2, \dots, v.$$

For BIB design, $V(\hat{T}_{1h}) = \frac{2\hat{\sigma}_h^2}{rE}$ and for RB design, $V(\hat{T}_{1h}) = \frac{2\hat{\sigma}_h^2}{b}$.

For the example cited above the \hat{t}_2 statistics for each of $\gamma_j - \gamma_s (j \neq s = 1, 2, \dots, v)$ contrast are shown below :

Contrast	\hat{t}_2	Contrast	\hat{t}_2	Contrast	\hat{t}_2	Contrast	\hat{t}_2
$\gamma_1 - \gamma_2$	5.92	$\gamma_2 - \gamma_3$	21.57	$\gamma_3 - \gamma_5$	58.37	$\gamma_5 - \gamma_6$	95.12
$\gamma_1 - \gamma_3$	10.21	$\gamma_2 - \gamma_4$	3.02	$\gamma_3 - \gamma_6$	5.16	$\gamma_5 - \gamma_7$	57.45
$\gamma_1 - \gamma_4$	12.98	$\gamma_2 - \gamma_5$	0.76	$\gamma_3 - \gamma_7$	3.51	$\gamma_6 - \gamma_7$	12.55
$\gamma_1 - \gamma_5$	19.89	$\gamma_2 - \gamma_6$	47.55	$\gamma_4 - \gamma_5$	12.85		
$\gamma_1 - \gamma_6$	28.97	$\gamma_2 - \gamma_7$	13.73	$\gamma_4 - \gamma_6$	73.31		
$\gamma_1 - \gamma_7$	10.83	$\gamma_3 - \gamma_4$	39.62	$\gamma_4 - \gamma_7$	29.69		

The values of \hat{t}_2 are to be compared with $k_2 = 7.84$ at 5% level of significance. It is observed that except $\gamma_1 - \gamma_2, \gamma_2 - \gamma_4, \gamma_3 - \gamma_6$ and $\gamma_3 - \gamma_7$ all other contrasts are significantly different.

Example 9.1 In a poultry farm an experiment is conducted to study the impact of 5 varieties of dry food for layer hen. The foods are given to hen of 4 different ages and the

experiment is repeated over 3 different seasons. The number of eggs of each hen are recorded for analysis. The data of 3 experiments are shown below :

Data of number of eggs per hen in each season for different foods (y_{hij})

Age	Season-1						Season-2					
	Foods					Total, $y_{1..}$	Foods					Total, $y_{2..}$
	1	2	3	4	5		1	2	3	4	5	
1	100	95	110	85	85	475	95	92	100	80	80	447
2	98	90	90	86	80	444	95	87	80	82	75	419
3	105	98	112	88	90	493	100	95	100	85	85	465
4	95	90	85	80	80	430	90	85	80	85	78	418
Total Y_{hj}	398	373	397	339	335	1842	380	359	360	332	318	1749

Age	Season-3					Total $y_{3..}$
	Foods					
	1	2	3	4	5	
1	90	90	100	70	78	428
2	95	85	80	76	75	411
3	98	92	100	80	83	453
4	90	87	80	75	75	407
Total $y_{3.j}$	373	354	360	301	311	1699

Analyse the data and verify the suitability of dry food.

Solution : We have $h = 3, b = 4, v = 5, \sigma_1^2 = 276.2, \sigma_2^2 = 270.0, \sigma_3^2 = 306.7, f_1 = f_2 = f_3 = 12$. These error variances are homogeneous by Bartlett's (1937) χ^2 -test, where $\chi^2 = 0.06$. Hence, usual analysis of variance technique can be applied for the combined analysis of the data of 3 experiments. For combined data, we have

$$C.T. = 466401.67, SS (Total) = 4992.33.$$

$$SS (Ages within seasons) = \sum_{h=1}^3 \left[\frac{y_{h.}^2}{v} - \frac{y_{h..}^2}{bv} \right] = 1070.1.$$

$$SS (Food) = \frac{\sum y_{.j}^2}{pb} - C.T. = \frac{5625966}{3 \times 4} - 466401.67 = 2428.83.$$

$$SS (Seasons) = \frac{\sum y_{h..}^2}{bv} - C.T. = \frac{9338566}{4 \times 5} - 466401.67 = 526.63.$$

$$\begin{aligned} SS (Seasons \times Foods) &= \sum \sum \frac{y_{h.j}^2}{b} - C.T. - SS (Seasons) - SS (Foods) \\ &= \frac{1877884}{4} - 466401.67 - 526.63 - 2428.83 = 113.87. \end{aligned}$$

$$\begin{aligned} SS (Error) &= SS (Total) - SS (Seasons) - SS (Ages within seasons) \\ &\quad - SS (Foods) - SS (Seasons \times Foods) \\ &= 4992.33 - 526.63 - 1070.1 - 2428.83 - 113.87 = 852.9. \end{aligned}$$

ANOVA Table

Sources of variation	d.f.	SS	MS = $\frac{SS}{d.f.}$	F	$F_{0.05}$
Seasons	2	526.63	263.315	--	—
Ages within seasons	9	1070.1	118.900	—	—
Foods	4	2428.83	607.207	27.63	2.594
Seasons \times foods	8	113.87	14.234	0.60	2.216
Error	36	852.9	23.692		
Total	59				

It is noted that $H_0 : \sigma_{\alpha\gamma}^2 = 0$ is true since the F -statistic is $F = 0.60 < F_{0.05,8,36} = 2.216$. Therefore, to test the significance of

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5,$$

the test statistic is

$$F = \frac{s_3/(v-1)(pb-1)}{(p-1)(v-1)s_4 + p(b-1)(v-1)s_5} = \frac{607.207}{21.972} = 27.63.$$

Since $F = 27.63 > F_{0.05,4,44} = 2.594$, the foods are significantly different.

By Duncan's multiple range test we can compare the pairs of treatments, where the DMRT values are

$$D_k = d_{0.05} \sqrt{\frac{(p-1)(v-1)s_4 + p(b-1)(v-1)s_5}{(v-1)(pb-1)b}}, \quad k = 2, 3, 4, 5.$$

$$D_2 = 2.846 \sqrt{\frac{21.972}{4}} = 6.67, \quad D_3 = 7.05, \quad D_4 = 7.21, \quad D_5 = 7.26.$$

The means of foods are $\bar{y}_{..1} = 95.92$, $\bar{y}_{..2} = 90.50$, $\bar{y}_{..3} = 93.08$, $\bar{y}_{..4} = 81.00$, $\bar{y}_{..5} = 80.33$.

It is observed that the first 3 foods are similar and the last 2 foods are similar. But first 3 foods significantly differ from last 2 foods.

Chapter 10

Construction of Design

10.1 Introduction

The basic idea of construction of design was explained first by Sir R. A. Fisher. He contributed in the field of construction of design during 1919–30 and proposed some rules and regulations to conduct experiment. The basic 3 rules are (a) Randomisation, (b) Replication, and (c) Local Control. By local control the treatments are allocated in blocks of homogeneous plots. This in turn increases the precision of the contrast to be estimated. But if number of treatments are large, it is difficult to have large number of homogeneous plots in a block. With the increase in number of treatments, there is a chance of increasing heterogeneity in the plots of a block and hence, efficiency of the experiment is decreasing. To avoid this problem the block size can be reduced to allocate a group of selected treatments to the plots of a block and there may be more blocks in a replication of the experiment. The reduction of block size for each replication to allocate a part of treatment in a block is the technique what is known as *construction of design* and this technique provides the idea of incomplete block design.

In case of factorial experiment, if factors or level of a factor is increased, the total level combinations are increased creating a problem to have blocks of size to allocate all level combinations as treatment. As a result, all treatments are allocated in blocks of smaller size, where plots in a block are homogeneous. The allocation of treatments to blocks within a replication is done according to some principles. The basic principle of allocation of a group of treatments from all treatments in blocks is to be followed in such a way that pre-identified treatment contrast can be estimated with more efficiency. The selected group of treatments are allocated in plots of a block by a random process. The selection of group of treatments from all treatments under study and allocation of selected group of treatments to plots of a block are the aspects of design of experiment.

Depending on the allocation of selected group of treatments the construction of design is of different types. Yates (1936) has done work on the construction of design at first. The problem arisen in the proposed method of construction by Yates has been solved jointly by Fisher, Yates and Bose (1939). Later on many developments have been done in this field by Bose and Nair (1939), Nair and Rao (1942), Bose and Shimamoto (1952), Kempthorne (1953) and by many others. In this chapter, the idea of construction of balanced incomplete block design and partially balanced incomplete block design will be expressed.

10.2 Some Mathematical Concept Related to Construction of Design

Groups : Let S be the set of values a, b and c . If any operation $(*)$ is done on each pair of observations of S , then the new observations will be the unique observation of this set. Operation $(*)$ is called binary operation. Thus, if $a, b \in S$, then $a * b = g \in S$.

Let S be a set and $(*)$ is a binary operation. Consider that $E = \langle S, * \rangle$ is a mathematical rule. The mathematical rule will create group, if the following rules are valid :

Group 1 : In S there is an element e ($e \in S$) so that $e * S = x = x * e$, $x \in S$. Here e is called identity of group and it is unique.

Group 2 : For each value of x ($x \in S$) there is unique value x^{-1} ($x^{-1} \in S$) so that $x^{-1} * x = e = x * x^{-1}$. Here x^{-1} is the inverse of x .

Group 3 : The binary operation ($*$) follows associative rule. Let a , b and c be the elements in S . Then $a * b * c = (a * b) * c$.

Group 4 : The binary operation ($*$) follows commutative rule. Consider that a and b are the elements in S . Then $a * b = b * a$. The group under this rule is called abelian group.

Fields : Let F be the set of some elements and two binary operation ($+$) and (\cdot) are possible, then $f = \langle F, +, \cdot \rangle$ is called field, if the following rules are valid.

F-1 : $\langle F, + \rangle$ is the abelian group, the identity of which is denoted by zero and the inverse of $x \in F$ is denoted by $-x$.

F-2 : If $F_0 = \{x \in F/x \neq 0\}$, then $\langle F_0, \cdot \rangle$ will be abelian group.

F-3 : If in any set F there are limited elements, then F is called limited field or Galois field.

The construction of some design depends on properties of Galois field. Here some properties are discussed.

1. If any positive number N is divided by p , then the residue after division can be written equal to N . That is $R = N \pmod{p}$. Here p is the number of elements in a set and R is the value of the set under \pmod{p} .
2. If p is the prime number, then operations addition, subtraction, multiplication and division can be done on the elements of the set and the new elements obtained under the operation are the elements in the set under \pmod{p} . Let $p = 11$, then under \pmod{p} the elements will be $0, 1, 2, \dots, 10$.

Here, $5 + 6 = 11 = 0 \pmod{p}$

$$5 - 6 = (5 + 11 - 6) \pmod{p} = 10 \pmod{p}$$

$$5 \times 6 = 8 \pmod{p}$$

$$5 \div 6 = (5 + 5 \times 11) \div 6 = 10 \pmod{p}.$$

If each element in the set $\{0, 1, 2, \dots, 10\}$ is multiplied by non-zero element of the set, we have new result of the product and the products will be the elements of the set except zero. This is possible as p is a prime number. In such a situation the field generated by the elements is known as Galoi's Field. For example, if the element 4 in the set of elements $\{0, 1, 2, \dots, 10\}$ is multiplied by the elements in the set except zero, then under $\pmod{11}$, we get the elements in the set except zero. For example, $4 \times 10 = 7 \pmod{11}$. Here field is written as $GF(p = 11)$.

3. In every $GF(p)$ there is at least one observation, the value of different powers of it under \pmod{p} is the non-zero observation of $GF(p)$. Such value is called primitive root. For example, in $GF(p = 11)$, we have $2^0 = 1, 2^1 = 2, 2^2 = 4, 2^3 = 8, 2^4 = 5, 2^5 = 10, 2^6 = 9, 2^7 = 7, 2^8 = 3, 2^9 = 6$. Again, $6^0 = 1, 6^1 = 6, 6^2 = 3, 6^3 = 7, 6^4 = 9, 6^5 = 10, 6^6 = 5, 6^7 = 8, 6^8 = 4, 6^9 = 2$.

Here 2 and 6 of $GF(p = 11)$ are the two primitive roots. In this method an infinite number can also be expressed by p -value, if p is a prime number. However, if p is the power of a prime number, then a function can be considered for the number and the infinite number can be expressed by p according to the following method.

Let $S = p^n$, where p is a prime number and n is any positive integer value. Then

$$x^{p^n} - 1 = 0$$

equation can be considered and a function is also considered as follows :

$$\phi(x) = \frac{x^{p^{n-1}} - 1}{F(x)}$$

Let us consider that $F(x)$ is the greatest factor of $x^{p^n} - 1 = 0$. This $\phi(x)$ is called cyclotomic polynomial. If the function from $\phi(x)$ can be expressed as n degree polynomial, then the function is called minimum function. For example, let $S = p^n = 3^2$.

Then $\phi(x) = \frac{x^8 - 1}{x^4 - 1} = x^4 + 1$.

Again, $x^4 + 1 = (x^2 + x + 2)(x^2 + 2x + 2) \pmod{3}$.

Therefore, $(x^2 + x + 2)$ and $(x^2 + 2x + 2)$ are the minimum function of field 3^2 . Here for all powers of x up to $x^{p^{n-2}}$ the values except zero in the field 3^2 are available. For example, from the minimum function $x^2 + x + 2$, we have

$$x^0 = 1, x^1 = 1, x^2 = x + 1, x^3 = 2x + 1, x^4 = 2, x^5 = 2x, x^6 = 2x + 2, x^7 = x + 2.$$

Using these values orthogonal Latin square design can be constructed.

10.3 Construction of Incomplete Block Design Using Primitive Root

If p is a prime number and if x is the primitive root of $GF(p)$, then up to x^{p-2} the even or odd power of x can be used to form an initial block. Then adding 1 under mod p with the values of initial block symmetric BIB design can be constructed. The parameters of such BIB design are $b = v$, $r = k$ and λ .

Let us consider that $p = 11$. The primitive root of $GF(p = 11)$ is 6. Using the even powers of 6 under mod p , we have initial block as follows :

$$1 \ 3 \ 9 \ 5 \ 4.$$

Now, by adding 1 under mod 11 with all elements in the initial block, we have other block as follows :

Block	Treatments in Blocks				
1	1	3	9	5	4
2	2	4	10	6	5
3	3	5	0	7	6
4	4	6	1	8	7
5	5	7	2	9	8
6	6	8	3	10	9
7	7	9	4	0	10
8	8	10	5	1	0
9	9	0	6	2	1
10	10	1	7	3	2
11	0	2	8	4	3

For this BIB design the parameters are $b = v = 11$, $r = k = 5$ and $\lambda = 2$. This is a symmetric BIB design.

Now, if one block and its treatments are excluded from the above arrangement, we get another BIB design. Let us delete the first block and its treatments. Then

Block	Treatments in Blocks			
1	2	10	6	
2	0	7	6	
3	6	8	7	
4	7	2	8	
5	6	8	10	
6	7	0	10	
7	8	10	0	
8	0	6	2	
9	10	7	2	
10	0	2	8	

For this new design the parameters are :

$$b' = b - 1 = 10, v' = v - k = 11 - 5 = 6, r' = r = 5, k' = k - \lambda = 5 - 2 = 3, \lambda' = \lambda = 2.$$

Another design can be formed with the treatments which are not included in the above design. The new parameters are $b'' = b - 1, v'' = k, r'' = r - 1, k'' = \lambda$ and $\lambda'' = \lambda - 1$. We have not included 1, 3, 9, 5, 4. With this set of treatments the arrangements of treatments in blocks are as follows :

Block	Treatments			Block	Treatments		
1	4	5		6	4	9	
2	3	5		7	1	5	
3	1	4		8	1	9	
4	9	5		9	1	3	
5	3	9		10	3	4	

The parameters of the above design are :

$$b'' = 10, v'' = 5, r'' = 4, k'' = 2, \lambda = 1.$$

If $(4\lambda + 3)$ is a prime number or power of a prime number, then a symmetric BIB design can be constructed with parameters $b = v = 4\lambda + 3, r = k = 2\lambda + 1$ and λ . Let $\lambda = 4$, then $4\lambda + 3 = 19$, where the primitive root of 19 $[GF(19)]$ is 2. Now, using the even powers of 2 we can form an initial block and by adding 1 under mod 19 with treatment numbers. We can form a BIB design as shown below :

Block	Treatments									
1	1	4	16	7	9	17	11	6	5	
2	2	5	17	8	10	18	12	7	6	
3	3	6	18	9	11	0	13	8	7	
4	4	7	0	10	12	1	14	9	8	
5	5	8	1	11	13	2	15	10	9	
6	6	9	2	12	14	3	16	11	10	
7	7	10	3	13	15	4	17	12	11	
8	8	11	4	14	16	5	18	13	12	

Contd...

Block	Treatments									
9	9	12	5	15	17	6	0	14	13	
10	10	13	6	16	18	7	1	15	14	
11	11	14	7	17	0	8	2	16	15	
12	12	15	8	18	1	9	3	17	16	
13	13	16	9	0	2	10	4	18	17	
14	14	17	10	1	3	11	5	0	18	
15	15	18	11	2	4	12	6	1	0	
16	16	0	12	3	5	13	7	2	1	
17	17	1	13	4	6	14	8	3	2	
18	18	2	14	5	7	15	9	4	3	
19	0	3	15	6	8	16	10	5	4	

The parameters of the above design are

$$b = v = 4 \times 4 + 3 = 19; r = k = 2 \times 4 + 1 = 9; \lambda = 4.$$

If the parameters of a BIB design are b, v, r, k and λ , then if we discard the blocks in which a particular treatment is allocated, we get a PBIB design. Let us consider the BIB design with 7 treatments formed by initial block as follows :

Block	Treatments				Block	Treatments			
1	1	2	4		5	5	6	1	
2	2	3	5		6	6	0	2	
3	3	4	6		7	0	1	3	
4	4	5	0						

Now, let us delete the blocks having treatment 6. Then new blocks with treatments are as follows :

Block	Treatments			
1	1	2	4	
2	2	3	5	
3	4	5	0	
4	0	1	3	

This new design is a PBIB design having parameters.

$$b' = b - r = 7 - 3 = 4; v' = v - 1 = 7 - 1 = 6, r' = r - 1 = 3 - 1 = 2; k' = k = 3, \lambda_1 = 1, \lambda_2 = 0, n_1 = v - k = 7 - 3 = 4, n_2 = k - 2 = 3 - 2 = 1, p_{22}^2 = 0, p_{jk}^1 = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}, p_{jk}^2 = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}.$$

In some cases, instead of using all the even or odd powers of primitive root of $GF(p)$ one can use some of the even or odd powers of primitive root to form an initial block. In that case, PBIB design can be formed. For example, let us consider the 3 odd powers of the primitive root of $GF(11)$, where the primitive root is 2. The 3 odd powers of 2 under mod 11 are :

$$2 \ 8 \ 10.$$

Let us consider the initial block containing treatments 2, 8, 10 and other block contents as follows :

Block	Treatments			Block	Treatments		
1	2	8	10	7	8	3	5
2	3	9	0	8	9	4	6
3	4	10	1	9	10	5	7
4	5	0	2	10	0	6	8
5	6	1	3	11	1	7	9
6	7	2	4				

The parameters of the above PBIB design are $b = v = 11$, $r = k = 3$, $\lambda_1 = 1$, $\lambda_2 = 0$, $n_1 = 6$, $n_2 = 4$, $p_{jk}^2 = \begin{bmatrix} 3 & 3 \\ 3 & 0 \end{bmatrix}$, $p_{jk}^1 = \begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix}$.

10.4 Construction of Design from Other Design

Let there be a BIB design having parameters b, v, r, k and λ . Consider that the blocks of this design are B_1, B_2, \dots, B_b . In block B_i there are k treatments ($k < v$). Now, another design can be formed using the other treatments which are not in i th block. As an example, let us consider the following two designs, where second one is constructed with the treatments which are not included in i -th block of the first design.

First Design				Second Design				
Block	Treatments			Block	Treatments			
1	1	2	4	1	0	3	5	6
2	2	3	5	2	1	4	0	6
3	3	4	6	3	0	1	2	5
4	4	5	0	4	1	2	3	6
5	5	6	1	5	0	2	3	4
6	6	0	2	6	1	3	4	5
7	0	1	3	7	2	5	4	6

The parameters of first design are $b = v = 7$, $r = k = 3$, $\lambda = 1$.

The parameters of the second design are $b' = b = 7$, $v' = v = 7$, $r' = b - r = 7 - 3 = 4$, $k' = v - k = 7 - 3 = 4$, $\lambda' = b - 2r + \lambda = 7 - 2 \times 3 + 1 = 2$

Here second design is complementary to first design.

The complementary design can also be formed from the construction of design obtained from the system of confounding of factorial experiment. For example, let us consider that for a 2^3 factorial experiment two interactions AC and BC will be confounded in a replication. Then their generalized interaction AB is automatically confounded in the replication. Now, considering AB, AC and BC as 3 separate treatments, we can allocate 3 treatments in an incomplete block of 3 plots. For 2^3 factorial experiments we have 7 effects and interactions; these are A, B, C, AB, AC, BC and ABC . Any two of these effects and interactions can be confounded in a replication. Then their generalized interaction will automatically be confounded. We have 7 systems of confounding. The seven effects and interactions can be considered as 7 treatments and these 7 treatments are denoted by, viz., 1, 2, 3, ..., 7. For example, if A and B are confounded in one replication, then their generalized interaction AB is automatically confounded. If we denote A by treatment 1, B by treatment 2 and AB by treatment 3, then, if treatment 1, 2 and 3 are allocated in a block of 3 plots, we shall get an incomplete block. In a

similar way, we shall get 7 blocks of 3 plots. The system of confounding and the arrangement of treatments from confounded effects and interactions are shown below :

Replication	Confounded Effects and Interactions							Block	Treatments		
	A	B	AB	C	AC	BC	ABC				
1	✓	✓	✓					1	1	2	3
2	✓			✓	✓			2	1	4	5
3		✓		✓		✓		3	2	4	6
4		✓			✓		✓	4	2	5	7
5	✓					✓	✓	5	1	6	7
6			✓	✓			✓	6	3	4	7
7			✓		✓	✓		7	3	5	6

The parameters of the above BIB design constructed from the system of confounding effects and interactions are $b = v = 7$; $r = k = 3$ and $\lambda = 1$.

Now, a complementary design can be constructed using the remaining treatments which are not included in a block. The new design is shown below :

Block	Treatments	Block	Treatments
1	4, 5, 6, 7	5	2, 3, 4, 5
2	2, 3, 6, 7	6	1, 2, 5, 6
3	1, 3, 5, 7	7	1, 2, 4, 7
4	1, 3, 4, 6		

The parameters of this complementary design are $b' = v' = 7$, $r' = k' = 4$, $\lambda' = 2$.

The incomplete block design can also be constructed by dual design. For this, the block numbers in which a particular treatment is allocated can be considered as the treatments in a block. For example, in the above design, treatment 1 is allocated in block numbers 3, 4, 6, 7. So, 3, 4, 6, 7 can be considered as treatments for a block. Thus, we have a new design as follows :

Block	Treatments in Block					Block	Treatments in Block				
1	1	4	5	7		5	2	5	6	7	
2	1	3	5	6		6	2	3	4	5	
3	1	2	4	6		7	3	4	6	7	
4	1	2	3	7							

This above design is also a BIB design having parameters $b' = v' = 7 = b = v$, $r' = r = 4$, $k' = k = 4$; $\lambda' = \lambda = 2$.

The above dual design is formed from the symmetric BIB design. But, if principal design is not symmetric, the dual design will not be symmetric.

If the parameters of a BIB design are b, v, r, k and $\lambda = 1$, then the dual design will be PBIB design of parameters :

$$b' = v, v' = b, r' = k, k' = r, \lambda_1 = 1, \lambda_2 = 0, n_1 = k(r - 1), n_2 = b - 1 - k(r - 1), p_{11}^2 = k^2.$$

The above parameters are the parameters of 2 associates PBIB design.

Again, if dual design is formed from a BIB design having parameters b, v, r, k and $\lambda = 2$, then the dual design will be 3 associates PBIB design. However, for $\lambda = 2$ the dual design

may be a 2 associates PBIB design. For example, let us consider the dual design obtained from principal design having parameter $b = 10$, $v = 5$, $r = 4$, $k = 2$ and $\lambda = 1$, where the principal and dual designs are as follows :

Principal Design			Dual Design			
Block	Treatments in Blocks		Block	Treatments in Blocks		
1	3	4	1	1	3	6 10
2	2	4	2	3	7	8 9
3	0	3	3	2	5	9 10
4	4	8	4	1	2	4 7
5	2	8	5	4	5	6 8
6	3	8				
7	0	4				
8	0	8				
9	0	2				
10	3	2				

The parameters of this dual design are $b' = 5$, $v' = 10$, $r' = 2$, $k' = 4$, $\lambda_1 = 1$, $\lambda_2 = 0$, $n_1 = 6$, $n_2 = 3$, $p_{11}^3 = 4$.

10.5 Construction of Incomplete Block Design from Orthogonal Latin Square Design

Let us discuss the construction of Latin square design before the construction of incomplete block design. Let us consider that F_i ($i = 1, 2, \dots, m$) is the prime number or power of a prime number and $F = (F_1, F_2, \dots, F_m)$, where F_i ($i = 1, 2, \dots, m$) are the components of F . The value of $GF(F_i)$ including minimum function is available based on the value of F_i . For example, let $F = 12$ or $F = 3 \times 4$, where $F_1 = 3$ and $F_2 = 4$. Again, the values of $GF(3)$ and $GF(2^2)$ are 0, 1, 2 and 0, 1, α , α^2 . The minimum function of $GF(2^2)$ is $\alpha^2 + \alpha + 1$. Then the combinations of the values of $GF(3)$ and $GF(2^2)$ are as follows :

$$00, 01, 0\alpha, 0\alpha^2, 10, 11, 1\alpha, 1\alpha^2, 20, 21, 2\alpha, 2\alpha^2.$$

These 12 values of combination can be used to form a 12×12 Latin square design, where each combination is considered a treatment.

In the same way using each value from m different fields the combination of values can be done and one can get F combined values. If F is a prime number or power of a prime number, then the results of combination for $m = 1$ and $GF(F_i)$ are the values of $GF(F)$.

Let us consider that the combination of F values are for the $F \times F$ Latin square design with F treatments. The combination of F values can be arranged in F rows and F columns so that paired value of row and column can be added, where all the results of all possible added values of 2 can be presented through a table. Here the column is called *principal column* and row is called *principal row*. The table obtained from the results of addition is a Latin square design. In finding the results of addition the minimum function of $GF(F)$ is considered zero.

The values of principal column are multiplied by the values of $GF(F)$, except the values 0 and 1, to get the values of new column. At this stage two values in pairs from new column and principal column are added to get a table of new values. This gives orthogonal Latin square design. As many as orthogonal Latin square designs are available depending on the number of times the values of principal column are multiplied and table of added values are obtained. As

every time multiplication is done by new value, not more than $(S - 1)$ multiple are available, where S is the minimum factor of F . The process can be explained for 4×4 Latin square design.

Let us consider $GF(2^2)$ for 4×4 Latin square design. The values of $GF(2^2)$ are 0, 1, α and α^2 and the minimum function of this field is $\alpha^2 + \alpha + 1 = 0$. Now, the values of 0, 1, α and α^2 are to be written in principal row and principal column and step-by-step the values of row column are to be added to get the following 4×4 Latin square design.

Principal Column	Principal Row			
	0	1	α	α^2
0	0	1	α	α^2
1	1	0	α^2	α
α	α	α^2	0	1
α^2	α^2	α	1	0

Now, the elements in principal column are multiplied by α and using principal row we shall add the two elements of row and column to have the following arrangements of treatments :

Second Principal Column	0	1	α	α^2
0	0	1	α	α^2
α	α	α^2	0	1
α^2	α^2	α	1	0
1	1	0	α^2	α

Again, multiplying the elements in the principal column by α^2 and following the similar method as discussed above, the following arrangement is obtained :

Third Principal Column	0	1	α	α^2
0	0	1	α	α^2
α^2	α^2	α	1	0
1	1	0	α^2	α
α	α	α^2	0	1

The above 3 arrangements of treatments are 3 orthogonal Latin square designs.

If there are S^2 treatment, they can be arranged into $S \times S$ square. If this $S \times S$ Latin square design is orthogonal, then the BIB design of parameters $v = S^2$, $b = S^2 + S$, $k = S$, $r = S + 1$ and $\lambda = 1$ can be formed.

As the treatments are allocated in $S \times S$ square, in each block there are S treatments and there are S blocks. The number of treatments which are allocated in rows of first $S \times S$ squares can be allocated in columns to get another $S \times S$ squares. The second square generates another S blocks having S plots in each block. After that a $S \times S$ orthogonal Latin square design can be formed and on this design one can superimpose the first $S \times S$ square. Due to this arrangement of squares, there will be treatments of square with a particular treatment of Latin square. These treatments can be allocated in blocks. Thus, for S treatments of Latin square design S blocks will be available. In a similar way, blocks will be formed from the available

orthogonal Latin square design. In such arrangements of treatment of first two squares and orthogonal Latin square design a BIB design of parameters : $v = 16, b = 20, k = 4, r = 5$ and $\lambda = 1$ is formed, where $S = 4$.

If $S = 4, v = S^2 = 16; b = S^2 + S = 20, k = 4, r = 5$ and $\lambda = 1$.

The treatments in blocks are as follows :

Block	Treatments in Block				Block	Treatments in Block			
1	1	2	3	4	5	1	5	9	13
2	5	6	7	8	6	2	6	10	14
3	9	10	11	12	7	3	7	11	15
4	13	14	15	16	8	4	8	12	16

The other blocks are available from orthogonal Latin square designs as follows :

Block	Treatments in Block				Block	Treatments in Block				Block	Treatments in Block			
9	1	6	11	16	13	1	7	12	14	17	1	8	10	15
10	2	5	12	15	14	2	8	11	13	18	2	7	9	16
11	3	8	9	14	15	3	5	10	16	19	3	6	12	13
12	4	7	10	15	16	4	6	9	15	20	4	5	11	14

Using the design so constructed as above another design can be constructed with another $S + 1$ treatments, where total treatments are $S + S + 1$. In such a case, the j th treatment ($j = 1, 2, \dots, S$) of new $S + 1$ treatments can be allocated to any set of S blocks in any order and with the new $S + 1$ treatments if a new block is constructed, then $S^2 + S + 1$ blocks will be available and in each block there will be $S + 1$ plots. The parameters of the design are $b = v = S^2 + S + 1, r = k = S + 1, \lambda = 1$.

Let us consider 5 new treatments, where the treatments are 17, 18, 19, 20, 21. Then, if these 5 treatments are allocated in the above design, we get the following design :

Block	Treatments in Block					Block	Treatments in Block					Block	Treatments in Block				
1	1	2	3	4	17	8	4	8	12	16	21	15	3	5	10	16	20
2	5	6	7	8	18	9	1	6	11	16	17	16	4	6	9	15	21
3	9	10	11	12	19	10	2	5	12	15	19	17	1	8	10	15	17
4	13	14	15	16	20	11	3	8	9	14	20	18	2	7	9	16	18
5	1	5	9	13	18	12	4	7	10	15	21	19	3	6	12	13	19
6	2	6	10	14	19	13	1	7	12	14	17	20	4	5	11	14	21
7	3	7	11	15	20	14	2	8	11	13	18	21	17	18	19	20	21

The parameters of this design are $b = v = 21, r = k = 5, \lambda = 1$.

10.6 Optimum Design

It is observed that the design of a set of treatments may be different. As there may be different designs for the same set of treatment, one needs to decide which design is optimum. Kieter (1958, 1959) has indicated some discriminating features to identify the optimum design. His proposed discriminating rules are A, D, E, L, M .

Let us consider that for a set of treatments there are D designs. If there are v treatments, one may be interested to estimate the impact of all treatments or to study any contrast of the

treatments. Let us consider a function of treatment effects as

$$\theta = L\gamma,$$

where γ is the treatment effect and L is $v \times v$ matrix and v is the number of treatments. If the elements in L are so chosen that θ indicates a vector of contrast, then significance of θ can be tested. Let us consider that θ can be estimated from any of the d designs ($d \in D$). Let the variance-covariance matrix of θ be $\text{Cov}(\theta)$. Here, let us discuss the discriminating features proposed by Kieter on the basis of the variance-covariance of θ .

A-Optimum : The design $d(d \in D)$ will be *A-Optimum*, if for the unconditional and limited values of l a design $d(d \in D)$ is such that

$$\text{tr Cov}(\theta) \leq \text{tr Cov}(\theta_0).$$

Here θ_0 is estimated from $d_0(d_0 \in D)$.

D-Optimum : Let $r(L) = l \leq v - 1$. A design $d(d \in D)$ will be called *D-Optimum*, if

$$|\text{Cov}(\theta)| \leq |\text{Cov}(\theta_0)|,$$

where θ_0 is estimated from $d_0(d_0 \in D)$.

E-Optimum : Let $r(L) = l \leq v - 1$. A design $d(d \in D)$ will be called *E-Optimum*, if the maximum latent roots of $\text{Cov}(\theta)$ are such that they are equal or less than the latent roots of a design $d_0(d_0 \in D)$.

For *L-Optimum* and *M-Optimum* Kieter (1958) can be discussed.

Exercises

1. What is meant by analysis of variance? Explain the concept of design of experiment. Illustrate the basic principles of experimental design.
2. What is analysis of variance? How does it differ from regression analysis? Explain the terms related to analysis of variance with example.
3. What is the difference between analysis of variance model and regression model? Write down the assumptions of analysis of variance. Explain the basic principles of experimental design.
4. Define linear model and analysis of variance model. Explain the conditions for efficient experiment. Discuss the technique of analysis of variance.
5. Distinguish among fixed effect, mixed effect and random effect models. What is the difference between analysis of variance model and regression model? Explain the basic principles of design of experiment.
6. What do you mean by experiment and design of experiment? What are the pre-requisites of a good experiment? What are the consequences of violation of assumptions in analysis of variance?
7. Discuss the analysis of variance technique. How does analysis of variance differ from regression analysis? Explain the uses of analysis of variance.
8. Define design of experiment. What is the use of design of experiment? Explain the basic principles of design of experiment. How would you estimate the number of replication per treatment in an experiment?
9. Explain the concept of parametric function, estimable function and contrast. Show that there are exactly k estimable functions, where k is the rank of the design matrix.
State and prove Cochran's theorem. Explain its application in analysis of variance.

10. Define contrast, estimable function, parametric function. Show that for a model $Y = XB + U$, the linear estimator of $\lambda'B$ is $r'X'Y$, where $X'Xr = \lambda$.
11. What do you mean by multiple comparison? What are the different methods of multiple comparison.
12. Distinguish between regression model and analysis variance model.

For the model $Y = X\beta + U$ under usual assumption and notation, show that

- (i) $r'X'Y$ is the unbiased linear estimator of $\lambda'B$, where $X'Xr = \lambda$,
- (ii) $X\beta$ and $X'X\beta$ are estimable functions,
- (iii) $r'X'Y$ is the BLUE,
- (iv) if β is a vector of k elements, there are $(k - 1)$ orthogonal contrasts of the elements in β .

13. What do you mean by multiple classification? For a multiple classification model,

$$y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl}; \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q; \quad l = 1, 2, \dots, n_{ij},$$

show that the rank of the design matrix is $p + q - 1$. Also show that there are exactly $p + q - 1$ estimable functions of the parameters in the model.

14. Define two-way classification. Give examples of two-way classified data. Describe different steps of two-way classification.
15. Distinguish between orthogonal and non-orthogonal designs. For the model,

$$y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl}; \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q; \quad l = 1, 2, \dots, n_{ij}.$$

Show that if $\sum c_i = \sum d_j = 0$, $\sum c_i \alpha_i$ and $\sum d_j \beta_j$ are estimable functions. Describe the procedure to find the adjusted sum of squares due to $\hat{\beta}_j$ and also discuss the test of significance of $H_0 : \beta_j = 0$.

16. Define three-way classification. Describe the different steps in analysing the data of three-way classification with several observations per cell.
17. Distinguish regression model and design of experiment model. Define two-way classification. Mention some examples of two-way classified data. Show that, for a two-way classified data $E[MS (\text{Treatment})] \geq E[MS (\text{Error})]$.
18. For a model $y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijl}$, $i = 1, 2, \dots, p; j = 1, 2, \dots, q; l = 1, 2, \dots, n_{ij}$, describe the analytical procedure of the data.
19. What is meant by linear statistical model? Write down the assumptions in analysing such a model. How covariance analysis model differs from analysis of variance model and regression model?

Establish a linear model for two-way classification with r observations per cell. Find the rank of the design matrix of such a model. Write down the different steps in analysing such a model.

20. What is meant by experiment and experimental design? What are the requisites for a good experiment? Explain clearly the basic principles of experimental design.
21. Distinguish between completely randomized design and randomized block design. How would you analyse the data of randomized block design with missing observations?

22. What is Latin square design? What is the need of such a design? Distinguish among standard, conjugate and self-conjugate Latin square design.

Write down the method of construction of $k \times k$ Latin square design. Discuss the different steps in analysing data of $k \times k$ Latin square design.

When do we use treatment mean square as error mean square?

23. Compare and contrast randomized block design and Latin square design. Write down the advantages and disadvantages of randomized block design.

Find relative efficiency of LSD compared to RBD and CRD.

24. What is meant by orthogonality of design. Discuss the different steps of analysis of orthogonal Latin square design. Write down the advantages and disadvantages of Latin square design.

Show that the orthogonality of LSD is lost if one observation is missing.

25. Define Latin square design, graeco Latin square design and orthogonal Latin square design. Explain the situation where these three designs can be used. Write down the different steps of analysis of LSD. Compare CRD, RBD and LSD.

26. Define randomized block design explaining its advantages and disadvantages.

Explain covariance technique in analysing data of RBD with missing observations.

27. What is meant by efficiency of a design? Find relative efficiency of RBD compared to CRD.

Explain the method of analysis of LSD with two missing observations.

28. What is the need of repeated LSD? Explain method of analysis of p LSD.

Define orthogonal Latin square design and graeco Latin square design.

29. Distinguish between factorial experiment and non-factorial experiment. Explain main effect, simple effect, total effect, mean effect and interaction in case of factorial experiment. Write down the Yates' algorithm in analysing data of factorial experiment.

30. What do you mean by factorial experiment? explain advantages, disadvantages and uses of factorial experiment. Establish analysis of variance table for 2^4 -factorial experiment.

31. Distinguish between symmetric and asymmetric factorial experiment. Give some practical examples of asymmetrical factorial experiment.

Explain different methods of analysis of 2^3 -factorial experiment.

32. What is confounding? Write down its advantages and disadvantages.

Find block contents so that two interactions of a 2^4 -factorial experiment are confounded in one replication. Explain the method of analysis of such experimental data.

33. Distinguish between partial confounding and total confounding. Find total effects of different effects and interactions using linear equations. Use 2^5 -factorial experiment.

34. Explain method of confounding. What is the need of it? How analysis of data of confounded factorial experiment differs from that of unconfounded factorial experiment? Discuss the efficiency of unconfounded factorial experiment.

35. What do you mean by generalized interaction? Distinguish fractional replication in factorial experiment and confounded factorial experiment.

Discuss the analysis of $\frac{1}{2^2} 2^5$ -factorial experiment.

36. Explain the concept of defining contrast, generalized interaction, alias and principal blocks related to factorial experiment.

If 2^3 -factorial experiment is conducted through LSD, write down the linear model of the data. Also write down the different steps in analysing data of such an experiment.

37. What is factorial experiment? Write down its advantages and disadvantages.

Find block contents to conduct 2^5 -factorial experiment in blocks of 2^3 -plot so that 2-factor and 3-factor interactions are balanced confounded.

38. What is meant by generalized interaction? Show that when two interactions are confounded simultaneously in one replication, their generalized interaction is also confounded in that replication. Find block contents to conduct 3^3 factorial experiment so that AB_2C and ABC_2 are confounded in one replication.

Discuss the analytical procedure of such experimental data.

39. Distinguish partial confounding and total confounding. Discuss in detail the analysis of partially confounded 2^4 -factorial experiment.

40. Explain the need of fractional factorial experiment. Write down the problems in using fractional replication of factorial experiment.

Show that in conducting $\frac{1}{2}2^6$ -factorial experiment in blocks of 2^3 -plots, all main effects and 2-factor interactions cannot be estimated.

How would you analyse the data of such an experiment?

41. Distinguish symmetrical and asymmetrical factorial experiment. Is there any difference in the analysis of such two types of experimental data?

Discuss the analysis of 2×5 factorial experiment.

42. Define split-plot design. Write down its advantages, disadvantages and uses.

Discuss the analysis of data obtained from split-plot experiment.

Give a comparative study of randomized block design, split-plot design and confounded factorial experiment.

43. What is meant by split-plot design? Explain its different types. Is split-plot design an orthogonal design? Justify your answer.

Discuss the analysis of split-plot design and find its efficiency compared to RBD.

44. Define split-split-plot design and explain its analytical procedure. Find relative efficiency of this design compared to split-plot design.

Discuss the method of estimation of missing observation in split-split-plot design.

45. Explain the situation where split-plot design is used by necessity.

Suggest the model for split-plot experiment in RBD and its analysis. Find the relative efficiency of split-plot design compared to RBD.

46. Is there any similarity between split-split-plot design and split-block design? Justify your answer. Discuss the efficiency of split-split-plot design.

Find the exact test statistic to compare two sub-plot treatments in the same main plot.

47. Define split-split-plot design. Why the data of such design are not orthogonal? Explain the reason why there are two-error variances in split-plot design.

48. Explain split-plot design with its advantages and disadvantages.

Discuss the analysis of split-plot design. Find exact test statistic to compare two whole-plot treatments in the same sub-plot.

49. Explain the concept of split-block design. How does it differ from split-split-plot design? Discuss the analytical procedure of split-block design.
50. What do you mean by incomplete block design? What is the need of such a design? How would you analyse the data obtained from incomplete block design?
51. Define BIB design. Why is it so called? For a BIB design show that (i) $b \geq v$, (ii) $\lambda(v - 1) = r(k - 1)$, (iii) $b \geq v + r - 1$.

What is the need of inter-block analysis of a BIB design? Explain intra- and inter-block analysis of BIB design.

52. Define BIB design. Establish the relations among the parameters of BIB design. Discuss the advantages and disadvantages of BIB design.

Find combined intra- and inter-block estimate of treatment effect.

53. Define BIB design. What are the different types of incomplete block design? Give a comparative study of these designs. Discuss the analysis of a simple lattice design.
54. Define PBIB design. How does it differ from BIB design? Explain the association scheme of PBIB design. Discuss the analysis of PBIB design.
55. What do you mean by incomplete block design? When it becomes partially balanced? Discuss the analysis of data obtained from two-association scheme PBIB design.
56. Define BIB and PBIB designs. Prove the following relations for a BIB design with parameters b, v, r, k and λ : (i) $\lambda(v - 1) = r(k - 1)$, (ii) $b \geq v$.

Construct a BIB design with parameters $b = v = 7, r = k = 3$ and $\lambda = 1$.

57. Show that, if a block of a symmetric BIB design is dropped, the resultant design is also BIB design. Explain the intra-block analysis of BIB design.
58. What is the need of incomplete block design? Distinguish between symmetric and asymmetric BIB design. How would you estimate the combined intra- and inter-block estimate of treatment effect?
59. Define BIB design. Establish the relations of parameters of BIB design. What is the need of inter-block analysis of a BIB design?

Discuss the intra-block analysis of BIB design.

60. What do you mean by variance component analysis? When do we use this analysis? What are the different methods of variance component analysis? Write down the advantages and disadvantages of method of variance component analysis.

Discuss the different steps of variance component analysis of three-way classification.

61. Distinguish among fixed effect model, mixed effect model and random effect model.

Why do we need to use mixed effect and random effect models? Explain the assumptions for variance component analysis. How would you estimate the variance components in case of two-way classification with several observations per cell. Find the variance of your estimates. Also suggest the estimators of variance of estimates of variance components.

62. Define random effect model. Give an example where we need to establish random effect model.

Discuss the different steps in analysing random effect model for three-way classification with several observations per cell.

63. Distinguish between fixed effect model and random effect model. What is the problem in estimating variance component by analysis of variance technique.

Explain different steps in finding $E(MS)$ for both mixed effect and random effect model.

64. Define nested classification. Explain the necessity of this classification. What is the difference between balanced and unbalanced nested classification? Discuss the different steps in analysing the data of a balanced two-stage nested design.

65. What is meant by balanced nested classification. How does it differ from unbalanced nested classification? Explain the need of nested and cross classification simultaneously.

66. What is the difference between nested and cross classifications. Discuss the analytical procedure of data of a balanced three-stage nested design.

67. Distinguish between balanced and unbalanced nested classification. Explain the method of three factor nested and cross classified data. Mention some practical situations where we need nested classification.

68. What is meant by concomitant variable? What is the need of using this variable in analysis of variance?

Explain the method of analysis of randomized block design with one concomitant variable.

69. Define covariance analysis? How would you test the significance of effect of concomitant variable. Discuss the different steps in analysing data of LSD with one concomitant variable.

70. Explain in detail the application of covariance technique to analyse the data with missing observations with special reference to RBD.

71. What is meant by covariance analysis? Explain different experimental situations where we can use covariance technique to analyse the data.

Discuss covariance analysis to analyse the data of a RBD with two missing observations.

72. Define covariance analysis. How does it differ from analysis of variance? Explain the assumptions needed for covariance analysis.

How would you justify the use of one regression coefficient for all treatments?

73. What do you mean by groups of experiments? Why do you need the analysis of groups of experiments? Mention the problems in analysing groups of experiments. Explain some suitable techniques to avoid the problems in analysing groups of experiments.

74. What is meant by stable treatment? Write down the steps to select stable treatments.

How would you analyse the data of a group of randomized block designs with stable group of treatments.

75. Explain the need of groups of experiment. What are the different problems that arise in weighted analysis of data of groups of experiments. Suggest methods to overcome these problems.

76. What is meant by combined analysis? What is the need to this analysis. Explain the method of combined analysis of a group of LS designs.

77. Define group of experiment. Why place/season effects are considered random? Also mention the reason why the places \times treatments interaction is considered random. How does this interaction affect the combined analysis of a group of experiment?

References

- Allan, F. F. and Wishart, J. (1930): A method of estimating yield of missing plot in field experimental work, *Jour. Agril. Sci.*, 20, 399-406.
- Afonja, B. (1968): Analysis of a group of balanced block experiments having error variance and some treatments in common. *Biometrics*, 24, 389-400.
- Bhuyan, K. C. (1982): Problems in Analysing Groups of Experiments. Unpublished PhD Thesis, B.C.K.V., India.
- Bhuyan, K. C. and Miah, A. B. M. A. S. (1989): A group of randomized block designs, *Sankhya*, 51(3), 429-433.
- Bose, R. C. and Niar, K. R. (1939): Partially balanced incomplete block design, *Sankhya*, 4, 307-372.
- Bose, R. C. and Shimamoto, T. (1952): Classification and analysis of partially balanced incomplete block designs with two associate classes. *Jour. Amer. Stats. Assoc.*, 47, 151-184.
- Calinski, T. (1966): On the distribution of the F-type statistics in the analysis of a group of experiments. *Jour. Roy. Stats. Soc. B.*, 28, 526-542.
- Cochran, W. G. (1937): Problems arising in the analysing of a series of similar experiments. *Jour. Roy. Stats. Soc. Suppl.*, 4, 102-118.
- Cochran, W. G. (1954): The combination of estimates from different experiments. *Biometrics*, 10, 101-129.
- Cochran, W. G. and Cox, G. M. (1957): *Experimental Designs*, John Wiley and Sons, Inc., New York.
- Das, M. N. and Giri, N. C. (1986): *Designs and Analysis of Experiments*. Wiley Eastern Limited, New Delhi.
- Duncan, B. P. (1952): On the properties of multiple comparison tests. *Virginia F. Sci.*, N.P., 3, 49.
- Duncan, B. P. (1955): Multiple range and multiple F-tests, *Biometrics*, 11, 1.
- Conover, J. (1999): *Practical Nonparametric Statistics*, John Wiley and Sons Inc., New York.
- Federer, W. T. (1955): *Experimental Design: Theory and Application*, Macmillan and Co.
- Fisher, R. A. (1935): *Design and Experiments*. Oliver and Boyd, Edinburg.
- Gomes, F. P. and Guimaraes, R. F. (1958): Joint analysis of experiments in complete randomized blocks with some common treatments. *Biometrics*, 14, 521-526.
- Graybill, F. A. (1961): *An Introduction to Linear Statistical Models*, Vol-1, McGraw-Hill Book Company, Inc.
- James, G. S. (1951): The comparison of several groups of observations when the ratios of population variances are unknown. *Biometrika*, 38, 324-329.
- James, G. S. (1954): Test of linear hypothesis in univariate analysis when the ratios of the population variances are unknown. *Biometrika*, 41, 19-43.

- Kempthorne, O. (1952): *The Design and Analysis of Experiments*, John Wiley and Sons, New York.
- Montgomery, D. C. (1984): *Design and Analysis of Experiments*, John Wiley and Sons Inc., New York.
- Scheffe, H. (1959): *The Analysis of Variance*, John Wiley and Sons, New York.
- Searle, S. R. (1971): *Linear Models*, John Wiley and Sons, New York.
- Tyagi, B. N., Kathuria, O. P. and Rao, P. P. (1970): The analysis of groups of experiments involving several factors. *Jour. Indian Soc. Agril. Stat.*, 22, 27-42.
- Tukey, J. W. (1949): One degree of freedom for non-additivity, *Biometrika*, 5, 232.
- Yates, F. and Cochran, W. G. (1938): The analysis of groups of experiments. *Jour. Agril. Sci.*, 28, 556-580.

Sampling Methods

"This page is Intentionally Left Blank"

Chapter 11

Elementary Discussion on Sampling Methods

11.1 Concept of Sampling

Statistics deals with collection, tabulation, analysis and interpretation of data, where data are collected from a population according to some pre-determined objective. Data can be collected from all population units or from a representative part of population units. The method of selection of a representative part of population units is known as sampling and selected units constitute a sample. For example, a cook decides about the well-boiled of rice when prepares a pan of rice. The decision is made observing few rice from the whole pan of rice. Here all rice in the pan may be considered as population rice and few selected rice constitute the sample rice. The method of selection of rice is known as sampling. The selected sample rice leads to conclude about the well-cooked of the pan of rice. However, the decision regarding population characteristic depends on sample size and on representativeness of population units. Deming (1950) mentioned that, "Sampling is not mere substitution of a partial coverage for a total coverage. It is the science and art of controlling and measuring reliability of useful statistical information through the theory of probability."

The sampling method does not provide precise information about population characteristic if any population unit is preferred than other. For example, the weaver usually shows a part of the cloth which is well prepared. But this well-prepared part of the cloth is not the representative part of the cloth and the decision will not be well-informative if anybody observes only the well-prepared part.

11.2 Scope of Sampling

In practice, the sampling method has been utilized since long, specially to estimate the population characteristic like mean, variance, proportion of some variable or attribute. The development of the theory of sampling has been started since 1940. Since then, it is used in the following aspects :

- (i) the bureau of statistics of any country utilizes sampling method to estimate the literacy rate, employment rate, amount of service holders, to study the income distribution of people, to study the health condition of people, to estimate death rate, birth rate, etc.
- (ii) The agricultural statistics unit of the government utilizes sampling method to estimate the total production of agricultural crop and to predict the future production.
- (iii) In industrial sector, the owner of the industry utilizes it in controlling the quality of production, to estimate the demand of the production.
- (iv) The bureau of economics can utilize this to study the condition of investment, savings, loan, deposit, etc.
- (v) Sampling method is also used for opinion survey, if it is needed to formulate a new policy by the government.

- (vi) To study the effectiveness of a new discovered medicine, sampling method may be utilized to select some units on which medicine is to be applied.
- (vii) For any research of academic interest or national interest the data can be collected through sampling method.
- (viii) To estimate the cost on food, education, residence, health at national level the data can be collected through sampling method.

11.3 Important Points to be Considered During Sampling

The sampling methods are in use in our day-to-day life since long. However, the method of estimation of population characteristic as proposed by Laplace in 1783 is in use since mid-thirty of the last century. Since then attempts have been made to develop the method of sampling and to improve the estimate of parameter. The following points are to be considered in improving the estimate of parameter :

- (i) the sampling unit must be representative of the population units,
- (ii) the estimator of the parameter must be efficient and its accuracy must be measured,
- (iii) the cost of sampling and sample survey should be minimum.
- (iv) the sample survey should be completed within a short stipulated time period.

To minimize the cost and time during sample survey the following aspects should be considered :

- (a) the mode of selection of sampling units,
- (b) the number of sampling units, and (c) the mode of analysis of collected data.

All the problems discussed above can be solved if probability theory is applied in method of sampling. The probability sampling leads us to estimate the parameter efficiently and also leads us to draw conclusion efficiently with more precision.

11.4 Different Sampling Methods

The objective of sampling is to select a representative part of the population units to collect information regarding any characteristic of the population. Fisher (1935) has shown that if the sampling units are selected randomly, the estimate of error in estimating population parameter is available and hence the precision of the estimator under study can be measured. Thus, as a sampling method, we can mention two methods, viz., (a) probability sampling and (b) sampling without probability. Deming (1950) has discussed in detail about probability sampling. Since 1940 the development of probability sampling has been continuing. In this sampling each population unit is selected with a known probability. Thus, different probability sampling methods and the estimation procedure of parameter will be discussed in the following chapters. The important probability sampling techniques are (i) simple random sampling, (ii) stratified sampling, (iii) systematic sampling, (iv) cluster sampling, (v) multi-stage sampling, (vi) multi-phase sampling.

The non-probability sampling is usually known as (a) judgement sampling, and (b) quota sampling. Besides these two non-probability sampling, there are other two non-probability sampling. These are (c) Convenience sampling and (d) Snowball sampling. In all the cases the units are not selected with some pre-assigned probability and hence, the reliability of estimator cannot be estimated. The non-probability sampling is used in large scale during 1920-1930.

Judgement Sampling or Purposive Selection : In this sampling, the researcher selects population units according to his personal judgement or choice. For example, to study the family planning adoption behaviour a researcher can select a group of couples from one area

according to his personal judgement. For the same study in the same area a second researcher may select another group of couples. But no one can claim that their selected couples represent the whole of the couples of the area under investigation. In reality, a list of couples, usually known as frame, of child-bearing ages is difficult to get and hence, the researcher as a simpler tool chooses the judgement sampling. But this judgement sampling does not provide reliable information on population characteristic. The analysis of data collected through judgement sampling is also not done according to pre-determined level of precision.

The judgement sampling, though has some limitations, is not strictly prohibited or is not accepted at all. In some instances where the population units are not well defined or are not well specified or are not well located or the frame is not updated, the judgement sampling is preferred. As mentioned earlier, the judgement sampling is suitably used in selecting the couples to study the family planning adoption behaviour.

Quota Sampling : This sampling is slightly different than judgement sampling. The frame of the population units is not available or it is not up dated. However, the units are classified into several groups. For example, the couples of child-bearing ages are not listed properly for any particular region. But they may be classified according to their duration of marriage. The couples who are in the beginning of their life are more prone in adopting any of the family planning methods than those who want child during 2-5 years of duration of marriage. Again, those who have at least one child are more prone in adopting family planning methods. Therefore, the couples can be classified at least into three groups and to study the adoption behaviour of the couples they may be selected in different amounts from 3 different groups. The number of couples to be investigated from each group is pre-determined. This number is considered a quota and survey is continued until a quota is fulfilled. This type of sampling is known as quota sampling.

Since quota sampling is not a probability sampling, its reliability and efficiency of the estimator is not beyond question. In this sampling there is a scope of investigator's bias in the information and as a result its application is reduced day by day. However, if the selected units of any group are consistent with the population units of that group, the selected sample may represent the population and the analytical result will be reliable. The quota sampling is widely used in opinion survey.

The method of mail questionnaire is also used in the field of sample survey. The cost and complexity in this type of survey is reduced. If questionnaire is sent to individual who is selected with probability, the survey is equivalent to that one which is done according to probability sampling. However, in this survey method there is more chances of non-response error. The individuals may not send back the filled in questionnaire. However, to avoid this problem Hansen and Hurwitz (1946) have proposed to send questionnaire through mail as well as to investigate the unit personally. According to them a group of individuals among non-respondents are to be selected by probability sampling and they are to be investigated and interviewed face to face. There are other methods of reducing error due to non-response. These will be discussed later on.

Convenience Sampling : This is also a non-random sampling, where the up-dated frame is not available. It is similar to purposive or judgement sampling, because the researcher select some units from a population according to his convenience. It is neither probability nor judgement sampling even the units are selected at random. For example, the cost of living index number is to be calculated for middle and upper class of people. The frame of these two groups of people is not available. However, on the assumption, that middle and upper class of people have land phone in their house and from telephone directory some of the owners of the

telephone set may be selected purposively or at random for investigation. This type of selection is done as it is convenient to the researcher to have a sample of middle and upper class of people. Thus, the technique of selection of units is called convenience sampling.

This technique is also suitable for pilot survey.

Snowball Sampling : The snow is usually falls on the top of the hill which can roll to the ground. The snow which falls on the ground does not roll. The snowball begins small but becomes bigger and bigger as it rolls downhill.

The snowball sampling means an unit is observed or smaller units are observed and units are increased as study proceeds. For example, some researchers need to study the social character of drug addicted people. The population of drug addicted people cannot be identified but data are needed from them. In such a situation, if you find one or some of the drug addicted people, you can collect information from him or from them about social characters. Also one can ask the selected drug addicted people about other similar group of people whom can be met. One drug addicted people may know other such group of people. So, if one can be selected he can be helpful to select others and in this way sample size will be bigger and bigger like snowball.

This snowball sampling is a non-random sampling. However, it can be called a pseudo-random sampling like systematic sampling if first unit is selected at random from a group of units. The other units are to be selected based on the information provided to the first unit.

11.5 Advantages and Limitations of Sampling

With increased use of statistics in different disciplines the use of sampling is also increased. The statistical data are essential in any development plan in any field. One of the important sources of statistical data is sample survey. The sampling techniques are applied profitably in the following fields :

- (i) it is advantageous to collect statistical data within a short period of time and with minimum cost,
- (ii) even with a smaller sample more reliable data are collected through sample survey if the sample is selected through probability sampling,
- (iii) if the frame is not known or if the size of the population units is not known or if the population is not properly identified along with its location, it is better to use sampling methods to estimate the population parameter,
- (iv) the non-sampling error is increased in census survey when the population is large. In such a case sampling is advantageous,
- (v) the sampling is more precise if there is more variability in the observations related to characteristics under study,
- (vi) since in sample survey the units to be covered are smaller in size, well-trained personnels can be appointed easily to get reliable information.

In spite of the above advantages, the sampling methods have some limitations. These are :

- (i) if the data in sample survey are collected by inefficient investigators, the reliability of the analytical results will be lost,
- (ii) the misplanning of the survey may lead to fallacious conclusion,
- (iii) if the survey is not conducted properly by well-trained and organized personnel, the survey result may lead to misleading conclusion,
- (iv) if the sampling method is not used properly, the sample units may not be representative of the population units.

11.6 Census and Sample Survey

The complete count of the population units in any survey is known as census. It has already been mentioned that the count of all population units may create the problem of non-sampling error as large scale survey needs large organizational set-up to complete the survey. With the increase in population size we need more well-trained man-power to complete the job. In practice, the bigger is the organizational set-up, the more is the chance of untrained personnels. However, the status of survey depends on the use of the results. In population planning the population census is a must. The sample survey is not sufficient in the field of total count of population. Similar is the case with agriculture census, housing census, economic census, etc.

The sample survey is used when we need to guess the population parameter. To estimate the production of wheat in a year in the country it is necessary to count the total production produced in all agricultural plots in a year. A representative part of the plots used in producing wheat is to be selected and estimate of total production can be made using the production of those selected plots. The sample survey is a technique where smaller representative units are selected and those selected units are investigated to collect information.

It has already been mentioned that the sample survey is advantageous in collecting information with minimum cost and within short period of time. However, sample survey is not recommended if there is less chance to get reliable estimate of the population parameter from small scale investigation. In some cases, the sample survey is conducted in deciding the reliability of the result of census.

From the above discussion it is clear that both sample survey and census have same merits and demerits. These are discussed in the next section.

11.7 Merits and Demerits of Census and Sample Survey

Mahalanobis (1950), Yates (1953), Zarkovich (1961) and Lahiri (1963) have discussed the merits and demerits of census and sample survey. Cochran (1977) has classified the merits of sample survey into 5 classes. These are :

- (i) the survey is conducted with reduced cost,
- (ii) greater speed is observed in getting result,
- (iii) sample survey results are more accurate,
- (iv) the scope is greater, and
- (v) the adaptability of the result is increased.

The other merits and demerits of both census and survey have already been mentioned in the previous section.

11.8 Principal Steps in a Sample Survey

Since the adaptability of the survey result is increasing day by day in any scientific investigation, the sample survey must be well planned so that reliable estimate of population parameter is available. The following steps are to be considered to conduct a good survey.

1. **The objective of the survey should be well stated :** If the objective of the sample survey is not well mentioned, the data to be collected and the population to be covered are not decided properly. The pre-determined objective helps in preparing a well-designed questionnaire which will contain different questions consistent with the objective of the survey. Inconsistent information create the complexity in the analysis and it does not help in taking appropriate decision. The objective of the survey is formulated in such a way that it is consistent with the available resources of the survey.

2. Population to be surveyed should be identified : The size and location of the population units to be covered in any survey must be well known and well identified. For example, if a survey is needed to study the success of white revolution in milk production of an area, then all the farmers producing milk for commercial purpose will constitute the population. A part of the producers from any part of the area under study will not be sufficient to form the population.

3. Pilot survey : For efficient management of survey work, clear and concise idea about population to be surveyed is necessary. The information related to location of the population units, the mode of transport in the area, the level of communication of the units are the pre-requisites to conduct an efficient survey. In case of human population, the behaviour and the work status of the majority should be known by the investigators. These information are collected before conducting the main survey. A survey prior to the main survey to collect necessary information about population is known as pilot survey. The pilot survey also helps in estimating the cost of the survey. Also, it helps in giving training to the investigators. Questionnaire is pre-tested during pilot survey.

4. Reference of survey and period of report : The time period to conduct the survey should be consistent with the time period for which the survey result will be used. If the objective of the survey to collect economic information prior to the preparation of budget of fiscal year 2016–17 (say), the economic survey should be completed before formulation of the budget. The suitable time period for such a budget is the end of financial year 2015–16. The report of the survey should also be prepared before formulation of the budget so that reference can be made for any component of the budget. Thus, the reference period of survey should be well mentioned, since the survey results are used in different policy making activities.

5. Decision on nature of information to be collected : The decision regarding information which are to be collected through sample survey or census must be well decided before conducting the survey. Accordingly, the questionnaire should contain the questions related to the variables for which data are sought. For example, if data are needed on total expenditure of a family in household expenditure survey, there should be questions related to family expenditure either by a single question or by main components of expenditure. But, if detailed information regarding different components of expenditure are needed, questions should be set up in that way. The collected information should be consistent with the objective of the survey.

In large scale survey, it is better to collect more information as per as possible so that those information can be used in any future research activities. However, the intensive data collection should be planned within the budgetary limit of the survey. The investigator who will use the survey result or who is responsible for the analysis must be consulted at the planning stage of the survey so that questions on required variables are included in the questionnaire.

6. Method of data collection : Before conducting the survey the method of data collection is to be decided. Data can be collected through personnel interview by the investigator or it can be collected by mail questionnaire method. These are the mode of primary data collection.

In some instances, primary data may not be needed. Secondary data from official publication or from the publication of some previous survey report are sufficient for a survey. In such cases, the method of data collection is decided according to the objective of the survey.

7. Preparation of frame : The complete list of population units is known as frame. It is necessary to select the sampling units through any probability sampling method. Frame is also constructed according to the objective of the survey. For example, if the objective of the

survey is to estimate the per hectare cost of cultivation of maize, the population units are the farmers who cultivate maize during a particular period. Again, if the objective is to estimate the agriculture production of a crop by crop-cutting system, then there is no need of frame of agricultural land.

The frame should be updated and complete. However, the frame may be defective due to the following causes :

(a) inaccuracy, (b) incomplete, (c) duplication of units, (d) inadequate, (e) out of data. Frame should be prepared so that all the above defects are avoided.

8. Choice of sample design : It has already been defined that a sample is a representative part of population units. The selected population units are called sampling units. The sampling units are usually selected by probability sampling. There are different probability sampling schemes, viz., simple random sampling, stratified sampling, cluster sampling, systematic sampling, two-stage sampling, three-stage sampling, double sampling, etc. The sample is selected using any of the sampling schemes. But the scheme is decided in such a way that it is consistent with the objective of the survey and with the resources available for the survey.

9. Training of personnel engaged in survey work : This is an important component of the survey work at the planning stage. Many people are involved in survey work. The persons who are supposed to collect information from sampling units are usually called enumerators. There are supervisors who are responsible to supervise the field work of the enumerators. Both the groups must be well trained so that data are collected accurately. If data are collected by enumerators who are not well trained, their personal bias or ignorance may distort the collected information. The inaccurate data will not be fruitful to draw a valid conclusion corresponding to the objective of the survey.

10. Preparation of questionnaire : The schedule which is used to collect information on different aspects of the sampling units is called questionnaire. The questionnaire should be simple and comprehensive. The simple and comprehensive questionnaire is helpful to the sampling units if he wishes to fill in the questionnaire himself.

The questionnaire which is to be used in the survey must be pre-tested so that data are collected without any hindrance. The questions related religious sentiment, or personal belief should be avoided.

11. Different stages of analysis of data : The collected data are analysed to prepare report according to the objective of the survey. For proper and unbiased analysis the analytical stages are classified as follows :

- (a) data should be edited and scrutinised,
- (b) data are tabulated, and
- (c) statistical analysis of data is performed.

The data must be edited to avoid the missing information and the inconsistency in information. If necessary, the incorrect and inconsistent information are recollected from the sampling unit. If it is not possible, the inconsistent information of any sampling unit is dropped from the analysis. However, the dropping of any information may create the problem of non-response.

The tabulation of data is necessary for any particular pre-determined analytical procedure. The mode of analysis becomes easier in case of tabulated information.

Tabulation of data is a preliminary step of analysis. Detailed statistical analysis is done according to the objective of the survey.

12. **Preparation of report :** The analytical results are reported to infer about the population characteristics. The final survey report is presented in chapters. The contents of the report should include :

- (a) Introduction and objective of the survey
- (b) Method of data collection and analysis
- (c) Results
- (d) Accuracy of analysis
- (e) Expenditure
- (f) Future scope of research
- (g) Merits and demerits of the survey
- (h) Summary of analytical results
- (i) Reference of research
- (j) Appendix containing questionnaire, sample and map of area covered, if any, in the survey.

Report can also be presented according to the UNO (1949) instruction.

11.9 Sampling Error and Precision

Let the parameter to be estimated for any population be θ and the estimate of θ be $\hat{\theta}$. The estimate is derived from the sample observations. If the sample observations represent the population observations, the sample estimate is, usually, unbiased [$E(\hat{\theta}) = \theta$], but it is not exactly equal to the parameter. The discrepancy in estimate and parameter is known as sampling error and it is estimated by [$\hat{\theta} - E(\hat{\theta})$]. This error arises since we do not find the value of θ from population observations. The average value of sampling error depends on sample size, sampling method and method of estimation.

The sampling error [s.e.] is defined by

$$\text{s.e.} = \hat{\theta} - E(\hat{\theta}) = \hat{\theta} - \theta, \quad \because E(\hat{\theta}) = \theta, \text{ if } \hat{\theta} \text{ is unbiased.}$$

$$\therefore \theta = \hat{\theta} - \text{s.e.}$$

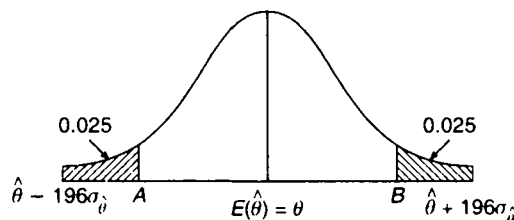
It is clear that if θ is known the value of s.e. can be calculated. In practice, $\hat{\theta}$ is unknown and s.e. is not calculated. However, $\hat{\theta}$ follows same sampling distribution, where mean of $\hat{\theta}$ is $E(\hat{\theta})$. This mean of $\hat{\theta}$ can be used to estimate the confidence interval of θ .

Let us consider that $\hat{\theta}$ follows normal distribution with mean $E(\hat{\theta}) = \theta$ and variance of $\hat{\theta}$ is $\sigma_{\hat{\theta}}^2$.

$$\text{Then } P \left[-1.96 < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < 1.96 \right] = 0.95 \quad \text{or, } P \left[-1.96\sigma_{\hat{\theta}} < \hat{\theta} - \theta < 1.96\sigma_{\hat{\theta}} \right] = 0.95$$

$$\text{or, } P[\hat{\theta} - 1.96\sigma_{\hat{\theta}} < \theta < \hat{\theta} + 1.96\sigma_{\hat{\theta}}] = 0.95.$$

The result can be shown graphically as below :



The graph shows that $\hat{\theta}$ will lie between A and B with probability 0.95 and with this probability $|\hat{\theta} - \theta| \leq 1.96\sigma_{\hat{\theta}}$. The maximum value of $|\hat{\theta} - \theta|$ is $1.96\sigma_{\hat{\theta}}$, where $|\hat{\theta} - \theta|$ is the sampling error and $1.96\sigma_{\hat{\theta}}$ is the half of the confidence interval.

The above analytical results indicate that $\hat{\theta} - \theta$ is not only the difference of estimate and parameter, it indicates the maximum variation in $\hat{\theta}$ and θ in case of repeated samples. This amount of $\hat{\theta} - \theta$ is the amount of precision of the estimate. If $\hat{\theta} - \theta$ is estimated from a sample, it is called sampling error and if it is estimated from repeated samples and if confidence limit is related to it, it is called precision. Since $1.96\sigma_{\hat{\theta}}$ is the maximum value of $|\hat{\theta} - \theta|$, the precision is estimated by $\sigma_{\hat{\theta}}[V(\hat{\theta})]$.

The $V(\hat{\theta})$ is also used to compare the precision of two or more estimates. The reliability of estimate is increased if $V(\hat{\theta})$ is decreased.

11.10 Reliability

The precision of the estimate is measured by $|\hat{\theta} - \theta| = 1.96\sigma_{\hat{\theta}}$, when the confidence interval is calculated with probability 0.95. For 90% confidence interval $|\hat{\theta} - \theta| = 1.64\sigma_{\hat{\theta}}$. We have already mentioned that reliability of the estimate is increased if $V(\hat{\theta})$ is decreased and it is observed that reliability is related with a coefficient, where this coefficient is 1.96 for 95% confidence interval and 1.64 for 90% confidence interval. The value $\hat{\theta} - \theta$ is known as precision and 95% reliability of precision is half of the confidence limit.

Therefore, $|\hat{\theta} - \theta| = \frac{1}{2}$ confidence limit = $1.96\sigma_{\hat{\theta}}$, for 95% confidence limit.

We can write, $d = z\sigma_{\hat{\theta}}$, where d = precision, z = reliability coefficient and $\sigma_{\hat{\theta}}$ is the variance of $\hat{\theta}$.

11.11 Determination of Sample Size

It has already been mentioned that the parameter θ of a population is estimated by $\hat{\theta}$ from sample observations. Whatever be the size of sample, if it is not equal to population size, there is a discrepancy between $\hat{\theta}$ and θ . This discrepancy depends on sample size. The discrepancy between parameter and estimate is known as sampling error and it also depends on the variability in the population observations. More variation in observations results in more value in $V(\hat{\theta})$. However, the sample size is to be estimated in such a way that the accuracy of the estimate is increased and the precision of $\hat{\theta} - \theta$ is decreased.

Let us consider that we have a population of size N and we need to select a sample of size n ($n \leq N$) to estimate the population mean \bar{X} , where the sample mean is $\bar{x} = \frac{1}{n} \sum x_i$. The variance of \bar{x} is $\left(\frac{\sigma^2}{n}\right)$, where $v(x) = \sigma^2$. Then the precision of the estimate \bar{x} is given by

$$\begin{aligned} d &= z\sqrt{V(\bar{x})}, \text{ where } z \text{ is the confidence coefficient} \\ &= z\frac{\sigma}{\sqrt{n}} \end{aligned}$$

$$\text{or, } |\bar{x} - \bar{X}| = z\sigma/\sqrt{n}.$$

This implies that $\bar{x} - \bar{X}$ will be minimum, if n is large.

Let us consider that we need a sample to estimate the average birth-weight (in lb) of newborn babies in a hospital within a year. Assume that the variance of birth-weight of babies is 0.25 lb^2 . If confidence coefficient z is taken as 2, then

$$d = 2 \frac{0.5}{\sqrt{n}} \quad (95\% \text{ value of normal variable is approximately } 2)$$

and for $n = 100$, $d = 0.1$ lb. Thus, with 95% reliability the estimate of precision is ± 0.1 . Now, for different values of n , d can be estimated as follows :

$$n = 150, \quad d = \frac{2 \times 0.5}{\sqrt{150}} = 0.08$$

$$n = 200, \quad d = \frac{2 \times 0.5}{\sqrt{200}} = 0.07$$

$$n = 400, \quad d = \frac{2 \times 0.5}{\sqrt{400}} = 0.05.$$

It is noted that with the increase in the value of n , the precision is decreased and \bar{x} tends to \bar{X} . Now, for a certain value of precision we can estimate the value of n as follows :

Let the precision be $d = |\bar{x} - \bar{X}| = 0.2$ lb.

$$\text{Then } d = z \frac{\sigma}{\sqrt{n}}$$

$$\text{or, } 0.2 = \frac{2 \times 0.5}{\sqrt{n}} \quad \text{or, } n = 25.$$

Thus, we have $n = \frac{(z\sigma)^2}{d^2}$, where $V(x) = \sigma^2$.

Here z is used assuming the distribution of \bar{x} as normal. The distribution of \bar{x} may follow Student's t -distribution. In that case

$$n = \frac{(t_{\frac{\alpha}{2}} \sigma)^2}{d^2},$$

where $t_{\frac{\alpha}{2}}$ is the tabulated value of Student's t to construct $100(1 - \alpha)\%$ confidence limit for \bar{X} .

We have considered the determination of sample size to estimate population mean \bar{X} . The parameter to be estimated may be of proportion P of any characteristic. Let P be the proportion of newborn babies who by birth are affected by jaundice. Consider that p is the sample estimate of P , where $V(p) = \frac{PQ}{n}$. In such a case, the value of n is

$$n = \frac{z^2 PQ}{d^2}.$$

If it is assumed that 60% babies are affected by jaundice and we need to estimate the proportion of jaundice affected babies with precision 0.05, then

$$n = \frac{2^2 \times 0.6 \times 0.4}{(0.05)^2} = 384.$$

This method of estimation of n is used if sample is selected by simple random sampling. The method of estimation of sample size becomes complicated if multi-stage sampling plan is used.

Example 11.1 : In a college, there are 4000 students. The average educational expenditure of these students per month are to be estimated in such a way that the discrepancy in the average

estimate and the population average does not exceed Rs. 10.00. Find the value of sample size n so that the average is estimated with 95% reliability. What will be the sample size, if average is sought with 99% reliability?

Solution : Given $d = |\bar{x} - \bar{X}| \leq \text{Rs. } 10.00$.

We know, $n = \frac{(z\sigma)^2}{d^2}$.

If the variance of expenditure per month of the students is Rs. 4900.00

$$n = \frac{(1.96)^2 \times 4900}{(10)^2} = 188.$$

If estimate is sought with 99% reliability, then

$$n = \frac{(2.57)^2 4900}{(10)^2} = 324.$$

If variance of expenditure of students is Rs. 10000.00, then for 95% reliability,

$$n = \frac{(1.96)^2 10000}{(10)^2} = 384.$$

For 99% reliability,

$$n = \frac{(2.57)^2 10000}{(10)^2} = 660.$$

It is noted that the sample needs to be increased if the variability in the population observations increases. Also n needs to be increased if the level of reliability increases.

11.12 Non-Sampling Error

The different types of errors that creep in survey results during data collection and data analysis are known as non-sampling errors. The sources of non-sampling errors are :

- (a) Error in sample selection.
- (b) Failure to collect information from some sampling units.
- (c) Error in reporting.
- (d) Non-response error.
- (e) Error in pre-analysis of data.

Error in sample selection : The sample is a representative part of population units. If sampling units are selected purposively or if it is selected to represent some units having a special character or if it is selected in such a way that only sampling units possess the character under study, the sample will not be a representative part of the population and due to the above type of sampling the error will creep in the estimate. However, random selection of sampling units avoids this type of problem.

Failure to collect information from some sampling units : The sampling unit may be located but information from it may not be collected due to its absence during survey period or due to refusal to provide information or due to ignorance of any question. In case of absence of any unit it may be revisited, otherwise due to lack of information from such type of unit the error will creep in the estimate. The failure of collection of information occurs mostly in case of mail questionnaire.

Error in reporting : The respondent may provide wrong information due to ignorance or willingly. If questions regarding income or expenditure are asked, the respondents may provide downward or upward bias information in respective cases. Information on age may not be reported properly and accurately. If measuring devices are not used properly, the information on height, weight are not recorded correctly. These latter source of error is due to ignorance of the respondents. The experienced enumerator is capable to collect correct information in such cases.

Non-response error : The respondents may refuse to be interviewed, specially in surveys related to social, economical and political aspects. They may refuse to provide information if there are questions against their belief and custom. The answer to questions related to politics is avoided by many respondents. Mail questionnaire may not be returned by the respondents. If the respondents refuse to provide information in any of the above mentioned ways, the error creeps in the estimate.

Error in pre-analysis of data : The collected data are edited, scrutinized or coded before these are used for analysis. Any type of mistake in editorial work or in scrutinizing or in coding work may distort the analysis. This type of error arises due to lack of seriousness of the researcher or the supervisor responsible for analytical work.

11.13 Bias

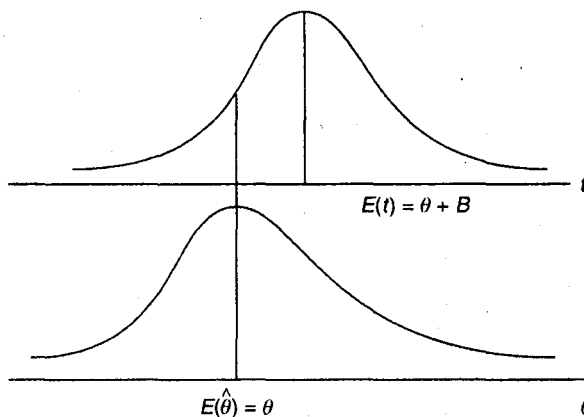
The survey is conducted to estimate the parameter, say θ . Let us consider that $\hat{\theta}$ is the estimate of θ . This estimate is unbiased if $E(\hat{\theta}) = \theta$. Let us consider another estimator of θ , say t . If $E(t) \neq \theta$, t is called *biased estimator* of θ . The amount, if bias, is measured by

$$B = E(t) - \theta = E(t - \theta).$$

Let us consider that the variance of $\hat{\theta}$ is $V(\hat{\theta})$. The mean square of error of t is given by

$$\begin{aligned} \text{MSE}(t) &= E(t - \theta)^2 = E[\{t - E(t)\} + \{E(t) - \theta\}]^2 \\ &= E[t - E(t)]^2 + E[E(t) - \theta]^2 + 2E[t - E(t)][E(t) - \theta] \\ &= V(t) + B^2. \end{aligned}$$

Here t is biased estimator of θ . This biased estimator may be preferred, if B is smaller. Sometimes $V(t)$ may be less than $V(\hat{\theta})$. Let us explain the facts by graph as follows :



It is observed that $\hat{\theta}$ is more dispersed than t . In such a case if B^2 is small t may be preferred than θ , if any such t is available.

Chapter 12

Simple Random Sampling

12.1 Definition and Estimation of Parameter

The simplest and widely used method of sampling is simple random sampling if frame is known. Let there be N units in population and we need to draw a sample of size n ($n \leq N$) from this population. The probable samples of size n are $\binom{N}{n}$. If the probabilities of selecting each of the samples are equal, then the sampling method is known as *simple random sampling*.

The units of a simple random sample are drawn one by one. To select the units two methods are followed. These are : (i) each unit is selected replacing the preceding selected one in the population (SRWR), (ii) each unit is selected without replacing the previous or any unit selected before (SRWOR). In both the methods the units are selected until required n units are included in the sample. For example, let us consider that we have a population of size $N = 3$, where the values of the characteristic under study are 2, 4, 6. We need to select a sample of size $n = 2$. The selected sample observations are :

Sample observations without replacement		Sample observations with replacement			
Sl. No.	Observations	Sl.No.	Observations	Sl.No.	Observations
1	2,4	1	2,2	7	4,2
2	2,6	2	2,4	8	6,2
3	4,6	3	4,4	9	6,4
		4	2,6		
		5	6,6		
		6	4,6		

In the former case, the number of probable samples is 3 and in the latter case this number is 9. If the probability of selection of any of the sample is $\frac{1}{3}$ in the first case and $\frac{1}{9}$ in the second case, the sample is called simple random sample and the method of drawing sample is known as simple random sampling (SRS).

Method of selection of sample : The sample is to be selected at random. The random selection is done by two methods. These are (i) Lottery method, (ii) Use of random number table.

Lottery method : It has already been mentioned that the SRS is used if frame is available. The serial number in the list of population units is written in a piece of paper. Serial numbers of all units are written in pieces of paper of same size and same colour. These pieces of paper are folded separately and put in a box so that all pieces can be well mixed. The pieces of paper are folded in such a way that the number written in it is not seen in any way. For a population of size N there will be N pieces of paper. If we need to select a sample of size n ($n \leq N$), n pieces of paper are to be taken out of the box one by one. At each step of selection, the remaining

pieces are well mixed. The units with serial number written in the selected pieces of paper constitute the sampling unit and the sample is known as a simple random sample.

In lottery method of selection, no population units is preferred during selection procedure. However, in case of large population size the pieces of paper are more and the person involved in taking out of pieces may try to take a piece from a particular part of the box where pieces are kept. This personal attitude of the researcher creates the problem of bias in sample selection.

Use of random number table : The random number table is one in which the digits from 0 to 9 are arranged in rectangular form. The digits are so arranged that using digits of any row or column, any number of required digit can be formed. The most widely used random numbers are presented in

- (i) Tippets Random Number Table,
- (ii) Tables of Random Sampling Numbers—by Kendall and Smith (1954).
- (iii) Statistical Tables for Biological, Agricultural and Medical Research—by R. A. Fisher and F. Yates.
- (iv) Tracts for Computers : Tables of Random Numbers—by M. G. Kendall and B. Babington Smith.

As an example, one of such table is given in appendix. At present random numbers are generated using computer programming. Random number is also available in scientific calculator.

The population units are listed giving serial number against each unit. For example, if $N = 100$, the units are numbered by 00, 01, 02, ..., 99. Let us consider that out of $N = 100$ units a random sample of $n = 10$ units are to be selected. This N is of two-digit number and hence, random number of two digits is to be chosen using any row or any column of the random number table. For a sample of size $n = 10$, ten two-digit numbers are to be selected (i) without selecting a number more than once (if sampling is WOR), (ii) selecting the same number repeatedly (if sampling is WR), if it is observed again and again. In some cases, the selected number is more than the number of units in the population, in that case, the random number is dropped, or subtracting N repeatedly from the selected number the residue is considered as selected random number. For example, let us consider that $N = 50$ and the serial numbers of the units are 01, 02, ..., 50. The first selected random number from the above given random number table is 51 (using first two columns of the first row). To select the sample this number 51 may be dropped, or subtracting $N = 50$ from this number we have residue = $51 - 50 = 01$. Therefore, the unit bearing serial number 01 is to be selected in the sample.

Method of estimation of parameters : Let there be N units in a population. The values of the characteristic under study of these N units are y_1, y_2, \dots, y_N and the values of the characteristic under study of the n sample units are y_1, y_2, \dots, y_n .

Population	Sample
Population total, $Y = y_1 + y_2 + \dots + y_N$ $= \sum_{i=1}^N y_i = N\bar{Y}.$	Sample total, $y = y_1 + y_2 + \dots + y_n$ $= \sum_{i=1}^n y_i = n\bar{y}.$
Population mean, $\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i.$	Sample mean, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$

Population variance :	Sample variance :
$S^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2$, if N is finite.	$s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$.
$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{Y})^2$, if N is large.	

It is assumed that a random sample of size n is drawn without replacement from the population. The objective is to estimate the parameters \bar{Y} and S^2 .

Theorem : If a simple random sample of size n is drawn without replacement from a population of size N , then sample mean is an unbiased estimator of population mean.

Proof : Let the population observations be y_1, y_2, \dots, y_N . The population mean is $\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$. The sample mean is $\bar{y} = \frac{1}{n} \sum y_i$. We need to prove that $E(\bar{y}) = \bar{Y}$.

$$\text{We know, } E(\bar{y}) = E\left[\frac{1}{n} \sum_{i=1}^n y_i\right] = \frac{1}{n} \sum_{i=1}^n E(y_i) = \frac{1}{n} \sum_{i=1}^n \left[\sum_{j=1}^N p_j y_j\right],$$

since p_i is the probability of selection of y_i from the population. In simple random sampling $p_i = \frac{1}{N}$ for all $i = 1, 2, \dots, N$. Therefore,

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{N} \sum_{j=1}^N y_j\right) = \frac{1}{n} \sum_{i=1}^n \bar{Y} = \bar{Y}.$$

Corollary : If \bar{y} is an unbiased estimator of \bar{Y} , then the unbiased estimator of population total Y is $\hat{Y} = N\bar{y}$.

We have $E(\hat{Y}) = E(N\bar{y}) = NE(\bar{y}) = N\bar{Y} = Y$.

Theorem : If a simple random sample of size n is drawn without replacement from a population of size N , then the sample variance s^2 is an unbiased estimator of population variance S^2 .

Proof : Let the population observations be y_1, y_2, \dots, y_N . Then sample variance is $s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$ and the population variance is $S^2 = \frac{1}{N-1} \sum (y_i - \bar{Y})^2$. We need to prove that $E(s^2) = S^2$.

$$\text{We have } s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2 = \frac{1}{n-1} \sum [(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2$$

$$= \frac{1}{n-1} \left[\sum (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right].$$

$$E(s^2) = \frac{1}{n-1} \left[\sum E(y_i - \bar{Y})^2 - nE(\bar{y} - \bar{Y})^2 \right] = \frac{1}{n-1} [n\sigma^2 - nV(\bar{y})]$$

$$= \frac{1}{n-1} \left[n \frac{N-1}{n} S^2 - n \frac{N-n}{Nn} S^2 \right], \quad \because \sigma^2 = \frac{N-1}{N} S^2, \quad V(\bar{y}) = \frac{N-n}{Nn} S^2$$

$$= \frac{nS^2}{n-1} \frac{nN - n - N + n}{Nn} = S^2.$$

Theorem : If a simple random sample of size n is drawn without replacement from a population of size N , then the variance of sample mean \bar{y} is given by

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

Proof : Let y_1, y_2, \dots, y_N be population observations, where $S^2 = \frac{1}{N-1} \sum (y_i - \bar{Y})^2$ and $\bar{y} = \frac{1}{n} \sum^n y_i$. We need to show that

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

$$\begin{aligned} \text{We have } V(\bar{y}) &= E(\bar{y} - \bar{Y})^2 = E \left[\frac{\sum^n y_i}{n} - \bar{Y} \right]^2 \\ &= \frac{1}{n^2} E \left[\sum^n y_i - n\bar{Y} \right]^2 \\ &= \frac{1}{n^2} E[(y_1 - \bar{Y}) + (y_2 - \bar{Y}) + \dots + (y_n - \bar{Y})]^2 \\ &= \frac{1}{n^2} E \left[\sum^n (y_i - \bar{Y})^2 + \sum_{i \neq j}^n (y_i - \bar{Y})(y_j - \bar{Y}) \right] \\ &= \frac{1}{n^2} \left[\sum^n \sigma^2 + E \sum_{i \neq j}^n (y_i - \bar{Y})(y_j - \bar{Y}) \right], \quad \because E(y_i - \bar{Y})^2 = \sigma^2 \\ &= \frac{1}{n^2} \left[\sum^n \frac{N-1}{N} S^2 + \sum_{i \neq j}^n E(y_i - \bar{Y})(y_j - \bar{Y}) \right]. \end{aligned}$$

$$\text{Now, } E(y_i - \bar{Y})(y_j - \bar{Y}) = \frac{1}{N(N-1)} \sum_{i \neq j}^N (y_i - \bar{Y})(y_j - \bar{Y}),$$

since the probability of selecting y_j after y_i without replacement is $\frac{1}{N(N-1)}$. Again,

$$\begin{aligned} \frac{1}{N(N-1)} \sum_{i \neq j}^N (y_i - \bar{Y})(y_j - \bar{Y}) &= \frac{1}{N(N-1)} \left[\left\{ \sum_i^N (y_i - \bar{Y}) \right\}^2 - \sum (y_i - \bar{Y})^2 \right] \\ &= -\frac{(N-1)S^2}{N(N-1)}, \quad \because \sum (y_i - \bar{Y}) = 0 \\ &= -\frac{S^2}{N} = -\frac{N\sigma^2}{N(N-1)} = -\frac{\sigma^2}{N-1}. \end{aligned}$$

$$\begin{aligned} \text{Therefore, } V(\bar{y}) &= \frac{1}{n^2} \left[\frac{n(N-1)}{N} S^2 - \sum_{i \neq j}^N \frac{S^2}{N} \right] \\ &= \frac{1}{n^2} \left[\frac{n(N-1)S^2}{N} - \frac{n(n-1)}{N} S^2 \right] = \frac{N-n}{Nn} S^2. \end{aligned}$$

$$\begin{aligned} \text{Again, } V(\bar{y}) &= \frac{1}{n^2} \left[\sum^n \sigma^2 - \sum_{i \neq j} \frac{\sigma^2}{N-1} \right] = \frac{\sigma^2}{n^2} \left[n - \frac{n(n-1)}{N-1} \right] \\ &= \frac{N-n}{n(N-1)} \sigma^2 = \frac{1 - \frac{n}{N}}{n \left(1 - \frac{1}{N}\right)} \sigma^2 = \frac{1-f}{1 - \frac{1}{N}} \frac{\sigma^2}{n} \\ &= \frac{\sigma^2}{n}, \text{ if finite population correction (f.p.c.) } f = \frac{n}{N} \text{ is neglected.} \end{aligned}$$

$$\text{Again, } V(\bar{y}) = \frac{1 - \frac{n}{N}}{n} S^2 = (1-f) \frac{S^2}{n} = \frac{S^2}{n}, \text{ if f.p.c. is neglected.}$$

Here $f = \frac{n}{N}$ is also called sampling fraction.

Corollary : The variance of the estimator of population total in simple random sampling is

$$V(\hat{Y}) = V(N\bar{y}) = N^2 \frac{N-n}{Nn} S^2 = \frac{N(1-f)}{n} S^2 = \frac{NS^2}{n}, \text{ if f.p.c. is neglected.}$$

Corollary : The standard error of estimator of population mean in simple random sampling is given by

$$\text{S.E.}(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{\frac{N-n}{Nn} S^2} = \sqrt{(1-f) \frac{S^2}{n}} = \frac{S}{\sqrt{n}}, \text{ if f.p.c. is neglected.}$$

Corollary : The standard error of estimator of population total in simple random sampling is given by

$$\text{S.E.}(\hat{Y}) = \sqrt{V(\hat{Y})} = \sqrt{\frac{N(1-f)}{n} S^2} = \sqrt{\frac{NS^2}{n}}, \text{ if f.p.c. is neglected.}$$

Corollary : In simple random sampling with replacement the standard error of sample mean is given by

$$\text{S.E.}(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{\frac{N-n}{n(N-1)} \sigma^2} = \sqrt{\frac{1-f}{1 - \frac{1}{N}} \frac{\sigma^2}{n}} = \frac{\sigma}{n}, \text{ if f.p.c. is neglected.}$$

Corollary : In simple random sampling with replacement the standard error of estimator of population total is given by

$$\text{S.E.}(Y) = \sqrt{V(\hat{Y})} = \sqrt{\frac{N^2(N-n)}{n(N-1)} \sigma^2} = \sqrt{\frac{N^2(1-f)}{n \left(1 - \frac{1}{N}\right)} \sigma^2} = \frac{N\sigma}{\sqrt{n}}, \text{ if f.p.c. is neglected.}$$

Corollary : In simple random sampling with replacement the estimator of variance of sample mean is given by

$$v(\bar{y}) = \frac{N-n}{Nn} s^2 = \frac{1-f}{n} s^2 = \frac{s^2}{n}, \text{ if f.p.c. is neglected.}$$

$$\text{Here } E[v(\bar{y})] = \frac{N-n}{Nn} E(s^2) = \frac{N-n}{Nn} S^2, \because E(s^2) = S^2.$$

Corollary : In simple random sampling without replacement the estimator of variance of the estimator of population total is given by

$$v(\hat{Y}) = \frac{N^2(N-n)}{Nn} s^2 = (1-f) \frac{N^2 s^2}{n} = \frac{N^2 s^2}{n}, \text{ if f.p.c. is neglected.}$$

Corollary : In simple random sampling without replacement the estimator of standard error of the estimators of population means and population total are given, respectively by

$$\text{s.e.}(\bar{y}) = \sqrt{\frac{N-n}{Nn} s^2} = \sqrt{(1-f) \frac{s^2}{n}} = \frac{s}{\sqrt{n}}, \text{ if f.p.c. is neglected.}$$

$$\text{s.e.}(\hat{Y}) = \sqrt{\frac{N^2(N-n)s^2}{Nn}} = \sqrt{(1-f) \frac{N^2 s^2}{n}} = \frac{Ns}{\sqrt{n}}, \text{ if f.p.c. is neglected.}$$

Confidence interval of population mean : Let y_1, y_2, \dots, y_N observations follow normal distribution. Then $\bar{y}_i, i = 1, 2, \dots, N_{c_n}$ are NID $(\bar{Y}, \frac{N-n}{Nn} S^2)$, where $V(\bar{y})$ is estimated by

$$v(\bar{y}) = (1-f) \frac{s^2}{n}.$$

Then $100(1-\alpha)\%$ confidence interval of \bar{Y} is given by

$$\bar{y} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{(1-f)s^2}{n}}.$$

Here $\hat{Y}_L = \bar{y} - t_{\frac{\alpha}{2}, n-1} \sqrt{(1-f) \frac{s^2}{n}}$ and $\hat{Y}_U = \bar{y} + t_{\frac{\alpha}{2}, n-1} \sqrt{(1-f) \frac{s^2}{n}}$.

In a similar way, the $100(1-\alpha)\%$ confidence interval of Y is written as

$$\hat{Y} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{(1-f) \frac{N^2 s^2}{n}},$$

where $\hat{Y}_L = \hat{Y} - t_{\frac{\alpha}{2}, n-1} \sqrt{(1-f) \frac{N^2 s^2}{n}}$ and $\hat{Y}_U = \hat{Y} + t_{\frac{\alpha}{2}, n-1} \sqrt{(1-f) \frac{N^2 s^2}{n}}$.

Example 12.1 : Assume that in a population there are $n = 3$ units. The values of the characteristic under study of these three units are 4, 8, and 6.

- (i) Draw all possible simple random samples of size $n = 2$ (a) with replacement, (b) without replacement.
- (ii) Show that sample mean is an unbiased estimator of population mean.
- (iii) Show that sample variance is an unbiased estimator of population variance.
- (iv) Find $V(\bar{y}), V(\hat{Y}), \text{S.E.}(\bar{y}), \text{S.E.}(\hat{Y})$.
- (v) Estimate $V(\bar{y}), V(\hat{Y}), \text{S.E.}(\bar{y}), \text{S.E.}(\hat{Y})$.
- (vi) Estimate 95% confidence interval for \bar{Y} .

Solution : (i) The selected sample observations, sample means, sample variances, probability of selection are shown below :

Sample information								
Without replacement (a)				With replacement (b)				
Sl. No.	Observations	\bar{y}_i	$s_i^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$	Observations	\bar{y}_i	s_i^2	Probability of selection for	
							(a)	(b)
1	4, 8	6	8	4, 4	4	0	$\frac{1}{3}$	$\frac{1}{9}$
2	4, 6	5	2	4, 8	6	8	$\frac{1}{3}$	$\frac{1}{9}$

Contd...

Table contd...

Sample information								
Without replacement (a)				With replacement (b)				
Sl. No.	Observations	\bar{y}_i	$s_i^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$	Observations	\bar{y}_i	s_i^2	Probability of selection for	
							(a)	(b)
3	8, 6	7	2	8, 4	6	8	$\frac{1}{3}$	$\frac{1}{9}$
4	—	—	—	6, 6	6	0	—	$\frac{1}{9}$
5	—	—	—	6, 8	7	2	—	$\frac{1}{9}$
6	—	—	—	8, 8	8	0	—	$\frac{1}{9}$
7	—	—	—	8, 6	7	2	—	$\frac{1}{9}$
8	—	—	—	4, 6	5	2	—	$\frac{1}{9}$
9	—	—	—	6, 4	5	2	—	$\frac{1}{9}$

$$(ii) \text{ The population mean, } \bar{Y} = \frac{1}{N} \sum Y_i = \frac{4+8+6}{3} = 6$$

$$E(\bar{y}) = \sum_{i=1}^3 p_i \bar{y}_i = \frac{1}{3}(6+5+7) = 6 = \bar{Y}.$$

∴ sample mean is an unbiased estimator of population mean.

The distribution of \bar{y}_i (in case of with-replacement sampling)

\bar{y}_i	4	5	6	7	8
$P(\bar{y}_i) = p_i$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

$$E(\bar{y}) = \sum p_i \bar{y}_i = 4 \times \frac{1}{9} + 5 \times \frac{2}{9} + 6 \times \frac{3}{9} + 7 \times \frac{2}{9} + 8 \times \frac{1}{9} = 6.$$

The population variance is

$$\begin{aligned} S^2 &= \frac{1}{N-1} \sum (y_i - \bar{Y})^2 = \frac{1}{N-1} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{N} \right] \\ &= \frac{1}{3-1} [116 - 108] = 4. \end{aligned}$$

$$E(s^2) = \sum p_i s_i^2 = \frac{1}{3}(8+2+2) = 4 = S^2.$$

Hence, sample variance is an unbiased estimator of population variance.

Again, for sampling with replacement

$$E(s^2) = \sum p_i s_i^2 = 0 \times \frac{3}{9} + 2 \times \frac{4}{9} + 8 \times \frac{2}{9} = \frac{8}{3}.$$

The sampling with replacement is considered as sampling from infinite population and in that case the population variance is

$$\sigma^2 = \frac{1}{N} \sum (y_i - \bar{Y})^2 = \frac{1}{N} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{N} \right] = \frac{1}{3} \left[116 - \frac{(18)^2}{3} \right] = \frac{8}{3}.$$

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

The sample variance is an unbiased estimator of population variance.

$$(iv) V(\bar{y}) = \frac{N-n}{Nm} S^2 = \frac{3-2}{3 \times 2} \times 4 \quad (\because S^2 = 4 \text{ in case of sampling without replacement}).$$

$$= 0.67.$$

$$V(\hat{Y}) = N^2 V(\bar{y}) = 3^2 \times \frac{6}{4} = 13.5$$

$$\text{S.E.}(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{0.67} = 0.82$$

$$\text{S.E.}(\hat{Y}) = \sqrt{V(\hat{Y})} = \sqrt{13.5} = 3.67.$$

(v) Estimate of $V(\bar{y})$, $V(\hat{Y})$ are, respectively :

$$v(\bar{y}) = \frac{N-n}{Nn} s^2 = \frac{3-2}{3 \times 2} \times 8 \quad [\text{using } s_1^2 = 8 \text{ from sampling without replacement}]$$

$$= 1.33.$$

$$v(\hat{Y}) = N^2 v(\bar{y}) = 3^2 \times \frac{8}{6} = 12.0.$$

Also, we have

$$\text{s.e.}(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{1.33} = 1.15, \quad \text{s.e.}(\hat{Y}) = \sqrt{v(\hat{Y})} = \sqrt{12.0} = 3.46.$$

(vi) 95% confidence interval is given by $\bar{y} \pm t_{0.025,1} \sqrt{\text{s.e.}(\bar{y})}$.

$$\text{Thus, } \hat{Y}_L = \bar{y} - 12.706 \times 1.15 \quad \hat{Y}_U = \bar{y} + t_{0.025,1} \text{s.e.}(\bar{y})$$

$$= 6 - 14.61 = -8.61. \quad = 6 + 12.706 \times 1.15 = 20.61.$$

The $V(\bar{y})$ is also calculated from sampling with replacement information, where

$$V(\bar{y}) = \frac{\sigma^2}{n} = \frac{\frac{8}{3}}{2} = \frac{8}{6} = 1.33.$$

[Here sampling with replacement is equivalent to sampling from infinite population].

$V(\bar{y})$ can also be calculated from the probability distribution of \bar{y} , where $E(\bar{y}) = 6$.

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

$$E(\bar{y}^2) = \sum \bar{y}^2 P(\bar{y}) = 4^2 \times \frac{1}{9} + 5^2 \times \frac{2}{9} + 6^2 \times \frac{3}{9} + 7^2 \times \frac{2}{9} + 8^2 \times \frac{1}{9} = \frac{336}{9}.$$

$$V(\bar{y}) = E(\bar{y}^2) - [E(\bar{y})]^2 = \frac{336}{9} - (6)^2 = 1.33.$$

$$\therefore V(\hat{Y}) = N^2 V(\bar{y}) = N^2 \frac{\sigma^2}{n} = 3^2 \times \frac{8}{6} = 12.$$

Example 12.2 : The following data represent the net area sown (in '000 hectares) in different financial year since 1950-51 to 1993-94. The area sown are presented serially.

Sl. No.	Area sown	Sl. No.	Area sown	Sl. No.	Area sown	Sl. No.	Area sown	Sl. No.	Area sown
01	118746	10	132939	19	137313	28	141953	37	139578
02	119400	11	133199	20	138772	29	142981	38	134085
03	123442	12	135399	21	140267	30	138903	39	141891
04	126806	13	136341	22	139721	31	140002	40	142339
05	127845	14	136483	23	137144	32	141928	41	142999
06	129156	15	133120	24	142416	33	140220	42	141632
07	130848	16	136198	25	137791	34	142841	43	142645
08	129080	17	137232	26	141652	35	140892	44	142095
09	131828	18	139876	27	139476	36	140901		

[Source : Agricultural statistics at a glance, Ministry of Agriculture].

- Select a simple random sample of size $n = 10$ years
- Estimate the average net area sown per year.
- Estimate the variance of the estimated net area sown.
- Estimate the total area sown during study period.
- Estimate the variance of the estimate of total area sown.
- Find 95% confidence interval of the average area sown.

Solution : (i) We have $N = 44$, $n = 10$. To select 10 years ten random numbers without replacement are selected. Vide random number table in appendix.

Selected random number	07	16	04	39	11	06	20
Selected year	07	16	04	39	11	06	20
Area sown, y_i	130848	136198	126806	141891	133199	129156	138772
Selected random number	28	10	41				
Selected year	28	10	41				
Area sown, y_i	141953	132939	142999				

(ii) The estimate of average net area sown is given by

$$\bar{y} = \frac{1}{n} \sum y_i = \frac{1354761}{10} = 135476.1 \text{ ('000 hectares)}$$

(iii) The estimate of variance of estimated average net area sown is given by

$$\begin{aligned} v(\bar{y}) &= \frac{N-n}{Nn} s^2, \text{ where } s^2 = \frac{1}{n-1} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right] = \frac{299234965}{9} = 33248329.44 \\ &= \frac{44-10}{44 \times 10} \times 33248329.44 \\ &= 2569189.093 \text{ ('000 hectares)}^2 \end{aligned}$$

$$\text{s.e. } (\bar{y}) = \sqrt{v(\bar{y})} = 1602.87 \text{ ('000 hectares).}$$

(iv) The estimate of total area sown during study period is given by

$$\hat{Y} = N\bar{y} = 44 \times 135476.1 = 5960948.4 \text{ ('000 hectares)}$$

(v) The estimate of variance of the estimated total area sown is given by

$$v(\hat{Y}) = N^2 v(\bar{y}) = 44^2 \times 2569189.093 = 4973950084 \text{ ('000 hectares)}^2.$$

(vi) 95% confidence interval for mean area sown is given by $\bar{y} \pm t_{0.025, 98} \text{s.e.}(\bar{y})$, where

$$\hat{Y}_L = \bar{y} - t_{0.025, 98} \text{s.e.}(\bar{y}) = 135476.1 - 2.262 \times 1602.87 = 131850.41 \text{ ('000 hectares)}$$

$$\hat{Y}_U = \bar{y} + t_{0.025, 98} \text{s.e.}(\bar{y}) = 135476.1 + 2.262 \times 1602.87 = 139101.79 \text{ ('000 hectares)}.$$

12.2 Estimation of Proportion in Case of Simple Random Sampling

If the variable under study is qualitative in nature, the parameter to be estimated is the population proportion or the variance of the estimated proportion. For example, let us consider that in a population there are N units and the values of the variable under study are y_i , where $y_i = 1$, if the characteristic is present in i -th unit or $y_i = 0$, if the unit does not possess the character. The characteristic may be HIV positive among the patients in a hospital, or family planning adoption among couples living in an area, or first division in B Sc (Hons.) examination among students of a university, or affected by yellow fever, etc.

Let A be the number of units in the population possessing the character under study.

Then $A = \sum_{i=1}^N y_i$ and $N - A =$ the units who do not possess the character. The proportion of units possessing the character is

$$P = \frac{A}{N} = \frac{1}{N} \sum_{i=1}^N y_i.$$

Then $Q = \frac{N - A}{N} = 1 - P =$ proportion of units in the population who do not possess the character. The problem is to estimate P and to estimate the variance of the estimated P .

Let a sample of size n be drawn. Then the sample proportion is given by

$$p = \frac{1}{n} \sum_{i=1}^n y_i = \frac{a}{n},$$

where $a =$ number of units in the sample possessing the character under study. We have

$$q = \frac{n - a}{n} = 1 - p$$

as proportion of units not possessing the character.

Theorem : In simple random sampling, p is an unbiased estimator of P .

Proof : We know that the sample mean $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is an unbiased estimator of $\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$. That is

$$E(\bar{y}) = \bar{Y}.$$

Here $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{a}{n} = p$ and $\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i = P$.

$\therefore E(p) = P$.

Corollary : The population total units possessing the character under study is A and its unbiased estimator is $\hat{A} = Np$,

since $E(\hat{A}) = NE(p) = NP = A$.

$$\begin{aligned} \text{We have } S^2 &= \frac{1}{N-1} \sum (y_i - \bar{Y})^2 = \frac{1}{N-1} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{N} \right] \\ &= \frac{1}{N-1} \left[A - \frac{A^2}{N} \right] = \frac{A}{N-1} \left[1 - \frac{A}{N} \right] = \frac{NPQ}{N-1}. \end{aligned}$$

Also, we have

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum (y_i - \bar{y})^2 = \frac{1}{n-1} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right] \\ &= \frac{1}{n-1} \left[a - \frac{a^2}{n} \right] = \frac{a}{n-1} \left[1 - \frac{a}{n} \right] = \frac{npq}{n-1}, \text{ where } q = 1 - p = 1 - \frac{a}{n}. \end{aligned}$$

Theorem : In simple random sampling without replacement the variance of the sample proportion p is given by

$$V(p) = (1-f) \frac{NPQ}{n(N-1)}.$$

Proof : We know $V(\bar{y}) = \frac{N-n}{Nn} S^2$, where in estimating proportion $S^2 = \frac{NPQ}{N-1}$ and $p = \frac{1}{n} \sum y_i = \frac{a}{n}$. Therefore,

$$V(p) = \frac{N-n}{Nn} \frac{NPQ}{N-1} = (1-f) \frac{NPQ}{(N-1)n} = \frac{NPQ}{n(N-1)}, \text{ if f.p.c. is neglected.}$$

Theorem : In simple random sampling with replacement the variance of the sample proportion is given by

$$V(p) = \frac{N-n}{N-1} \frac{PQ}{n} = \frac{NPQ}{n(N-1)} (1-f).$$

Proof : In simple random sampling with replacement, $V(\bar{y}) = \frac{N-n}{(N-1)n} \sigma^2$, where

$$\sigma^2 = \frac{1}{N} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{N} \right].$$

In case of estimation of proportion, we have

$$\sigma^2 = \frac{1}{N} \left[A - \frac{A^2}{N} \right] = \frac{A}{N} \left[1 - \frac{A}{N} \right] = PQ.$$

$$\therefore V(p) = \frac{N-n}{n(N-1)} PQ, \text{ where } p = \bar{y} = \frac{1}{n} \sum y_i = \frac{a}{n} = \frac{NPQ(1-f)}{n(N-1)}.$$

Corollary : In simple random sampling without replacement the variance of the estimator of population total is given by

$$V(\hat{A}) = \frac{N^2(N-n)}{n(N-1)} PQ.$$

$$\begin{aligned}
 V(\hat{A}) &= V(Np) = N^2V(p) = \frac{N^2(N-n)}{N(N-1)} \frac{NPQ}{n} \\
 &= \frac{N^2(N-n)}{n(N-1)} PQ = \frac{N^3(1-f)}{n(N-1)} PQ \\
 &= \frac{N^2(1-f)}{n} PQ, \text{ if } N \text{ is large.}
 \end{aligned}$$

Theorem : In simple random sampling without replacement $\frac{npq}{n-1}$ is an unbiased estimator of $\frac{NPQ}{N-1}$.

Proof : We know that $E(s^2) = S^2$, where

$$s^2 = \frac{1}{n-1} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right] = \frac{npq}{n-1}, \text{ when } y_i = 1 \text{ or } 0.$$

$$\text{Also } S^2 = \frac{1}{N-1} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{N} \right] = \frac{NPQ}{N-1}.$$

$$\therefore E\left(\frac{npq}{n-1}\right) = \frac{NPQ}{N-1}.$$

Corollary : In simple random sampling without replacement the unbiased estimator of $V(p)$ is given by

$$v(p) = (1-f) \frac{pq}{n-1} = \frac{pq}{n-1}, \text{ if f.p.c. is neglected.}$$

$$\text{We have } V(p) = \frac{N-n}{n(N-1)} PQ. \text{ Again, } E\left[\frac{npq}{n-1}\right] = \frac{NPQ}{N-1}.$$

$$\therefore E[v(p)] = E\left[\frac{N-n}{N(n-1)} pq\right] = \frac{N-n}{Nn(N-1)} NPQ = \frac{N-n}{n(N-1)} PQ = V(p).$$

Corollary : In simple random sampling with replacement the unbiased estimator of $V(p)$ is given by

$$v(p) = \frac{pq}{n-1}.$$

Sampling with replacement is equivalent to sampling from infinite population and hence f.p.c. is neglected. Therefore,

$$v(p) = \frac{np}{n-1}.$$

Corollary : In simple random sampling without replacement the unbiased estimator of $V(\hat{A})$ is given by

$$v(\hat{A}) = \frac{N(N-n)}{n-1} pq = (1-f) \frac{N^2 pq}{n-1} = \frac{N^2 pq}{n-1}, \text{ if f.p.c. is neglected.}$$

Corollary : In simple random sampling without replacement the estimator of standard error of sample proportion is given by

$$\text{s.e.}(p) = \sqrt{(1-f) \frac{pq}{n-1}}.$$

Corollary : In simple random sampling without replacement the estimator of standard error of \hat{A} is given by

$$\text{s.e.}(\hat{A}) = \sqrt{(1-f) \frac{N^2 pq}{n-1}}$$

Corollary : In simple random sampling without replacement the $100(1-\alpha)\%$ confidence interval of population proportion P is given by

$$p \pm Z_{\frac{\alpha}{2}} \text{s.e.}(p).$$

Thus, we have

$$\hat{P}_L = p - Z_{\frac{\alpha}{2}} \text{s.e.}(p) \quad \text{and} \quad \hat{P}_U = p + Z_{\frac{\alpha}{2}} \text{s.e.}(p).$$

If $\alpha = 0.05$, $Z_{0.025} = 1.96$.

Example 12.3 : The following data represent the birth-weight (in lb) of some new born babies born in a hospital.

SL.No. of baby	Birth-weight	SL. No. of baby	Birth-weight	SL. No. of baby	Birth-weight	SL. No. of baby	Birth-weight	SL. No. of baby	Birth-weight
01	6.5	27	7.3	53	6.6	79	6.8	105	5.8
02	5.8	28	6.1	54	6.0	80	6.7	106	5.9
03	6.2	29	5.5	55	5.8	81	7.2	107	6.3
04	6.8	30	5.2	56	7.2	82	6.6	108	6.6
05	7.2	31	6.1	57	7.0	83	6.8	109	6.7
06	5.0	32	6.4	58	7.1	84	6.4	110	6.8
07	5.2	33	6.6	59	6.9	85	5.8	111	6.9
08	5.1	34	7.0	60	5.2	86	7.2	112	7.2
09	5.2	35	6.6	61	5.8	87	7.1	113	7.5
10	5.2	36	6.0	62	5.5	88	5.9	114	6.1
11	5.6	37	6.2	63	5.7	89	5.9	115	6.2
12	5.8	38	7.1	64	5.5	90	6.4	116	6.4
13	6.1	39	7.0	65	7.2	91	6.6	117	6.6
14	5.2	40	5.1	66	6.2	92	6.0	118	6.8
15	6.4	41	5.2	67	6.8	93	5.5	119	6.2
16	6.3	42	6.4	68	5.8	94	5.8	120	5.2
17	7.0	43	6.8	69	5.9	95	6.0	121	5.4
18	7.1	44	7.0	70	7.2	96	6.2	122	5.8
19	7.0	45	7.5	71	7.4	97	6.4	123	7.0
20	6.8	46	7.1	72	5.8	98	6.8	124	6.8
21	5.2	47	6.8	73	5.5	99	6.6	125	6.6
22	6.8	48	7.0	74	6.4	100	7.2	126	6.9
23	5.5	49	6.6	75	6.8	101	7.5	127	7.2
24	5.8	50	6.5	76	7.5	102	7.0	128	7.0
25	7.0	51	5.8	77	7.1	103	5.5	129	7.0
26	7.2	52	6.4	78	6.9	104	5.8	130	7.0

SL.No. of baby	Birth weight	SL. No. of baby	Birth weight	SL. No. of baby	Birth weight	SL. No. of baby	Birth weight	SL. No. of baby	Birth weight
131	7.1	145	6.8	159	6.3	173	6.4	187	5.9
132	6.6	146	6.9	160	6.4	174	7.0	188	6.5
133	6.2	147	6.8	161	6.2	175	7.1	189	6.6
134	6.4	148	6.9	162	6.2	176	6.2	190	6.2
135	6.8	149	5.8	163	6.8	177	6.8	191	6.0
136	6.2	150	5.5	164	5.8	178	6.0	192	5.8
137	6.1	151	6.0	165	5.9	179	6.2	193	5.6
138	6.0	152	6.2	166	6.0	180	6.4	194	5.7
139	6.2	153	6.8	167	5.7	181	6.2	195	5.8
140	6.5	154	6.7	168	7.2	182	6.0	196	5.9
141	7.0	155	6.6	169	7.4	183	6.0	197	7.2
142	7.2	156	6.0	170	7.0	184	5.8	198	7.3
143	6.2	157	6.4	171	7.0	185	5.5	199	6.8
144	6.4	158	6.2	172	7.2	186	5.8	200	6.2

- (i) Draw a simple random sample of 20 babies and estimate the proportion of babies with weight more than 6.5 lb.
- (ii) Estimate the standard error of your estimate.
- (iii) Find 95% confidence interval for the population proportion of babies having weight more than 6.5 lb.
- (iv) Estimate the total number of babies having weight more than 6.5 lb.
- (v) Estimate the standard error of the estimator of total number of babies having weight more than 6.5 lb.
- (vi) Estimate the average weight of babies.
- (vii) Find 95% confidence interval for the population total weight of babies.

Solution : (i) The simple random sample of $n = 20$ observations is drawn using random number table given in appendix taking first three columns of the table.

Random number	114	161	081	037	153	111	146	042	121	164	100	053
Body weight, y_i	6.1	6.2	7.2	6.2	6.8	6.9	6.9	6.4	5.4	5.8	7.2	6.6
$x_i = y_i > 6.5$	0	0	1	0	1	1	1	0	0	0	1	1
Random number	040	167	190	127	115	175	095	016				
Body weight, y_i	5.1	5.7	6.2	7.2	6.2	7.1	6.0	6.3				
$x_i = y_i > 6.5$	0	0	0	1	0	1	0	0				

$x_i = 0$, if $y_i \leq 6.5$, $x_i = 1$, if $y_i > 6.5$.

The estimate of proportion of babies having weight more than 6.5 lb is given by

$$p = \frac{a}{n} = \frac{1}{n} \sum x_i = \frac{8}{20} = 0.4.$$

(ii) We have $n = 20$, $N = 200$. The estimate of standard error of p is

$$\text{s.e.}(p) = \sqrt{v(p)}, \text{ where } v(p) = (1-f) \frac{pq}{n-1} = \left(1 - \frac{20}{200}\right) \frac{0.4 \times 0.6}{20-1} = 0.011368.$$

$$\therefore \text{s.e.}(p) = \sqrt{v(p)} = \sqrt{0.011368} = 0.1066.$$

(iii) The 95% confidence interval of population proportion P is given by

$$p \pm Z_{0.025} \text{s.e.}(p), \text{ where } \hat{P}_L = p - Z_{0.025} \text{ and } \text{s.e.}(p) = 0.4 - 1.96 \times 0.1066 = 0.19.$$

$$\hat{P}_U = p + Z_{0.025}; \text{ s.e.}(p) = 0.4 + 1.96 \times 0.1066 = 0.61.$$

(iv) The estimate of total number of babies having weight more than 6.5 lb is given by

$$\hat{A} = Np = 200 \times 0.4 = 80.$$

(v) The estimate of standard error of the estimate of total number of babies having weight more than 6.5 lb is given by

$$\text{s.e.}(\hat{A}) = \sqrt{v(\hat{A})}, \text{ where } v(\hat{A}) = N^2 v(p) = (200)^2 0.011368 = 454.72.$$

(vi) The estimate of average weight of babies is given by

$$\bar{y} = \frac{1}{n} \sum y_i = \frac{127.5}{20} = 6.375 \text{ lb.}$$

(vii) The estimated variance of \bar{y} is

$$\begin{aligned} v(\bar{y}) &= \frac{N-n}{Nn} s^2, \text{ where } s^2 = \frac{1}{n-1} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right] \\ &= \frac{200-20}{200 \times 20} 0.3714, \quad = \frac{1}{20-1} \left[819.87 - \frac{(127.5)^2}{20} \right] = 0.3714. \\ &= 0.016713 \text{ (lb)}^2 \end{aligned}$$

$$\text{s.e.}(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{0.016713} = 0.1293 \text{ lb.}$$

95% confidence interval of \bar{Y} is given by $\bar{y} \pm t_{0.025, 19} \text{s.e.}(\bar{y})$.

We have $\hat{\bar{Y}}_L = \bar{y} - t_{0.025, 19} \text{s.e.}(\bar{y}) = 6.375 - 2.093 \times 0.1293 = 6.10 \text{ lb.}$

$$\hat{\bar{Y}}_U = \bar{y} + t_{0.025, 19} \text{s.e.}(\bar{y}) = 6.375 + 2.093 \times 0.1293 = 6.65 \text{ lb.}$$

(viii) The estimate of population total weight is given by

$$\hat{Y} = N\bar{y} = 200 \times 6.375 = 1275 \text{ lb.}$$

$$v(\hat{Y}) = N^2 v(\bar{y}) = (200)^2 \times 0.016713 = 668.52 \text{ (lb)}^2.$$

$$\text{s.e.}(\hat{Y}) = \sqrt{v(\hat{Y})} = \sqrt{668.52} = 25.86 \text{ lb.}$$

Therefore, 95% confidence interval for Y is given by

$$\hat{Y} \pm t_{0.025, 19} \text{s.e.}(\hat{Y}),$$

where $\hat{Y}_L = \hat{Y} - t_{0.025} \text{s.e.}(\hat{Y}) = 1275 - 2.093 \times 25.86 = 1220.88 \text{ lb.}$

$$\hat{Y}_U = \hat{Y} + t_{0.025} \text{s.e.}(\hat{Y}) = 1275 + 2.093 \times 25.86 = 1329.12 \text{ lb.}$$

Example 12.4 : There are 20000 families living in an area. Five hundred families are randomly selected without replacement from this area and found that out of 500 selected families 300 families have number of ever born children more than 2.

- (i) Estimate the proportion of families having more than 2 ever born children.
- (ii) Find 95% confidence interval for population proportion of families having ever born children more than 2.
- (iii) Estimate the total number of families having number of ever born children more than 2.

Solution : (i) Given $N = 20000$, $n = 500$, $a = 300$. The estimate of proportion of families having more than 2 ever born children is

$$p = \frac{a}{n} = \frac{300}{500} = 0.60.$$

(ii) We have $v(p) = (1 - f) \frac{pq}{n - 1}$, where $f = \frac{n}{N} = \frac{500}{20000} = 0.025$.

This f can be neglected. Then

$$v(p) = \frac{pq}{n - 1} = \frac{0.6 \times 0.4}{200 - 1} = 0.001206.$$

$$\therefore \text{s.e.}(p) = \sqrt{v(p)} = \sqrt{0.001206} = 0.0347.$$

Therefore, 95% confidence interval for population proportion of families having more than 2 ever born children is given by

$$p \pm Z_{.025} \text{s.e.}(p), \text{ where } \hat{P}_L = p - Z_{.025} \text{s.e.}(p) = 0.6 - 1.96 \times 0.0347 = 0.53.$$

$$\hat{P}_U = p + Z_{0.025} \text{s.e.}(p) = 0.6 + 1.96 \times 0.0347 = 0.67.$$

(iii) The estimate of total number of families having more than 2 ever born children is given by

$$\hat{A} = Np = 20000 \times 0.6 = 12000.$$

Chapter 13

Stratified Random Sampling

13.1 Definition

In simple random sampling the variance of the estimator of population mean or population total increases with the increase in the value of S^2 , the population variance. Again, S^2 increases with the increase in heterogeneity in the population observations. Therefore, in case of heterogeneity in the population observations the efficiency of simple random sample estimator is decreased. This problem can be overcome using alternative sampling plan to select the sample.

Let us consider that the population size of a population under study be N , where the population units are classified into k classes (strata) according to their affinity to be included in a class. For example, let us consider a population of students who appear at an entrance examination for higher studies. The students may be originated from colleges of rural area or from colleges of urban area. It is expected that or assumed that the performance of urban students is better than that of rural students. In such a case, if average performance is under study, heterogeneity in performance of students is assumed to be present. The variance in performance of all students is expected more. However, the variation in performance of rural students or the variation in performance of urban students are expected to be less. Hence, the rural students can be included in a strata and urban students can be included in another strata ($k = 2$). Now, if separate sample is drawn from two separate strata, the estimate of population parameter is expected to be more efficient (at least S^2 is expected to be less). Now, if separate simple random sample is drawn from each stratum, the sampling is known as stratified random sampling.

The stratified random sampling is a widely used random sampling technique. It helps in estimating the population parameter more efficiently, specially if the population units are more heterogeneous in characteristic under study.

Method of estimation of parameters : Let the population units be divided into k strata, where h -th stratum has N_h population units such that $\sum_h N_h = N$, $h = 1, 2, \dots, k$.

The strata are non-overlapping. No unit should be included in more than one stratum. The problem is to draw a random sample of size n in such a way that n_h units are to be drawn from h -th stratum so that $\sum_h n_h = n$. If n_h units are drawn using simple random sampling technique from h -th stratum, then sampling procedure is known as stratified random sampling.

Let y_{hi} be the value of variable under study for i -th unit of h -th stratum; $i = 1, 2, \dots, N_h$; $h = 1, 2, \dots, L$. Then the population mean of h -th stratum is given by

$$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}, \quad Y_h = N_h \bar{Y}_h = \sum_{i=1}^{N_h} y_{hi}$$

is the population total of h -th stratum. The population variance of h -th stratum is given by

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2.$$

The sample mean, sample variance and sample total of h -th stratum are given, respectively by

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}, \quad s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2, \quad \text{and} \quad y_h = n_h \bar{y}_h = \sum_{i=1}^{n_h} y_{hi}.$$

The population mean, population total and population variance are given, respectively by

$$\bar{Y} = \frac{1}{N} \sum_{h=1}^L \sum_{i=1}^{N_h} y_{hi} = \frac{1}{N} \sum_{h=1}^L N_h \bar{Y}_h, \quad \text{where} \quad N = \sum_{h=1}^L N_h$$

$$Y = \sum_{h=1}^L \sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^L N_h \bar{Y}_h, \quad \text{and}$$

$$S^2 = \frac{1}{N - 1} \sum_{h=1}^L \sum_{i=1}^{N_h} (y_{hi} - \bar{Y})^2.$$

The sample mean is defined by

$$\bar{y} = \frac{1}{n} \sum_{h=1}^L \sum_{i=1}^{n_h} y_{hi} = \frac{1}{n} \sum_{h=1}^L n_h \bar{y}_h = \sum w_h \bar{y}_h,$$

where $w_h = \frac{n_h}{n} =$ proportion of sample observations from h -th stratum. Let $W_h = \frac{N_h}{N}$ be the proportion of observations of h -th stratum and $f_h = \frac{n_h}{N_h}$ be the sampling fraction of h -th stratum. The stratified sample mean is defined by

$$\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h = \sum_{h=1}^L W_h \bar{y}_h, \quad \text{where} \quad W_h = \frac{N_h}{N}.$$

Here \bar{y} and \bar{y}_{st} will be same if $\frac{n_h}{n} = \frac{N_h}{N}$

or, $n_h = \frac{n}{N} N_h$ or, $n_h \propto N_h$, where $\frac{n}{N}$ is proportionality constant. Thus, \bar{y} and \bar{y}_{st} give the same result if sample size of h -th stratum is allocated under proportional allocation.

Let us now investigate the characteristics of the estimator \bar{y}_{st} under proportional allocation of sample size in h -th stratum.

Theorem : In stratified random sampling under proportional allocation the sample estimator \bar{y}_{st} is an unbiased estimator of population mean \bar{Y} .

Proof : We define $\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h = \sum_{h=1}^L W_h \bar{y}_h$, where $W_h = N_h/N$. It is assumed that the sample of size n_h is drawn from h -th stratum using simple random sampling scheme. So \bar{y}_h is an unbiased estimator of \bar{Y}_h [$E(\bar{y}_h) = \bar{Y}_h$]. Therefore,

$$E[\bar{y}_{st}] = E \sum_{h=1}^L W_h \bar{y}_h = \sum_{h=1}^L W_h E(\bar{y}_h) = \sum_{h=1}^L W_h \bar{Y}_h = \bar{Y}.$$

Corollary : In stratified random sampling under proportional allocation the unbiased estimator of population total is given by

$$\hat{Y} = N\bar{y}_{st}.$$

Theorem : In stratified random sampling if sample is selected independently from each stratum by simple random sampling scheme, then

$$V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 V(\bar{y}_h) = \sum_{h=1}^L \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h n_h} S_h^2,$$

where $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2 = \frac{1}{N^2} \sum N_h(N_h - n_h) \frac{S_h^2}{n_h}$.

Proof : Let us first prove that $V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 V(\bar{y}_h)$.

We have $\bar{y}_{st} = \frac{1}{N} \sum N_h \bar{y}_h = \sum_{h=1}^L W_h \bar{y}_h$.

Now,
$$V(\bar{y}_{st}) = V \left[\sum_{h=1}^L W_h \bar{y}_h \right] = \sum_{h=1}^L W_h^2 V(\bar{y}_h) + 2 \sum_{h=1}^L \sum_{j>h}^L W_h W_j \text{Cov}(\bar{y}_h, \bar{y}_j)$$

$$= \sum_{h=1}^L W_h^2 V(\bar{y}_h), \quad \because \bar{y}_h \text{ and } \bar{y}_j \text{ are independent}$$

$$= \frac{1}{N^2} \sum N_h^2 V(\bar{y}_h).$$

Again, simple random sample is drawn from h -th stratum and hence,

$$V(\bar{y}_h) = \frac{N_h - n_h}{N_h n_h} S_h^2, \quad \text{where } S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2.$$

$$\begin{aligned} \therefore V(\bar{y}_{st}) &= \frac{1}{N^2} \sum_{h=1}^L N_h^2 \frac{N_h - n_h}{N_h n_h} S_h^2 = \frac{1}{N^2} \sum_{h=1}^L N_h(N_h - n_h) \frac{S_h^2}{n_h} \\ &= \frac{1}{N^2} \sum N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\ &= \sum W_h^2 (1 - f_h) \frac{S_h^2}{n_h}, \quad f_h = \frac{n_h}{N} \\ &= \sum W_h^2 \frac{S_h^2}{n_h}, \text{ if } f_h \text{ is neglected.} \end{aligned}$$

Corollary : If n_h from h -th stratum is allocated under proportional allocation,

$$V(\bar{y}_{st}) = \sum W_h (1 - f) \frac{S_h^2}{n} = \sum W_h \frac{S_h^2}{n}, \text{ if } f \text{ is neglected.}$$

Proof : Under proportional allocation, we have $n_h \propto N_h$ or, $n_h = \frac{n}{N} N_h$.

Now, let us replace n_h by $\frac{n}{N}N_h$ in $V(\bar{y}_{st})$.

$$\begin{aligned} \text{We have } V(\bar{y}_{st}) &= \frac{1}{N^2} \sum N_h^2 \frac{N_h - \frac{n}{N}N_h}{N_h} \frac{S_h^2}{\frac{n}{N}N_h} = \frac{1}{N} \sum N_h(1-f) \frac{S_h^2}{n} \\ &= \sum_{h=1}^L W_h \frac{S_h^2}{n}, \text{ if } f = \frac{n}{N} \text{ is neglected.} \end{aligned}$$

Corollary : In stratified random sampling the variance of \hat{Y}_{st} is given by

$$\begin{aligned} V(\hat{Y}_{st}) &= N^2 V(\bar{y}_{st}) = \sum_{h=1}^L N_h(N_h - n_h) \frac{S_h^2}{n_h} \\ &= \sum N_h^2(1-f_h) \frac{S_h^2}{n_h} \\ &= \sum N_h^2 \frac{S_h^2}{n_h}, \text{ if } f_h = \frac{n_h}{N_h} \text{ is small.} \end{aligned}$$

This variance is given by

$$V(\hat{Y}_{st}) = \frac{N}{n} \sum N_h(1-f) S_h^2 = \frac{N}{h} \sum N_h S_h^2, \text{ if } f = \frac{n}{N} \text{ is neglected.}$$

When $n_h \propto N_h$.

$$\text{We have } V(\hat{Y}_{st}) = \sum N_h(N_h - n_h) \frac{S_h^2}{n_h}.$$

In case of proportional allocation, $(n_h \propto N_h)n_h = \frac{n}{N}N_h$.

Replacing n_h by $\frac{n}{N}N_h$ in $V(\hat{Y}_{st})$, we get

$$\begin{aligned} V(\hat{Y}_{st}) &= \sum N_h(N_h - \frac{n}{N}N_h) \frac{S_h^2}{\frac{n}{N}N_h} \\ &= \frac{N}{n} \sum N_h(1-f) S_h^2, \text{ where } f = \frac{n}{N} \\ &= \frac{N}{n} \sum N_h S_h^2, \text{ if } f \text{ is neglected.} \end{aligned}$$

Theorem : In stratified random sampling the estimate of variance of \bar{y}_{st} is given by

$$v(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h(N_h - n_h) \frac{s_h^2}{n_h}, \text{ where } s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2.$$

Proof : Since simple random sample is drawn from h -th stratum in stratified random sampling, the sample variance

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

is an unbiased estimator of $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$, $[E(s_h^2) = S_h^2]$.

$$\begin{aligned}\text{Now } E[v(\bar{y}_{st})] &= \frac{1}{N^2} \sum_{h=1}^L N_h(N_h - n_h) \frac{E(s_h^2)}{n_h} \\ &= \frac{1}{N^2} \sum_{h=1}^L N_h(N_h - n_h) \frac{S_h^2}{n_h} \\ &= V(\bar{y}_{st}).\end{aligned}$$

Hence, $v(\bar{y}_{st})$ is an unbiased estimator of $V(\bar{y}_{st})$.

Further, we have

$$\begin{aligned}v(\bar{y}_{st}) &= \frac{1}{N^2} \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h} = \sum W_h^2 (1 - f_h) \frac{s_h^2}{n_h} \\ &= \sum W_h^2 \frac{s_h^2}{n_h}, \text{ if } f_h = \frac{n_h}{N_h} \text{ is neglected.}\end{aligned}$$

Corollary : In stratified random sampling, if sample units from h -th stratum are selected with proportional allocation ($n_h \propto N_h$), then

$$v(\bar{y}_{st}) = \frac{1-f}{n} \sum_{h=1}^L W_h s_h^2 = \frac{1}{n} \sum W_h s_h^2, \text{ if } f \text{ is neglected,}$$

Proof : We have $v(\bar{y}_{st}) = \frac{1}{N^2} \sum N_h(N_h - n_h) \frac{s_h^2}{n_h}$, where $n_h = \frac{n}{N} N_h$.

$$\begin{aligned}\therefore v(\bar{y}_{st}) &= \frac{1}{N^2} \sum_{h=1}^L N_h \left(N_h - \frac{n}{N} N_h\right) \frac{s_h^2}{\frac{n}{N} N_h} = \sum W_h (1 - f) \frac{s_h^2}{n} \\ &= \frac{1}{n} \sum W_h s_h^2, \text{ if } f \text{ is neglected.}\end{aligned}$$

Corollary : In stratified random sampling the estimated variance of \hat{Y}_{st} is given by

$$\begin{aligned}v(\hat{Y}_{st}) &= N^2 v(\bar{y}_{st}) = \sum_{h=1}^L \frac{N_h^2 s_h^2}{n_h} (1 - f_h) \\ &= \sum_{h=1}^L \frac{N_h^2 s_h^2}{n_h}, \text{ if } f_h = \frac{n_h}{N_h} \text{ is neglected.}\end{aligned}$$

Further, this $v(\hat{Y}_{st})$ is given by

$$\begin{aligned}v(\hat{Y}_{st}) &= \frac{N(1-f)}{n} \sum_{h=1}^L N_h s_h^2, \text{ if } n_h \propto N_h \\ &= \frac{N}{n} \sum_{h=1}^L N_h s_h^2, \text{ if } f \text{ is neglected.}\end{aligned}$$

$$\begin{aligned}\text{We have } v(\hat{Y}_{st}) &= \sum_{h=1}^L N_h(N_h - n_h) \frac{s_h^2}{n_h} = \sum N_h \left(N_h - \frac{n}{N} N_h\right) \frac{s_h^2}{\frac{n}{N} N_h}, \text{ if } n_h = \frac{n}{N} N_h \\ &= \frac{N}{n} \sum_{h=1}^L N_h (1 - f) s_h^2 = \frac{N}{n} \sum_{h=1}^L N_h s_h^2, \text{ if } f \text{ is neglected.}\end{aligned}$$

Corollary : In stratified random sampling, if sampling units from h -th stratum are selected according to proportional sampling scheme ($n_h \propto N_h$), then $100(1 - \alpha)\%$ confidence intervals of \bar{Y} and Y are respectively given by

$$\bar{y}_{st} \pm t_{\frac{\alpha}{2}} \sqrt{\frac{1-f}{n} \sum W_h s_h^2} \quad \text{or,} \quad \bar{y}_{st} \pm t_{\frac{\alpha}{2}} \sqrt{\frac{1}{n} \sum_h W_h s_h^2}, \text{ if } f \text{ is neglected}$$

and $\hat{Y}_{st} \pm t_{\frac{\alpha}{2}} \sqrt{\frac{N}{h} \sum N_h (1-f) s_h^2}$

or, $\hat{Y}_{st} \pm t_{\frac{\alpha}{2}} \sqrt{\frac{N}{n} \sum N_h s_h^2}$, if f is neglected.

If it is assumed that \bar{y}_{st} and \hat{Y}_{st} are normally distributed, the value of $t_{\frac{\alpha}{2}}$ is to be replaced by $Z_{\frac{\alpha}{2}}$, where Z is a normal variate. The assumption will be true, if all n_h are sufficiently large. For small value of n_h , t is to be used from Student's t -table, where the d.f. of t is to be approximated [Satterthwaite (1946)] by

$$n_a = \frac{\left(\sum_{h=1}^L g_h s_h^2 \right)^2}{\sum_{h=1}^L \frac{g_h^2 s_h^2}{n_h - 1}}$$

where $g_h = \frac{N_h(N_h - n_h)}{n_h} = \frac{N N_h(1-f)}{n} = \frac{N N_h}{n}$, if f is neglected.

Example 13.1 : In a rural area there are 345 farmers. These farmers are classified into 3 classes according to their amount of cultivable land. The small scale farmers have less than 1 hectare of land, the medium farmers have 1 to 3 hectares of land and large scale farmers have 3 and more hectares of land. The amount of land cultivated for paddy by these farmers in a season are shown below :

Amount of land (y_{hi} , in hectares)

Small farmers	S.L. No.	01	02	03	04	05	06	07	08	09	10	11
	y_{1i}	0.5	0.6	0.4	0.7	0.6	0.7	0.5	0.5	0.6	0.7	0.3
	S.L. No.	12	13	14	15	16	17	18	19	20	21	22
	y_{1i}	0.5	0.7	0.6	0.4	0.4	0.4	0.6	0.2	0.3	0.4	0.6
	S.L. No.	23	24	25	26	27	28	29	30	31	32	33
	y_{1i}	0.6	0.6	0.7	0.7	0.5	0.4	0.3	0.5	0.5	0.2	0.1
	S.L. No.	34	35	36	37	38	39	40	41	42	43	44
	y_{1i}	0.7	0.6	0.5	0.5	0.5	0.4	0.3	0.1	0.2	0.1	0.2
	S.L. No.	45	46	47	48	49	50	51	52	53	54	55
	y_{1i}	0.2	0.1	0.4	0.4	0.3	0.5	0.3	0.6	0.3	0.5	0.5
	S.L. No.	56	57	58	59	60	61	62	63	64	65	66
	y_{1i}	0.4	0.3	0.3	0.3	0.4	0.4	0.2	0.4	0.6	0.6	0.5

Small farmers	S.L. No.	67	68	69	70	71	72	73	74	75	76	77				
	y_{1i}	0.2	0.6	0.5	0.5	0.4	0.3	0.6	0.7	0.2	0.6	0.4				
	S.L. No.	78	79	80	81	82	83	84	85	86	87	88				
	y_{1i}	0.5	0.5	0.4	0.6	0.6	0.5	0.4	0.4	0.3	0.2	0.1				
	S.L. No.	89	90	91	92	93	94	95	96	97	98	99				
	y_{1i}	0.6	0.4	0.3	0.4	0.4	0.4	0.2	0.1	0.2	0.7	0.6				
	S.L. No.	100	101	102	103	104	105	106	107	108	109	110				
	y_{1i}	0.3	0.5	0.4	0.2	0.4	0.2	0.6	0.7	0.4	0.4	0.5				
	S.L. No.	111	112	113	114	115	116	117	118	119	120	121				
y_{1i}	0.6	0.5	0.4	0.5	0.5	0.5	0.6	0.6	0.6	0.4	0.3					
Medium farmers	S.L. No.	01	02	03	04	05	06	07	08	09	10	11				
	y_{2i}	2.5	2.0	1.5	1.6	1.0	2.4	2.0	2.0	1.8	1.9	1.5				
	S.L. No.	12	13	14	15	16	17	18	19	20	21	22				
	y_{2i}	1.0	1.6	2.0	1.8	1.5	1.6	1.0	1.2	1.4	1.5	1.0				
	S.L. No.	23	24	25	26	27	28	29	30	31	32	33				
	y_{2i}	1.6	1.8	1.9	2.0	2.2	2.4	2.0	2.5	1.6	1.7	1.5				
	S.L. No.	34	35	36	37	38	39	40	41	42	43	44				
	y_{2i}	1.0	1.2	1.4	2.0	1.7	1.8	1.6	1.9	2.0	1.8	1.5				
	S.L. No.	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
	y_{2i}	1.8	1.5	1.6	1.5	1.0	1.2	1.4	1.2	1.0	1.1	1.3	1.2	1.6	1.8	1.9
	S.L. No.	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
	y_{2i}	1.0	1.1	1.2	1.3	1.0	1.4	1.5	1.1	1.6	1.7	1.6	1.7	1.8	1.9	1.2
	S.L. No.	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89
	y_{2i}	1.6	1.0	1.2	1.2	1.3	1.4	1.8	1.4	1.5	1.6	1.7	2.0	2.1	2.2	2.2
	S.L. No.	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104
	y_{2i}	2.4	1.6	1.7	2.0	2.1	2.2	2.0	1.6	1.8	1.7	1.8	1.0	1.6	1.4	1.5
	S.L. No.	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119
y_{2i}	1.8	2.0	1.9	2.4	2.8	2.2	2.6	2.3	2.0	1.9	1.6	2.0	2.4	2.5	1.7	
S.L. No.	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	
y_{2i}	1.0	1.2	1.4	1.3	1.5	1.6	1.4	1.6	1.4	1.7	1.2	1.1	1.2	1.3	1.2	

Medium farmers	S.L. No.	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149
	y_{2i}	1.8	1.6	1.7	2.0	2.1	2.2	2.4	1.8	1.6	2.2	2.0	1.6	1.2	1.5	1.5
Large farmers	S.L. No.	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
	y_{3i}	4.5	4.0	3.8	5.5	6.8	7.5	4.6	5.5	5.6	5.0	4.2	4.5	6.2	4.8	5.6
	S.L. No.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	y_{3i}	5.0	4.5	4.2	5.0	5.0	5.0	5.0	6.2	5.5	5.2	5.1	4.8	4.0	4.5	4.6
	S.L. No.	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
	y_{3i}	4.0	4.2	6.0	6.1	6.2	7.5	7.2	6.0	4.8	4.4	5.0	5.1	5.0	5.0	4.5
	S.L. No.	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
	y_{3i}	4.0	4.2	4.8	5.0	5.2	5.2	5.1	4.2	4.0	4.6	4.7	4.6	4.0	4.8	4.9
S.L. No.	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	
y_{3i}	5.0	5.1	5.2	5.0	4.6	4.1	4.2	4.0	4.2	4.1	4.2	4.5	5.0	5.5	6.0	

- (i) Select 20% farmers by stratified random sampling scheme to estimate the average land cultivated for paddy. Select sample by proportional allocation.
- (ii) Find 95% confidence interval for the average land cultivated for paddy.
- (iii) Find 95% confidence interval for the total land cultivated for paddy.

Solution : (i) We have $N = 345$, $N_1 = 121$, $N_2 = 149$, $N_3 = 75$. We need a sample of $n = 69$ (20% of 345). Under proportional allocation,

$$n_h \propto \frac{n}{N} N_h, \quad h = 1, 2, 3.$$

$n_1 = 24$, $n_2 = 30$, $n_3 = 15$.

Small farmers	Rn. No.	514	161	481	837	953	946	642	721	164	100	853	840	767	190
	SL. No of units	030	040	118	111	106	099	037	116	043	100	006	114	041	069
	y_{1i}	0.5	0.3	0.6	0.6	0.6	0.6	0.5	0.5	0.1	0.3	0.7	0.5	0.1	0.5
	Rn. No.	727	115	714	175	095	816		972	262	532	461			
	SL. No of units	001	115	109	054	095	090		004	020	48	098			
	y_{1i}	0.5	0.5	0.4	0.5	0.2	0.4		0.7	0.3	0.4	0.7			
Medium farmers	Rn. No.	532	816	678	504	597	947	018	851	977	433	478	242	319	379
	SL. No of units	085	071	082	057	001	053	018	106	083	135	031	093	021	081
	y_{2i}	1.7	1.7	1.4	1.6	2.5	1.0	1.0	2.0	1.5	1.8	1.6	2.0	1.5	1.8
	Rn. No.	829	659	569	162	963	841	490	435	252	392	568	482	483	384
	SL. No of units	084	063	122	013	069	096	043	137	103	094	121	035	036	086
	y_{2i}	1.6	1.3	1.4	1.6	1.7	2.0	1.8	1.7	1.4	2.1	1.2	1.2	1.4	2.0

	Rn. No.	524		102											
	SL. No of units	078		102											
	y_{2i}	-	1.2	1.6											
Large farmers	Rn. No.	192	420	662	514	820	397	734	590	001	703	101	234	011	922
	SL. No of units	042	045	062	064	070	022	059	065	001	028	026	009	011	022
	y_{3i}	5.1	4.5	5.1	5.0	4.1	5.0	4.8	4.6	4.5	4.0	5.1	5.6	4.2	5.0
	Rn. No.	774													
	SL. No. of units	94													

Here $\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{24} y_{1i} = \frac{11}{24} = 0.46$ hectare

$\bar{y}_2 = \frac{1}{n_2} \sum_{i=1}^{30} y_{2i} = \frac{48.3}{30} = 1.61$ hectares, $\bar{y}_3 = \frac{1}{n_3} \sum_{i=1}^{15} y_{3i} = \frac{68.4}{15} = 4.56$ hectares.

The estimated average of land cultivated for paddy is given by

$$\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{Y}_h = \frac{1}{345} [121 \times 0.46 + 149 \times 1.61 + 75 \times 4.56]$$

= 1.85 hectares.

(ii) We have $s_h^2 = \frac{1}{n_h - 1} \left[\sum y_{hi}^2 - \frac{(\sum y_{hi})^2}{n_h} \right], h = 1, 2, 3$

$$s_1^2 = \frac{1}{24 - 1} \left[5.35 - \frac{(11)^2}{24} \right], \quad s_2^2 = \frac{1}{30 - 1} \left[81.17 - \frac{(48.3)^2}{30} \right] = 0.1175$$

= 0.0134.

$$s_3^2 = \frac{1}{15 - 1} \left[322.78 - \frac{(68.4)^2}{15} \right] = 11.9467.$$

The estimate of variance of estimated average amount of land (\bar{y}_{st}) is given by

$$v(\bar{y}_{st})_{prop} = \frac{1-f}{n} \sum W_h s_h^2$$

$$W_1 = \frac{N_1}{N} = \frac{121}{345} = 0.35, \quad W_2 = \frac{N_2}{N} = \frac{149}{345} = 0.43, \quad W_3 = \frac{N_3}{N} = \frac{75}{345} = 0.22.$$

$$v(\bar{y}_{st})_{prop} = \frac{(1 - \frac{69}{345})}{69} [0.35 \times 0.0134 + 0.43 \times 0.1175 + 0.22 \times 11.9467]$$

$$= \frac{1 - 0.20}{69} \times 2.683489 = 0.031113.$$

s.e. (\bar{y}_{st})_{prop} = $\sqrt{v(\bar{y}_{st})_{prop}} = \sqrt{0.031113} = 0.1764.$

The 95% of confidence interval of population mean \bar{Y} is

$$\bar{y}_{st} \pm t_{0.025} \text{s.e.}(\bar{y}_{st})_{\text{prop}},$$

where $t_{0.025}$ is the tabulated value of t at 5% level of significance with n_e d.f., where

$$n_e = \frac{(\sum g_h s_h^2)^2}{\sum \frac{g_h^2 s_h^4}{n_h - 1}}$$

$$g_h = \frac{N N_h}{n} (1 - f)$$

$$g_1 = \frac{345 \times 121}{69} (1 - 0.2), \quad g_2 = \frac{345 \times 149}{69} (1 - 0.2), \quad g_3 = \frac{345 \times 75}{69} (1 - 0.2)$$

$$= 484 \qquad \qquad \qquad = 596 \qquad \qquad \qquad = 300$$

$$n_e = \frac{(484 \times 0.0134 + 596 \times 0.1175 + 300 \times 11.9467)^2}{\frac{(484)^2 (0.0134)^2}{24-1} + \frac{(596)^2 (0.1175)^2}{30-1} + \frac{(300)^2 (11.9467)^2}{15-1}}$$

$$= \frac{13399447.67}{917680.0592} = 15.$$

Therefore, $\hat{Y}_L = \bar{y}_{st} - t_{0.025, 15} \text{s.e.}(\bar{y}_{st}) = 1.85 - 2.131 \times 0.1764 = 1.47$ hectares.

$$\hat{Y}_U = \bar{y}_{st} + t_{0.025, 15} \text{s.e.}(\bar{y}_{st}) = 1.85 + 2.131 \times 0.1764 = 2.23 \text{ hectares.}$$

(iii) 95% confidence interval for total land cultivated for paddy is given by

$$\hat{Y}_{st} \pm t_{0.025, 15} \text{s.e.}(\hat{Y}_{st}).$$

Here $\hat{Y}_{st} = N \bar{y}_{st} = 345 \times 1.85 = 638.25$ hectares

$$v(\hat{Y}_{st}) = N^2 v(\bar{y}_{st}) = (345)^2 (0.031113) = 3703.2248$$

$$\text{s.e.}(\hat{Y}_{st}) = \sqrt{v(\hat{Y}_{st})} = \sqrt{3703.2248} = 60.8541.$$

Now $\hat{Y}_L = \hat{Y}_{st} - t_{0.025, 15} \text{s.e.}(\hat{Y}_{st}) = 638.25 - 2.131 \times 60.8541 = 508.57$ hectares.

$$\hat{Y}_U = \hat{Y}_{st} + t_{0.025, 15} \text{s.e.}(\hat{Y}_{st}) = 638.25 + 2.131 \times 60.8541 = 767.93 \text{ hectares.}$$

13.2 Allocation of Sample Size in Different Strata

In stratified random sampling the sample units are selected from stratum by simple random sampling scheme. The sample size n_h is to be selected from h -th stratum of size N_h . The problem is to decide the value of n_h for h -th stratum. This problem is known as problem of allocation of sample size in h -th stratum. One solution of this problem is to select n_h proportional to N_h . In general, the value of n_h is allocated in such a way that the estimate is more precise. However, the decision regarding n_h is so made that it is consistent with the overall resources of the survey.

The important points to be considered in allocating n_h are :

(i) stratum size, (ii) variability of variable observed within stratum units, (iii) cost of survey of sampling unit in the stratum. Considering all the above points, the following four methods of allocation are, usually, considered. These are :

- (a) Equal sample size in each stratum.
- (b) Proportional allocation.
- (c) Optimum allocation.

Optimum allocation is again of two types, viz.,

(i) Neyman allocation ($n_h \propto N_h S_h$).

(ii) Optimum allocation for a fixed cost $\left[n_h \propto \frac{N_h S_h}{\sqrt{C_h}} \right]$

Here S_h is the standard deviation of variable observed in h -th stratum.

Equal sample size in each stratum : Let a population be of size N . The population units are divided into strata according to the homogeneity of the values of variable under study. Let the population units of h -th stratum be $N_h (h = 1, 2, \dots, L)$. We need a sample of size n . Then

$$n_h = \frac{n}{L} \quad (h = 1, 2, \dots, L)$$

indicates that sample units are equally allocated to h -th stratum.

This allocation is done to avoid the problem that arises in selecting sample. It is an administrative advantage in sample selection. However, the estimate is not efficient if N_h 's differ much and if the variability in observations is more.

The variance of the estimated mean is given by

$$V(\bar{y}_{st}) = \sum_h W_h^2 \left(L - \frac{n}{N_h} \right) \frac{S_h^2}{n}$$

The estimated variance is given by

$$v(\bar{y}_{st}) = \sum_h W_h^2 \left(L - \frac{n}{N_h} \right) \frac{s_h^2}{n}$$

Proportional allocation : Bowley (1926) has proposed this method of allocation. In this method the sample is allocated by

$$n_h \propto N_h \quad \text{or,} \quad n_h = \frac{n}{N} N_h,$$

where $\frac{n}{N}$ is proportionality constant.

The estimate of variance and the variance of the estimate of population mean under this allocation scheme has already been discussed in the previous section.

Optimum allocation : In this method the sample size n_h is selected according to

(i) $n_h \propto N_h S_h$ [allocation under fixed sample size]

and (ii) $n_h \propto \frac{N_h S_h}{\sqrt{C_h}}$ [allocation under fixed cost]

where $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$

and $C = C_0 + \sum_{h=1}^L C_h n_h$, where C_h is the cost of selection of sample from h -th stratum.

The first method of optimum allocation is known as Neyman allocation and has been first proposed by Tschuprow (1923). In this allocation the variability of observation is also under

consideration. Under Neyman (1934) allocation,

$$n_h = \frac{nN_h S_h}{\sum N_h S_h}.$$

In this allocation, it is assumed that the cost of sampling in each stratum remains same. The variance of the estimator under this sampling scheme becomes minimum.

Theorem : In stratified random sampling the variance of \bar{y}_{st} is minimum if $n_h \propto N_h S_h$, where \bar{y}_{st} is an estimate of population mean from a specified sample of size n .

Proof : We know $V(\bar{y}_{st}) = \frac{1}{N^2} \sum N_h(N_h - n_h) \frac{S_h^2}{n_h}$.

The sample size $n = \sum_{h=1}^L n_h$. The value of n_h is to be selected in such a way that $V(\bar{y}_{st})$ is minimum. This is possible, if

$$\phi = V(\bar{y}_{st}) + \lambda \left(\sum n_h - n \right)$$

is minimum. Here λ is called Lagrange's Multiplier.

$$\begin{aligned} \text{We have } \phi &= \frac{1}{N^2} \sum N_h(N_h - n_h) \frac{S_h^2}{n_h} + \lambda \left(\sum n_h - n \right) \\ &= \frac{1}{N^2} \sum \frac{N_h^2 S_h^2}{n_h} - \frac{1}{N^2} \sum N_h S_h^2 + \lambda \left(\sum n_h - n \right) \\ \frac{\partial \phi}{\partial n_h} &= -\frac{1}{N^2} \frac{N_h^2 S_h^2}{n_h^2} + \lambda. \end{aligned}$$

Using the principle of maxima and minima, we can write

$$\frac{\partial \phi}{\partial n_h} = 0 \Rightarrow \lambda = \frac{1}{N^2} \frac{N_h^2 S_h^2}{n_h^2}.$$

We have $n_h = \frac{N_h S_h}{N \sqrt{\lambda}}$.

Taking sum on both sides, we get

$$n = \frac{\sum N_h S_h}{N \sqrt{\lambda}} \quad \text{or, } \sqrt{\lambda} = \frac{\sum N_h S_h}{N n}.$$

Putting the value of $\sqrt{\lambda}$ in the equation of n_h above, we get

$$n_h = \frac{N_h S_h N n}{N \sum N_h S_h} = \frac{n N_h S_h}{\sum N_h S_h}.$$

$\therefore n_h \propto N_h S_h$, where $\frac{n}{\sum N_h S_h}$ is proportionality constant.

The minimum variance under Neyman allocation is given by

$$\begin{aligned} V(\bar{y}_{st})_{\min} &= \frac{1}{nN^2} \left(\sum N_h S_h \right)^2 - \frac{1}{N^2} \sum N_h S_h^2 \\ &= \frac{1}{n} \left(\sum W_h S_h \right)^2 - \frac{1}{N} \sum W_h S_h^2 \\ &= \frac{1}{n} \left(\sum W_h S_h \right)^2, \text{ if f.p.c. is neglected.} \end{aligned}$$

The estimate of $V(\bar{y}_{st})_{\min}$ is obtained by

$$\begin{aligned} v(\bar{y}_{st})_{\min} &= \frac{1}{n} \left(\sum W_h s_h \right)^2 - \frac{1}{N} \sum W_h s_h^2 \\ &= \frac{1}{n} \left(\sum W_h s_h \right)^2, \text{ it f.p.c. is neglected} \end{aligned}$$

The minimum variance of \hat{Y}_{st} under Neyman allocation is given by

$$\begin{aligned} V(\hat{Y}_{st})_{\min} &= N^2 V(\bar{y}) \\ &= \frac{1}{n} \left(\sum N_h S_h \right)^2 - \sum N_h S_h^2. \end{aligned}$$

The unbiased estimate of $V(\hat{Y}_{st})_{\min}$ is

$$v(\hat{Y}_{st})_{\min} = \frac{1}{n} \left(\sum N_h s_h \right)^2 - \sum N_h s_h^2,$$

where $s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$.

The variance of estimate under Neyman allocation becomes minimum. But there is one disadvantage of this method of allocation, since the value of n_h involves S_h . In practice, the value of S_h is not known. However, S_h can be replaced by its estimate, if it is available from any previous survey conducted to study the same variable. The value S_h can also be estimated in the first stage of sampling if the sampling is done in different stages.

Allocation under fixed cost : Neyman allocation does not take into consideration of cost involvement in selecting sample from h -th stratum. We need minimum variance of the estimator, but the survey should be finished within a fixed amount of cost. Let us consider that for a survey the fixed cost is

$$C = C_0 + \sum C_h n_h,$$

where C_h is the cost of taking sample from h -th stratum, C is the total cost of the survey and C_0 is the overhead cost of survey. The sample size n_h is to be allocated in h -th stratum so that the variance of the estimator is minimum under the given fixed cost.

Theorem : Under a fixed cost function of the type

$$C = C_0 + \sum n_h C_h,$$

the variance of the estimator of population mean is minimum, if $n_h \propto \frac{N_h S_h}{\sqrt{C_h}}$.

The minimum variance is given by

$$V(\bar{y}_{st})_{\min} = \frac{1}{N^2 n} \sum (N_h S_h / \sqrt{C_h}) \sum \frac{N_h S_h}{\sqrt{C_h}} - \frac{1}{N^2} \sum N_h S_h^2.$$

Proof : The variance of the estimator of population mean is

$$\begin{aligned} V(\bar{y}_{st}) &= \frac{1}{N^2} \sum (N_h - n_h) \frac{N_h S_h^2}{n_h} \\ &= \frac{1}{N^2} \sum \frac{N_h^2 S_h^2}{n_h} - \frac{1}{N^2} \sum N_h S_h^2. \end{aligned}$$

To allocate sample size n_h in h -th stratum so that $V(\bar{y}_{st})$ is minimum, let us consider a function

$$\phi = \frac{1}{N^2} \sum \frac{N_h^2 S_h^2}{n_h} - \frac{1}{N^2} \sum N_h S_h^2 + \lambda [C_0 + \sum C_h n_h - C],$$

where λ is Lagrange's multiplier. The value of n_h is to be found out from the equation

$$\frac{\partial \phi}{\partial n_h} = 0 \text{ (by principle of maxima and minima).}$$

$$\text{We have } \frac{\partial \phi}{\partial n_h} = - = - \frac{N_h^2 S_h^2}{N^2 n_h^2} + \lambda C_h = 0$$

$$\text{or, } \lambda C_h = \frac{N_h^2 S_h^2}{N^2 n_h^2} \quad \text{or, } n_h = \frac{N_h S_h}{N \sqrt{\lambda C_h}}.$$

Taking sum over h on both sides, we get

$$n = \sum \frac{N_h S_h}{N \sqrt{\lambda C_h}} \quad \text{or, } \sqrt{\lambda} = \frac{1}{N n} \sum \frac{N_h S_h}{\sqrt{C_h}}.$$

Substituting the value of $\sqrt{\lambda}$ in the equation of n_h given above, we get

$$n_h = \frac{N_h S_h}{N \sqrt{C_h} \frac{1}{N n} \sum \frac{N_h S_h}{\sqrt{C_h}}} = \frac{n}{\sum \frac{N_h S_h}{\sqrt{C_h}}} \frac{N_h S_h}{\sqrt{C_h}}$$

$$\text{or, } n_h \propto \frac{N_h S_h}{\sqrt{C_h}}, \quad \text{where } \frac{n}{\sum \frac{N_h S_h}{\sqrt{C_h}}} \text{ is called } \textit{proportionality constant}.$$

Replacing the value of n_h in $V(\bar{y}_{st})$, we get

$$V(\bar{y}_{st})_{\min} = \frac{1}{n N^2} \left(\sum N_h S_h \sqrt{C_h} \right) \sum \frac{N_h S_h}{\sqrt{C_h}} - \frac{1}{N^2} \sum N_h S_h^2.$$

If all C_h 's are same [$C_1 = C_2 = \dots = C_L = C$ (say)],

$$V(\bar{y}_{st})_{\min} = \frac{1}{n N^2} \left(\sum N_h S_h \right)^2 - \frac{1}{N^2} \sum N_h S_h^2 = V(\bar{y}_{st})_{\min} \text{ [by Neyman allocation].}$$

Again, if all S_h 's are same, then n_h under Neyman allocation becomes

$$n_h = \frac{n}{N} N_h.$$

From this optimum allocation, it may be concluded as follows :

- (a) If N_h is large, n_h should be large.
- (b) If S_h^2 is large, n_h should be large.
- (c) If cost C_h is smaller, n_h may be larger.

The conclusions (a) and (b) are applicable in case of Neyman allocation also.

For optimum allocation the value of n should be pre-determined. However, if n is not known, it is estimated by

$$n = \frac{(C - C_0)(\sum N_h S_n / \sqrt{C_h})}{\sum N_h S_h \sqrt{C_h}}, \text{ if cost is fixed}$$

$$\text{and } n = \frac{(\sum W_h S_h \sqrt{C_h})(\sum W_h S_h / \sqrt{C_h})}{V(\bar{y}_{st}) + \frac{1}{N} \sum W_h S_h^2}, \text{ if } V(\bar{y}_{st}) \text{ is fixed.}$$

Example 13.2 : Draw a stratified random sample of size $n = 50$ by Neyman allocation using the data of Example 13.1. Find 95% confidence interval for population mean and population total.

Solution : In the given example $N_1 = 121, N_2 = 149, N_3 = 75, N = 345$. Also, we have

$$S_1^2 = \frac{1}{N_1 - 1} \left[\sum y_{1i}^2 - \frac{(\sum y_{1i})^2}{N_1} \right] = \frac{1}{120} \left[25.86 - \frac{(52.8)^2}{121} \right] = 0.0235$$

$$S_2^2 = \frac{1}{N_2 - 1} \left[\sum y_{2i}^2 - \frac{(\sum y_{2i})^2}{N_2} \right] = \frac{1}{148} \left[434.06 - \frac{(252.8)^2}{149} \right] = 0.0394$$

$$S_3^2 = \frac{1}{N_3 - 1} \left[\sum y_{3i}^2 - \frac{(\sum y_{3i})^2}{N_3} \right] = \frac{1}{74} \left[1901.17 - \frac{(372.8)^2}{73} \right] = 0.6501$$

$$S_1 = 0.1533, S_2 = 0.1985, S_3 = 0.8063, \sum N_h S_h = 108.5983.$$

According to Neyman allocation $n_h = \frac{n N_h S_h}{\sum N_h S_h}$.

$$\therefore n_1 = \frac{50 \times 18.5493}{108.5983}, n_2 = \frac{50 \times 29.5765}{108.5983}, n_3 = \frac{50 \times 60.4725}{108.5983}$$

$$= 8.54 \approx 8 \quad = 13.62 \approx 14 \quad = 27.84 \approx 28$$

Strata	The Sample observations are (y_{hi})								
Small farmers	Rn. No.	514	161	481	837	953	946	642	721
	SL. No. of units	30	40	118	111	106	99	37	116
	y_{1i}	0.5	0.3	0.6	0.6	0.6	0.6	0.5	0.5

Strata	The sample observations (y_{hi})																
Medium farmers	Rn. No.	164	100	853	840	767	190	727	115	714	175	816	458	972	262		
	SL. No.	015	100	108	95	022	041	131	115	118	026	071	011	078	113		
	y_{2i}	1.8	1.8	2.4	2.2	1.0	1.9	1.1	1.6	2.5	2.0	1.7	1.5	1.2	2.0		
Large farmers	Rn. No.	877	461	532	816	678	504	597	947	018	851	977	433	478	242	319	379
	SL. No.	052	011	007	066	003	054	072	047	018	026	002	058	028	017	019	004
	y_{3i}	5.1	4.2	4.6	4.1	3.8	4.0	4.5	4.2	4.2	5.1	4.0	4.0	4.0	4.5	5.0	5.5
	Rn. No.	827	659	569	162	963	841	490	435	252	568	482	483				
	SL. No.	002	059	044	012	063	016	040	060	027	043	032	033				
y_{3i}	4.0	4.8	5.0	4.5	5.2	5.0	4.4	4.9	4.8	5.0	4.2	6.0					

Now, $\bar{y}_1 = \frac{1}{n_1} \sum y_{1i} = \frac{4.2}{8} = 0.525$ hectare, $\bar{y}_2 = \frac{1}{n_2} \sum y_{2i} = \frac{24.7}{14} = 1.764$ hectares

$$\bar{y}_3 = \frac{1}{n_3} \sum y_{3i} = \frac{128.6}{28} = 4.593 \text{ hectares.}$$

The estimate of population mean is

$$\begin{aligned}\bar{y}_{st} &= \frac{1}{N} \sum N_h \bar{y}_h = \frac{1}{345} [121 \times 0.525 + 149 \times 1.764 + 75 \times 4.593] \\ &= \frac{728.836}{345} = 2.1 \text{ hectares.}\end{aligned}$$

Again, $s_h^2 = \frac{1}{n_h - 1} \left[\sum y_{hi}^2 - \frac{(\sum y_{hi})^2}{n_h} \right]$

$$s_1^2 = \frac{1}{8-1} \left[2.28 - \frac{(4.2)^2}{8} \right] = 0.01071, \quad s_2^2 = \frac{1}{14-1} \left[46.29 - \frac{(24.7)^2}{14} \right] = 0.2086,$$

$$s_3^2 = \frac{1}{28-1} \left[598.48 - \frac{(128.6)^2}{28} \right] = 0.2903.$$

$$s_1 = 0.1035, \quad s_2 = 0.4567, \quad s_3 = 0.5388.$$

The estimate of variance of \bar{y}_{st} by Neyman allocation is

$$\begin{aligned}v(\bar{y}_{st}) &= \frac{1}{n} \left(\sum W_h s_h \right)^2 - \frac{1}{N} \sum W_h s_h^2 \\ &= \frac{1}{50} \left[\frac{121 \times 0.1035}{345} + \frac{149 \times 0.4567}{345} + \frac{75 \times 0.5388}{345} \right]^2 \\ &\quad - \frac{1}{345} \left[\frac{121 \times 0.01071}{345} + \frac{149 \times 0.2086}{345} + \frac{75 \times 0.2903}{345} \right] \\ &= 0.0024594 - 0.0004549 = 0.0020045.\end{aligned}$$

$$\text{s.e.}(\bar{y}_{st}) = \sqrt{v(\bar{y}_{st})} = \sqrt{0.0020045} = 0.04477.$$

95% confidence interval of population mean (\bar{Y}) is

$$\bar{y}_{st} \pm t_{0.025, n_e} \text{s.e.}(\bar{y}_{st}),$$

where $n_e = \frac{(\sum g_h s_h^2)^2}{\sum \frac{g_h^2 s_h^4}{n_h - 1}}$, where $g_h = \frac{N N_h}{n} (1 - f)$

$$f = \frac{n}{N} = \frac{50}{345} = 0.145, \quad g_1 = \frac{345 \times 121}{50} (1 - 0.145) = 713.84,$$

$$g_2 = \frac{345 \times 149}{50} (1 - 0.145) = 879.02, \quad g_3 = \frac{345 \times 75}{50} (1 - 0.145) = 442.46,$$

$$\begin{aligned}n_e &= \frac{(713.84 \times 0.01071 + 879.02 \times 0.2086 + 442.46 \times 0.2903)^2}{\frac{(713.84)^2 (0.01071)^2}{8-1} + \frac{(879.02)^2 (0.2086)^2}{14-1} + \frac{(442.46)^2 (0.2903)^2}{28-1}} \\ &= \frac{102051.4564}{3205.7252} \approx 32, \quad t_{0.025, 32} = 2.037.\end{aligned}$$

Now, $\hat{Y}_L = \bar{y}_{st} - t_{0.025, 32} \text{s.e.}(\bar{y}_{st}) = 2.11 - 2.037 \times 0.04477 = 2.02$ hectares

$$\hat{Y}_U = \bar{y}_{st} + t_{0.025, 32} \text{s.e.}(\bar{y}_{st}) = 2.11 + 2.037 \times 0.04477 = 2.20$$
 hectares.

The estimate of population total is

$$\hat{Y}_{st} = N \bar{y}_{st} = 345 \times 2.11 = 727.95.$$

$$v(\hat{Y}_{st}) = N^2 v(\bar{y}_{st}) = (345)^2(0.0020045) = 238.585612.$$

$$\text{s.e.}(\hat{Y}_{st}) = \sqrt{v(\hat{Y}_{st})} = \sqrt{238.585612} = 15.4462.$$

95% confidence interval of population total (Y) is $\hat{Y}_{st} \pm t_{0.025,32} \text{ s.e.}(\hat{Y}_{st})$.

$$\text{Thus, } \hat{Y}_L = \hat{Y}_{st} - t_{0.025} \text{ s.e.}(\hat{Y}_{st}) = 727.95 - 2.037 \times 15.4462 = 696.49.$$

$$\hat{Y}_U = \hat{Y}_{st} + t_{0.025} \text{ s.e.}(\hat{Y}_{st}) = 727.95 + 2.037 \times 15.4462 = 759.41.$$

13.3 Estimation of Proportion from Stratified Random Sample

The parameters which are usually estimated from sample observations are population mean and population total, where the population mean is \bar{Y} and population total is Y . The variable under study is y , where y_{hi} is the value of i -th unit in h -th stratum ($h = 1, 2, \dots, h; i = 1, 2, \dots, N_h$). The total units in the population are $N = \sum_{h=1}^L N_h$. Here

$$\bar{Y} = \frac{1}{N} \sum_h^L \sum_i^{N_h} y_{hi}, \quad Y = \sum_h \sum_i y_{hi} = \sum_h N_h \bar{Y}_h, \quad \bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}.$$

Here values of y are y_{hi} and the variable is assumed to be quantitative in nature.

The variable may be qualitative in nature, where $y_{hi} = 1$, if the quality under study is present in i -th unit of h -th stratum $y_{hi} = 0$ otherwise.

Let A_h number of units in h -th stratum possess the characteristic (quality) under study, where

$$A_h = \sum_{i=1}^{N_h} y_{hi}.$$

Then A = total number of units in the population possessing the characteristic under study

$$= \sum_{h=1}^L A_h = \sum_h^L \sum_i^{N_h} y_{hi}.$$

The population proportions of units possessing the quality in h -th stratum and in the population are, respectively

$$P_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi} = \frac{A_h}{N_h} \quad \text{and} \quad P = \frac{1}{N} \sum_h^L \sum_i^{N_h} y_{hi} = \frac{1}{N} \sum A_h = \frac{A}{N}.$$

$$\text{We have } Q_h = 1 - P_h = 1 - \frac{A_h}{N_h}, \quad Q = 1 - P = 1 - \frac{A}{N}.$$

$$\text{Also, we have } S_h^2 = \frac{1}{N_h - 1} \left[\sum_i y_{hi}^2 - \frac{(\sum y_{hi})^2}{N_h} \right] = \frac{1}{N_h - 1} \left[A_h - \frac{A_h^2}{N_h} \right] = \frac{N_h P_h Q_h}{N_h - 1}.$$

$$\begin{aligned} S^2 &= \frac{1}{N - 1} \left[\sum_h \sum_i y_{hi}^2 - \frac{(\sum_h \sum_i y_{hi})^2}{N} \right] = \frac{1}{N - 1} \left[\sum A_h - \frac{(\sum A_h)^2}{N} \right] \\ &= \frac{1}{N - 1} \left[A - \frac{A^2}{N} \right] = \frac{NPQ}{N - 1}. \end{aligned}$$

Let a sample of n units be drawn from the population by stratified random sampling scheme.

The sample size from h -th stratum is n_h so that $n = \sum_{h=1}^L n_h$. Consider that a_h number of units in the sample from h -th stratum possess the quality under study and $a = \sum a_h$ is the number of units in the sample possessing the characteristic. Here

$$a_h = \sum_{i=1}^{n_h} y_{hi}, \quad a = \sum_h \sum_i y_{hi} = \sum_{h=1}^L a_h.$$

The sample proportion of h -th stratum is

$$p_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{a_h}{n_h}, \quad q_h = 1 - p_h = 1 - \frac{a_h}{n_h}.$$

The sample proportion from all sample units is

$$p = \frac{1}{n} \sum \sum y_{hi} = \frac{1}{n} \sum_h a_h = \frac{1}{n} \sum_{h=1}^L n_h p_h = \frac{a}{n}.$$

Also, we have

$$\begin{aligned} s_h^2 &= \frac{1}{n_h - 1} \left[\sum y_{hi}^2 - \frac{(\sum y_{hi})^2}{n_h} \right] \\ &= \frac{1}{n_h - 1} \left[a_h - \frac{a_h^2}{n_h} \right] = \frac{n_h p_h q_h}{n_h - 1}. \end{aligned}$$

The problem is to estimate P and $S^2 = \frac{NPQ}{N-1}$.

Corollary : In stratified random sampling the unbiased estimator of population proportion P is given by

$$p_{st} = \frac{1}{N} \sum_{h=1}^L N_h p_h = \sum_{h=1}^L W_h p_h.$$

Here $E(p_{st}) = \frac{1}{N} \sum N_h E(p_h) = \frac{1}{N} \sum N_h P_h$, [$\because p_h$ is an estimate obtained from simple random sampling from h -th stratum].

Corollary : In stratified random sampling under proportional allocation the variance of the estimate of population proportion is given by

$$\begin{aligned} V(p_{st})_{\text{prop}} &= \frac{1-f}{n} \sum_h \frac{W_h N_h P_h Q_h}{N_h - 1} \\ &= \frac{1-f}{n} \sum_h W_h P_h Q_h, \text{ if } N_h = N_h - 1, \text{ when } N_h \text{ is large enough} \\ &= \frac{1}{n} \sum_h W_h P_h Q_h, \text{ if f.p.c. is neglected.} \end{aligned}$$

We have $V(\bar{y}_{st}) = \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{S_h^2}{n_h}$

and
$$V(\bar{y}_{st})_{prop} = \frac{1}{N^2} \sum N_h \left(N_h - \frac{n}{N} N_h \right) \frac{S_h^2}{\frac{n}{N} N_h} = \sum W_h (1-f) \frac{S_h^2}{n}.$$

$$\therefore V(p_{st})_{prop} = \frac{1-f}{n} \sum_h \frac{W_h N_h P_h Q_h}{N_h - 1}, \quad \because S_h^2 = \frac{N_h P_h Q_h}{N_h - 1}.$$

Corollary : In stratified random sampling under proportional allocation the unbiased estimate of variance of estimate of population proportion (p_{st}) is given by

$$\begin{aligned} v(p_{st})_{prop} &= \frac{1-f}{n} \sum_h \frac{W_h n_h p_h q_h}{n_h - 1} \\ &= \frac{1}{n} \sum_h \frac{W_h n_h p_h q_h}{n_h - 1}, \text{ if f.p.c. is neglected.} \end{aligned}$$

Corollary : In stratified random sampling under proportional allocation and with replacement the variance of estimate of population proportion is given by

$$V(p_{st})_{prop} = \frac{1}{n} \sum_h W_h P_h Q_h.$$

Corollary : In stratified random sampling under proportional allocation and with replacement the estimate of $V(p_{st})$ is given by

$$v(p_{st})_{prop} = \frac{1}{n} \sum_h \frac{W_h n_h p_h q_h}{n_h - 1}.$$

Corollary : In stratified random sampling under proportional allocation the estimate of population total units possessing a characteristic is given by

$$\hat{A}_{st} = N p_{st} = \sum_h N_h p_h.$$

Corollary : In stratified random sampling under proportional allocation the variance of \hat{A}_{st} is given by

$$\begin{aligned} V(\hat{A}_{st}) &= \frac{N^2(1-f)}{n} \sum_h \frac{W_h N_h P_h Q_h}{N_h - 1} = \frac{N^2(1-f)}{n} \sum_h W_h P_h Q_h, \text{ if } N_h \approx N_h - 1 \\ &= \frac{N}{n} \sum_h N_h P_h Q_h, \text{ if f.p.c. is neglected.} \end{aligned}$$

Corollary : In stratified random sampling under proportional allocation and with replacement the variance of \hat{A}_{st} is given by

$$V(\hat{A}_{st}) = \frac{N}{n} \sum_h N_h P_h Q_h.$$

Corollary : In stratified random sampling under proportional allocation the estimate of $V(\hat{A}_{st})$ is given by

$$v(\hat{A}_{st}) = \frac{N^2(1-f)}{n} \sum_h \frac{W_h n_h p_h q_h}{n_h - 1} = \frac{N}{n} \sum_h \frac{N_h n_h p_h q_h}{n_h - 1}, \text{ if f.p.c. is neglected.}$$

Corollary : In stratified random sampling under Neyman allocation the variance of p_{st} is given by

$$\begin{aligned} V(p_{st})_{\text{opt}} &= \frac{1}{n} \left(\sum_h W_h \sqrt{\frac{N_h p_h q_h}{N_h - 1}} \right)^2 - \frac{1}{N} \sum W_h \frac{N_h p_h q_h}{N_h - 1} \\ &= \frac{1}{n} \left(\sum W_h \sqrt{P_h Q_h} \right)^2 - \frac{1}{N} \sum W_h P_h Q_h, \text{ if } N_h \approx N_h - 1 \\ &= \frac{1}{n} \left(\sum W_h \sqrt{\frac{N_h P_h Q_h}{N_h - 1}} \right)^2, \text{ if f.p.c. is neglected} \\ &= \frac{1}{n} \left(\sum W_h \sqrt{P_h Q_h} \right)^2, \text{ if } N_h \approx N_h - 1. \end{aligned}$$

Corollary : In stratified random sampling under Neyman allocation the estimate of $V(p_{st})$ is given by

$$\begin{aligned} v(p_{st})_{\text{opt}} &= \frac{1}{n} \left(\sum_h W_h \sqrt{\frac{n_h p_h q_h}{n_h - 1}} \right)^2 - \frac{1}{N^2} \sum_h \frac{N_h n_h p_h q_h}{n_h - 1} \\ &= \frac{1}{n} \left(\sum_h W_h \sqrt{\frac{n_h p_h q_h}{n_h - 1}} \right)^2, \text{ if f.p.c. is neglected.} \end{aligned}$$

Corollary : In stratified random sampling under Neyman allocation $V(\hat{A}_{st})$ is given by

$$\begin{aligned} V(\hat{A}_{st})_{\text{opt}} &= \frac{1}{n} \left(\sum_h N_h \sqrt{\frac{N_h P_h Q_h}{N_h - 1}} \right)^2 - \sum_h \frac{N_h^2 P_h Q_h}{N_h - 1} \\ &= \frac{1}{n} \left(\sum_h N_h \sqrt{P_h Q_h} \right)^2 - \sum_h N_h P_h Q_h, \text{ if } N_h \approx N_h - 1. \end{aligned}$$

Corollary : In stratified random sampling under Neyman allocation the estimate of $V(\hat{A}_{st})$ is given by

$$v(\hat{A}_{st}) = \frac{1}{n} \left(\sum_h N_h \sqrt{\frac{n_h p_h q_h}{n_h - 1}} \right)^2 - \sum_h N_h \frac{n_h p_h q_h}{n_h - 1}.$$

Corollary : In stratified random sampling with replacement and under Neyman allocation the estimate of $V(p_{st})$ is given by

$$v(p_{st})_{\text{opt}} = \frac{1}{n} \left(\sum_h W_h \sqrt{\frac{n_h p_h q_h}{n_h - 1}} \right)^2.$$

Corollary : In stratified random sampling the $100(1 - \alpha)\%$ confidence interval of P is given by

$$\begin{aligned} p_{st} \pm Z_{\frac{\alpha}{2}} \text{ s.e. } (p_{st}) \\ p_{st} - Z_{\frac{\alpha}{2}} \text{ s.e. } (p_{st}) < P < p_{st} + Z_{\frac{\alpha}{2}} \text{ s.e. } (p_{st}), \text{ where s.e. } (p_{st}) = \sqrt{v(p_{st})}. \end{aligned}$$

Example 13.3 : In a rural area there are 1500 couples of child-bearing ages. These couples are classified into 4 classes according to the levels of education of female. The number of couples in each class are as follows :

Illiterate, $N_1 = 225$

Primary educated, $N_2 = 432$

Secondary educated, $N_3 = 648$

Higher secondary and above, $N_4 = 195$.

Twenty five per cent couples are selected under proportional allocation scheme, and the family planning adoption behaviour of these selected couples are recorded on investigation. The number of selected couples from different strata and number of couples adopting family planning method are shown below :

$$\begin{array}{l} n_h : 56, 108, 162, 49, n = 375 \\ a_h : 21, 54, 96, 38, a = 209 \end{array}$$

Find 95% confidence interval for the population proportion of adopter couples. Also find 95% confidence interval for total number of adopter couples.

Solution : We have $p_1 = \frac{a_1}{n_1} = \frac{21}{56}$, $p_2 = \frac{a_2}{n_2} = \frac{54}{108}$, $p_3 = \frac{a_3}{n_3} = \frac{96}{162}$, $p_4 = \frac{a_4}{n_4} = \frac{38}{49}$

$$\begin{array}{llll} = 0.375 & = 0.50 & = 0.593 & = 0.776 \end{array}$$

$$s_1^2 = \frac{n_1 p_1 q_1}{n_1 - 1} = \frac{56 \times 0.375 \times 0.625}{56 - 1} = 0.23864, s_1 = 0.4885$$

$$s_2^2 = \frac{n_2 p_2 q_2}{n_2 - 1} = \frac{108 \times 0.50 \times 0.50}{108 - 1} = 0.25234, s_2 = 0.50233$$

$$s_3^2 = \frac{n_3 p_3 q_3}{n_3 - 1} = \frac{162 \times 0.593 \times 0.407}{162 - 1} = 0.24286, s_3 = 0.4928$$

$$s_4^2 = \frac{n_4 p_4 q_4}{n_4 - 1} = \frac{49 \times 0.776 \times 0.224}{49 - 1} = 0.17745, s_4 = 0.42124.$$

The estimate of population proportion of adopter couples is

$$\begin{aligned} p_{st} &= \frac{1}{N} \sum_{h=1}^4 N_h p_h = \frac{1}{1500} [225 \times 0.375 + 432 \times 0.5 + 648 \times 0.593 + 195 \times 0.776] \\ &= 0.557. \end{aligned}$$

Also, we have $f = \frac{n}{N} = \frac{375}{1500} = 0.25$.

$$W_1 = \frac{N_1}{N} = \frac{225}{1500} = 0.15, W_2 = \frac{N_2}{N} = \frac{432}{1500} = 0.288, W_3 = \frac{N_3}{N} = \frac{648}{1500} = 0.432,$$

$$W_4 = \frac{N_4}{N} = \frac{195}{1500} = 0.13.$$

The estimate of variance of p_{st} under proportional allocation is

$$\begin{aligned} v(p_{st}) &= \frac{1-f}{n} \sum_{h=1}^L \frac{w_h n_h p_h q_h}{n_h - 1} = \frac{1-f}{n} \sum W_h s_h^2 \\ &= \frac{1-0.25}{375} [0.15 \times 0.23864 + 0.288 \times 0.25234 + 0.432 \times 0.24286 + 0.13 \times 0.17745] \\ &= 0.0004729. \end{aligned}$$

$$\text{s.e.}(p_{st}) = \sqrt{v(p_{st})} = \sqrt{0.0004729} = 0.02175.$$

95% confidence interval for the population proportion $[P]$ of adopter couples is

$$p_{st} \pm Z_{0.025} \sqrt{\text{s.e.}(p_{st})},$$

where $\hat{P}_L = p_{st} - Z_{0.025} \sqrt{\text{s.e.}(p_{st})} = 0.557 - 1.96 \times 0.02175 = 0.514.$

$$\hat{P}_U = p_{st} + Z_{0.025} \sqrt{\text{s.e.}(p_{st})} = 0.557 + 1.96 \times 0.02175 = 0.600.$$

Again, the estimate of total adopter couples in the population is

$$\hat{A}_{st} = Np_{st} = 1500 \times 0.557 = 836.$$

The estimate of variance of \hat{A}_{st} is

$$v(\hat{A}_{st}) = N^2v(p_{st}) = (1500)^2 \times 0.0004729 = 1064.025.$$

$$\text{s.e.}(\hat{A}_{st}) = \sqrt{v(\hat{A}_{st})} = \sqrt{1064.025} = 32.6194.$$

95% confidence interval for population total $[A]$ of adopter couples is

$$\hat{A}_{st} \pm Z_{0.025} \text{ s.e.}(\hat{A}_{st}),$$

where $\hat{A}_L = \hat{A}_{st} - Z_{0.025} \text{ s.e.}(\hat{A}_{st}) = 836 - 1.96 \times 32.6194 = 772.$

$$\hat{A}_U = \hat{A}_{st} + Z_{0.025} \text{ s.e.}(\hat{A}_{st}) = 836 + 1.96 \times 32.6194 = 900.$$

Example 13.4 : In a diagnostic centre 2000 patients attended for blood test, specially for grouping of blood and for F.B.S. From this population 11% patients were selected under Neyman allocation, where patients were classified into 7 classes, viz., patients of blood group A^+ , A^- , B^+ , B^- , O^+ , O^- and others. The information of population units and sample units are shown below :

No. of units in		Number of units in the sample having normal F.B.S.	Blood group
Population N_h	Sample n_h		
815	98	62 = a_1	A^+
218	17	11 = a_2	A^-
337	42	23 = a_3	B^+
162	20	8 = a_4	B^-
352	30	18 = a_5	O^+
80	8	5 = a_6	O^-
36	5	2 = a_7	others
2000	220		

Find 95% confidence interval for population proportion of patients having normal F.B.S. Also find 95% confidence interval for the total patients having normal F.B.S.

Solution : We have

$$p_1 = \frac{a_1}{n_1} = \frac{62}{98} = 0.63, \quad p_2 = \frac{a_2}{n_2} = \frac{11}{17} = 0.65, \quad p_3 = \frac{a_3}{n_3} = \frac{23}{42} = 0.55,$$

$$p_4 = \frac{a_4}{n_4} = \frac{8}{20} = 0.40, \quad p_5 = \frac{a_5}{n_5} = \frac{18}{30} = 0.60, \quad p_6 = \frac{a_6}{n_6} = \frac{5}{8} = 0.625,$$

$$p_7 = \frac{a_7}{n_7} = \frac{2}{5} = 0.40, \quad s_1^2 = \frac{n_1 p_1 q_1}{n_1 - 1} = \frac{98 \times 0.63 \times 0.37}{98 - 1} = 23.5503,$$

$$s_2^2 = \frac{n_2 p_2 q_2}{n_2 - 1} = \frac{17 \times 0.65 \times 0.35}{17 - 1} = 0.2417, \quad s_3^2 = \frac{n_3 p_3 q_3}{n_3 - 1} = \frac{42 \times 0.55 \times 0.45}{42 - 1} = 0.2585,$$

$$s_4^2 = \frac{n_4 p_4 q_4}{n_4 - 1} = \frac{20 \times 0.40 \times 0.60}{20 - 1} = 0.2526, \quad s_5^2 = \frac{n_5 p_5 q_5}{n_5 - 1} = \frac{30 \times 0.60 \times 0.40}{30 - 1} = 0.2483,$$

$$s_6^2 = \frac{n_6 p_6 q_6}{n_6 - 1} = \frac{8 \times 0.625 \times 0.375}{8 - 1} = 0.2679, \quad s_7^2 = \frac{n_7 p_7 q_7}{n_7 - 1} = \frac{5 \times 0.4 \times 0.6}{5 - 1} = 0.3000.$$

The estimate of population proportion of patients is given by

$$p_{st} = \sum_{h=1}^7 \frac{N_h p_h}{N}$$

$$= \frac{1}{2000} [815 \times 0.63 + 218 \times 0.65 + 337 \times 0.55 + 162 \times 0.40$$

$$+ 352 \times 0.60 + 80 \times 0.625 + 36 \times 0.40]$$

$$= \frac{1278.5}{2000} = 0.64.$$

The estimate of variance of p_{st} under Neyman allocation is

$$v(p_{st}) = \frac{1}{n} \left(\sum W_h \sqrt{\frac{n_h p_h q_h}{n_h - 1}} \right)^2 - \frac{1}{N^2} \sum \frac{N_h n_h p_h q_h}{n_h - 1}$$

$$= \frac{1}{n} \left(\sum W_h s_h \right)^2, \text{ if } N \text{ is large (f.p.c. is neglected).}$$

We have $W_h = \frac{N_h}{N}$.

$$\therefore W_1 = \frac{N_1}{N} = \frac{815}{2000} = 0.4075, \quad W_2 = \frac{N_2}{N} = \frac{218}{2000} = 0.109, \quad W_3 = \frac{N_3}{N} = \frac{337}{2000} = 0.1685,$$

$$W_4 = \frac{N_4}{N} = \frac{162}{2000} = 0.081, \quad W_5 = \frac{N_5}{N} = \frac{352}{2000} = 0.176, \quad W_6 = \frac{N_6}{N} = \frac{80}{2000} = 0.04,$$

$$W_7 = \frac{N_7}{N} = \frac{36}{2000} = 0.018,$$

$$s_1 = 4.8529, \quad s_2 = 0.4916, \quad s_3 = 0.5035, \quad s_4 = 0.5026, \quad s_5 = 0.4983,$$

$$s_6 = 0.5176, \quad s_7 = 0.5477.$$

$$v(p_{st}) = \frac{1}{n} \left(\sum W_h s_h \right)^2, \quad \because N = 2000 \text{ is large}$$

$$= \frac{1}{220} [0.4075 \times 4.8529 + 0.109 \times 0.4916 + 0.1685 \times 0.5035$$

$$+ 0.081 \times 0.5026 + 0.176 \times 0.4983 + 0.04 \times 0.5176 + 0.018 \times 0.5477]^2$$

$$= 0.021067.$$

$$\text{s.e. } (p_{st}) = \sqrt{v(p_{st})} = \sqrt{0.021067} = 0.1451.$$

95% confidence interval of population proportion of patients having normal F.B.S. is given by

$p_{st} \pm Z_{0.025} \text{ s.e. } (p_{st})$, where

$$\hat{P}_L = p_{st} - Z_{0.025} \text{ s.e. } (p_{st}) = 0.64 - 1.96 \times 0.1451 = 0.356.$$

$$\hat{P}_U = p_{st} + Z_{0.025} \text{ s.e. } (p_{st}) = 0.64 + 1.96 \times 0.1451 = 0.924.$$

The estimate of total patient having normal F.B.S. is

$$\hat{A}_{st} = Np_{st} = 2000 \times 0.64 = 1280.$$

The estimate of variance of \hat{A}_{st} is

$$v(\hat{A}_{st}) = N^2 v(p_{st}) = (2000)^2 \times 0.021067 = 84268.$$

$$\text{s.e.}(\hat{A}_{st}) = \sqrt{v(\hat{A}_{st})} = \sqrt{84268} = 290.29.$$

95% confidence interval for total patients having normal F.B.S. is

$$\hat{A}_{st} \pm Z_{0.025} \text{s.e.}(\hat{A}_{st}),$$

where

$$\hat{A}_L = \hat{A}_{st} - Z_{0.025} \text{s.e.}(\hat{A}_{st}) = 1280 - 1.96 \times 290.29 = 711.$$

$$\hat{A}_U = \hat{A}_{st} + Z_{0.025} \text{s.e.}(\hat{A}_{st}) = 1280 + 1.96 \times 290.29 = 1849.$$

13.4 Relative Precision of Simple Random and Stratified Random Sampling

It is already mentioned that Neyman allocation is done by minimizing the variance of the estimator. Thus, it is expected that the stratified random sampling under Neyman allocation is more efficient than proportional allocation. Also, it is mentioned that stratified random sampling is more applicable than simple random sampling if the population observations are more heterogeneous. Therefore, we can compare the precision of simple random sampling, stratified random sampling under proportional allocation and under optimum allocation.

Theorem : If $\frac{1}{N_h}$ is negligible, then $V_{\text{ran}} \geq V_{\text{prop}} \geq V_{\text{opt}(N)}$, where

$V_{\text{ran}} = V(\bar{y})$ in simple random sampling,

$V_{\text{prop}} = V(\bar{y})$ under proportional allocation in stratified random sampling,

$V_{\text{opt}(N)} = V(\bar{y})$ under Neyman allocation in stratified random sampling.

Proof : If $\frac{1}{N_h}$ is negligible, $\frac{1}{N}$ is also negligible. We know that

$$V(\bar{y})_{\text{ran}} = \frac{1-f}{n} S^2, \quad V(\bar{y})_{\text{prop}} = \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 \text{ and}$$

$$V(\bar{y})_{\text{opt}(N)} = \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^L W_h S_h^2.$$

Let y_{hi} be the value of the variable under study for i -th unit in h -th stratum ($h = 1, 2, \dots, L$; $i = 1, 2, \dots, N_h$). Then, total sum of squares of N observations is ($N = \sum N_h$).

$$\begin{aligned} (N-1)S^2 &= \sum_{h=1}^L \sum_{i=1}^{N_h} (y_{hi} - \bar{Y})^2 \\ &= \sum_{h=1}^L \sum_{i=1}^{N_h} [(y_{hi} - \bar{Y}_h) + (\bar{Y}_h - \bar{Y})]^2 \\ &= \sum_{h=1}^L \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 \\ &= \sum_{h=1}^L (N_h - 1)S_h^2 + \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2. \end{aligned}$$

Dividing both sides by N and neglecting the term $\frac{1}{N}$ and $\frac{1}{N_h}$, we get

$$S^2 = \sum W_h S_h^2 + \sum W_h (\bar{Y}_h - \bar{Y})^2$$

$$(1-f) \frac{S^2}{n} = \frac{(1-f)}{n} \sum W_h S_h^2 + \frac{(1-f)}{n} \sum W_h (\bar{Y}_h - \bar{Y})^2.$$

$\therefore V_{\text{ran}} = V_{\text{prop}} + \text{positive quantity} \Rightarrow V_{\text{ran}} \geq V_{\text{prop}}$.

The equality sign holds good, if $\bar{Y}_h = \bar{Y}$.

Again,

$$V_{\text{prop}} - V_{\text{opt}(N)} = \frac{1-f}{n} \sum_h^L W_h S_h^2 - \frac{1}{n} \left(\sum_h^L W_h S_h \right)^2 + \frac{1}{N} \sum_h^L W_h S_h^2$$

$$= \frac{1}{n} \sum W_h S_h^2 - \frac{1}{n} \left(\sum W_h S_h \right)^2, \quad \because \text{f.p.c. is neglected}$$

$$= \frac{1}{Nn} \left[\sum N_h S_h^2 - \frac{(\sum N_h S_h)^2}{N} \right]$$

$$= \frac{1}{Nn} \sum N_h (S_h - \bar{S})^2, \quad \text{where } \bar{S} = \frac{1}{N} \sum N_h S_h.$$

$\therefore V_{\text{prop}} - V_{\text{opt}(N)} = +\text{ve quantity} \Rightarrow V_{\text{prop}} \geq V_{\text{opt}(N)}$.

Therefore, $V_{\text{ran}} \geq V_{\text{prop}} \geq V_{\text{opt}(N)}$.

From the above result, we have

$$V_{\text{ran}} = V_{\text{prop}} + \frac{1-f}{n} \sum (Y_h - \bar{Y})^2$$

$$= V_{\text{opt}(N)} + \frac{1}{Nn} \sum N_h (S_h - \bar{S})^2 + \frac{1-f}{n} \sum (\bar{Y}_h - \bar{Y})^2.$$

It is observed that the variance of \bar{y}_{st} is minimum if sample is selected under Neyman allocation. The variance is minimum compared to the variance of \bar{y} under simple random sampling. The minimum variance of $\bar{y}_{st}[V(\bar{y}_{st})_{\text{opt}(N)}]$ depends on two quantities, viz., S_h^2 and \bar{Y}_h . Therefore, it may be concluded that Neyman allocation will provide more efficient estimate of population mean if the variance of observations within a stratum is more and if the \bar{Y}_h 's are more heterogeneous. If the observations of different strata are homogeneous, the simple random sampling may be used as a better sampling technique.

We have $(N-1)S^2 = \sum (N_h - 1)S_h^2 + \sum N_h (\bar{Y}_h - \bar{Y})^2$.

$$\frac{1-f}{n} S^2 = \frac{1-f}{n(N-1)} \sum (N_h - 1)S_h^2 + \frac{1-f}{n(N-1)} \sum N_h (\bar{Y}_h - \bar{Y})^2$$

$$= \frac{1-f}{n} \sum W_h S_h^2 + \frac{1-f}{n(N-1)} \sum N_h (\bar{Y}_h - \bar{Y})^2$$

$$+ \frac{1-f}{n(N-1)} \sum (N_h - 1)S_h^2 - \frac{1-f}{n} \sum W_h S_h^2.$$

$$V_{\text{ran}} = V_{\text{prop}} + \frac{1-f}{n(N-1)} \left[\sum_h N_h (\bar{Y}_h - \bar{Y})^2 - \frac{1}{N} \sum (N - N_h) S_h^2 \right].$$

It is observed that if stratified random sampling is drawn under proportional allocation, then $V(\bar{y}_{st})_{prop}$ is greater than $V(\bar{y})_{ran}$, if

$$\sum N_h(\bar{Y}_h - \bar{Y})^2 < \frac{1}{N} \sum_h (N - N_h) S_h^2.$$

Let us consider that S_h^2 for every stratum is less and let us denote this by S_w^2 . Then

$$\frac{1}{N} \sum (N - N_h) S_h^2 = (L - 1) S_w^2.$$

This implies that

$$\sum N_h(\bar{Y}_h - \bar{Y})^2 < (L - 1) S_w^2 \quad \text{or,} \quad \frac{\sum N_h(\bar{Y}_h - \bar{Y})}{L - 1} < S_w^2.$$

That is, the mean square error within stratum is less than mean square error between strata. If this condition does not prevail, $V(\bar{y})$ will be greater than $V(\bar{y}_{st})_{prop}$.

13.5 Estimation of Gain in Precision due to Stratification

It is observed that estimator of population parameter under stratified random sampling is more efficient than that from simple random sampling. Let us now estimate this precision.

Let n_h ($h = 1, 2, \dots, L$) be the sample size from h -th stratum. The mean and variance from sample of h -th stratum are, respectively \bar{y}_h and s_h^2 . Then the estimate of variance of \bar{y}_{st} is

$$v(\bar{y}_{st}) = \sum \frac{W_h^2 s_h^2}{n_h} - \sum \frac{W_h^2 s_h^2}{N}.$$

Let us compare this variance with $v(\bar{y})$, where \bar{y} is a simple random sample calculated from a sample of n observations. Let

$$s^2 = \frac{1}{n - 1} \sum_{h=1}^L \sum_{i=1}^{n_h} (y_{hi} - \bar{y})^2.$$

We have $v(\bar{y}) = \frac{N - n}{Nn} s^2$, where $v(\bar{y})$ is an unbiased estimator of $V(\bar{y})$. Here

$$V(\bar{y}) = \frac{N - n}{Nn} S^2, \quad \text{where} \quad S^2 = \frac{1}{N - 1} \sum \sum (y_{hi} - \bar{Y})^2.$$

We known that

$$\frac{1}{N} E \left(\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}^2 \right) = \frac{1}{N} \sum_h \sum_i y_{hi}^2.$$

Also, it is known that \bar{y}_{st} is an unbiased estimator of \bar{Y} and $v(\bar{y}_{st})$ is an unbiased estimator of $V(\bar{y}_{st})$, where

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum N_h(N_h - n_h) \frac{S_h^2}{n_h}.$$

Again, $V(\bar{y}_{st}) = E(\bar{y}_{st}^2) - [E(\bar{y}_{st})]^2 = E(\bar{y}_{st}^2) - \bar{Y}^2$.

$\therefore E(\bar{y}_{st}^2) = V(\bar{y}_{st}) + \bar{Y}^2$.

Therefore, the unbiased estimator of \bar{Y}^2 is

$$\bar{y}_{st}^2 - v(\bar{y}_{st}), \quad \text{where } v(\bar{y}_{st}) = \sum_h \frac{W_h^2 s_h^2}{n_h} - \sum \frac{W_h^2 s_h^2}{N}$$

Now,
$$V(\bar{y})_{\text{ran}} = \frac{N-n}{n(N-1)} \left[\frac{1}{N} \sum^L \sum^{N_h} (y_{hi} - \bar{Y})^2 \right] = \frac{N-n}{N(N-1)} \left[\frac{1}{N} \left\{ \sum \sum y_{hi}^2 - \bar{Y}^2 \right\} \right].$$

The unbiased estimator of $V(\bar{y})_{\text{ran}}$ is

$$v(\bar{y})_{\text{ran}} = \frac{N-n}{Nn} s^2, \quad \text{where } s^2 = \frac{1}{n-1} \sum^L \sum^{n_h} (y_{hi} - \bar{y})^2.$$

Replacing s^2 and on simplification, we get

$$v(\bar{y})_{\text{ran}} = \frac{N-n}{n(N-1)} \left[\frac{n-1}{n} s^2 + v(\bar{y}_{st}) \right].$$

If n and N are large, then $n-1 \approx n$ and $N-1 \approx N$ and we have

$$v(\bar{y})_{\text{ran}} = \frac{N-n}{Nn} s^2.$$

Therefore, the relative efficiency of stratified random sampling compared to simple random sampling is

$$\frac{v(\bar{y})_{\text{ran}} - v(\bar{y}_{st})}{v(\bar{y}_{st})} \times 100\%.$$

The above quantity is gain in precision due to stratification compared to simple random sampling. This gain is achieved if stratification is done properly so that the observations within a stratum are more homogeneous and all the strata are heterogeneous.

In practice, the stratification is not possible using the values of the variable, since the values are not known before the survey. However, more precise estimate under stratified random sampling is found out if the following conditions exist in the population :

- (a) The population units are divided into several parts and the sizes of parts vary widely.
- (b) The estimator of variable which is needed to be estimated is associated with the size of the population units.
- (c) The sizes of the strata are estimated properly.

As an example of (a), we can mention the estimation problem of income of employees in different industries. The sizes of industry in terms of employees or in terms of economic activities vary widely. Each industry can be considered a stratum.

If number of employees of an industry is to be estimated, the number is associated with the size of industry. In such a case stratification will usually provide more precise estimate. To estimate the area under wheat cultivation in a state, the state can be divided into region according to the area under wheat cultivation.

Example 13.5 : Estimate the gain in precision due to stratification compared to simple random sampling using the data of example 13.2.

Solution : We have $v(\bar{y}_{st})_{\text{opt}(N)} = 0.0020045$.

$$v(\bar{y})_{\text{ran}} = \frac{N-n}{n(N-1)} \left[\frac{n-1}{n} s^2 + v(\bar{y}_{st}) \right],$$

$$\text{where } s^2 = \frac{1}{n-1} \left[\sum \sum y_{hi}^2 - \frac{(\sum \sum y_{hi})^2}{n} \right] = \frac{1}{50-1} \left[647.05 - \frac{(157.5)^2}{50} \right] = 3.0801.$$

$$\begin{aligned} \therefore v(\bar{y}) &= \frac{345-50}{50(345-1)} \left[\frac{50-1}{50} \times 3.0801 + 0.0020045 \right] \\ &= 0.0518051. \end{aligned}$$

Therefore, the relative gain in precision due to stratification compared to simple random sampling is

$$\frac{v(\bar{y})_{\text{ran}} - v(\bar{y})_{\text{opt}(N)}}{v(\bar{y})_{\text{st}}} \times 100\% = \frac{0.0518051 - 0.0020045}{0.0020045} \times 100\% = 2484.44\%.$$

13.6 Method of Construction of Strata

It has already been mentioned that the stratification is profitable if stratum variances ($S_h^2, h = 1, 2, \dots, L$) are more heterogeneous. Moreover, the technique is more profitable in a case where \bar{Y}_h 's are more heterogeneous. However, the observations within a stratum should be homogeneous. These points are to be kept in mind in constructing strata.

There were many suggestions in constructing strata. Sethi (1963) suggested the method of selecting optimum points for special kind of populations. Later on, Sethi et al. (1966) utilized his earlier technique in real world data. Delenius (1957) proposed the equation to decide the best limit of stratum when number of strata are known. He also found out the equations to decide the boundary of stratum in case of proportional and Neyman allocations.

In practice, the contiguous geographical regions are included in a stratum, specially in any agricultural survey research. The stratification is also done using the characteristics of some related variables related to the study variable. For example, to estimate the total production of jute in an area the stratification can be done using the information of total cultivated land in the smaller segments of the study area. To estimate the total maize production of an area, the area can be divided into strata according to the availability of irrigation facilities. However, the best method of stratification is the use of frequency distribution of the variable under study, though the technique is not feasible in most cases. Therefore, it is better to use the frequency distribution of some variables related to the variable under study. The frequency distribution of the variable observed in any previous survey may also be used as a basis of stratification. Let us discuss the method of stratification for Neyman allocation.

Let y_0 and y_2 be respectively the smallest and the largest values of the variable under study. The stratum boundary of the remaining values y_1, y_2, \dots, y_{L-1} is to be decided in such a way that $V(\bar{y}_{st})_{\text{opt}(N)}$ is minimum, where

$$\begin{aligned} V(\bar{y}_{st})_{\text{opt}(N)} &= \frac{1}{n} \left(\sum_h^L W_h S_h \right)^2 - \frac{1}{N} \sum_h^L W_h S_h^2 \\ &= \frac{1}{n} \left(\sum_h W_h S_h \right)^2, \text{ if f.p.c. is neglected.} \end{aligned}$$

The $V(\bar{y}_{st})$ will be minimum if $\sum W_h S_h$ is minimum. To minimize the quantity $\sum W_h S_h$, it is to be differentiated with respect to y_h , where y_h is used in calculating S_h as well as S_{h+1} . Therefore, to minimize $\sum W_h S_h$ with respect to y_h , we get

$$\frac{\partial}{\partial y_h} \sum W_h S_h = \frac{\partial}{\partial y_h} (W_h S_h) + \frac{\partial}{\partial y_h} (W_{h+1} S_{h+1}).$$

Let $f(y)$ be the frequency distribution of y , then

$$W_h = \int_{Y_{h-1}}^{y_h} f(t)dt, \quad \frac{\partial W_h}{\partial y_h} = f(y_h).$$

Again,
$$W_h S_h^2 = \int_{y_{h-1}}^{y_h} t^2 f(t)dt - \frac{\left[\int_{y_{h-1}}^{y_h} t f(t)dt \right]^2}{\int_{y_{h-1}}^{y_h} f(t)dt}.$$

Now, differentiating $W_h S_h^2$, we get

$$S_h^2 \frac{\partial W_h}{\partial y_h} + 2W_h S_h \frac{\partial S_h}{\partial y_h} = y_h^2 f(y_h) - 2y_h \mu_h f(y_h) + S_h^2 f(y_h).$$

Here μ_h is the mean of y_h in h -th stratum. On simplification, we get

$$\frac{\partial (W_h S_h)}{\partial y_h} = S_h \frac{\partial W_h}{\partial y_h} + W_h \frac{\partial S_h}{\partial y_h} = \frac{1}{2} f(y_h) \frac{(y_h - \mu_h)^2 + S_h^2}{S_h}.$$

Similarly,
$$\frac{\partial (W_{h+1} y_{h+1})}{\partial y_h} = -\frac{1}{2} f(y_h) \frac{(y_h - \mu_{h+1})^2 - S_{h+1}^2}{S_{h+1}}.$$

Therefore, the equation for y_h is

$$\frac{(y_h - \mu_h)^2 + S_h^2}{S_h} = \frac{(y_h - \mu_{h+1})^2 + S_{h+1}^2}{S_{h+1}}, \quad h = 1, 2, \dots, L - 1.$$

However, the application of this equation is not easy in practice. We need to discuss some simple technique to form strata. Some methods are stated below :

(i) **Equalization of $W_h S_h$** : Delinius and Gurney (1951) proposed the technique of construction of strata equalizing $W_h S_h$. They have also proposed to select equal number of units from each stratum. But the technique needs the value of S_h^2 and it is not available. So, the application of the technique is not suitable.

(ii) **Equalization of strata totals** : Mahalanobis (1952) proposed this technique. He has proposed to select equal units from those strata for which $W_h Y_h$ are equals. But the technique is not suitable to select units from normal distribution, Gamma distribution and Beta distribution.

(i) **Equalization of $W_h R_h$** : Ayoma (1954) proposed this technique. According to him $W_h R_h$'s are to be equalized and to form strata and equal units are to be selected from those strata for which $W_h R_h$'s are equal. Here R_h is the range of values of y_h .

13.7 Number of Strata

The number of strata is pre-determined when stratified random sample is selected. But the question arises how to decide the number of strata. The following two points are to be kept in mind in determining number of strata :

- (i) To observe whether $V(\bar{y}_{st})$ decreases with the increase in number of strata.
- (ii) To observe the change in cost of survey when strata increase.

There are many suggestions to decide the number of strata. Cochran (1977) explained the method as follows :

Consider that the values of the variable Y are used in constructing strata. Also consider that Y is distributed uniformly over the range $[a, a + d]$. Then the variance of Y is

$$S_y^2 = \frac{d^2}{12} \quad \text{and} \quad V(\bar{y}) = \frac{d^2}{12n};$$

where \bar{y} is the sample mean of n observations. If N_h 's are equal ($h = 1, 2, \dots, h$), then the variance of the observations in h -th stratum is

$$S_{yh}^2 = \frac{d^2}{12L^2}, \quad \text{where } W_h = \frac{1}{L}.$$

$$\text{Therefore, } V(\bar{y}_{st}) = \frac{1}{n} \left(\sum W_h S_{yh} \right)^2 = \frac{1}{n} \left(\sum \frac{1}{L} \frac{d}{L\sqrt{12}} \right)^2 = \frac{d^2}{12nL^2} = \frac{V(\bar{y})}{n}.$$

Thus, it is observed that, for a rectangular distribution $V(\bar{y}_{st})$ is decreased inversely to the square of strata number (L^2). Similar result is also observed in selecting sample under Neyman allocation from skewed population having finite range. Cochran (1961) studied the ratio $V(\bar{y}_{st})/V(\bar{y})$ for rectangular distribution and skewed distribution. He also found out $V(\bar{y}_{st})/V(\bar{y})$ in selecting sample from skewed distribution, when $L = 2, 3, 4$, where the ratios are 0.232, 0.098 and 0.053, respectively. The corresponding values in selecting sample from rectangular distribution are 0.250, 0.111 and 0.062. This results indicate that $V(\bar{y}_{st})$ decreases with the increase in number of strata. However, the argument is not always true, specially if stratification is done using information of variable correlated to the study variable. Cochran has showed that unless $\rho_{xy} > 0.95$ for $L = 6$, the $V(\bar{y}_{st})$ is not decreased so much.

Sethi (1963) has observed that if L exceeds 6, the sampling results are not improved so much. This is observed in analysing the cost function when strata sizes are changed. Moreover, if strata number increases, the sample size n needs to be decreased to complete the survey within the specified cost.

Let us study the relationship of cost function and sample size n . Let the cost function be

$$C = C_0 + LC_1 + nC_2,$$

where C_0 is the overall cost of survey, C_1 and C_2 are the costs of survey of each stratum and each unit, respectively. Sethi (1963) has showed that

$$V(\bar{y}_{st}) = \frac{S^2}{n} (bL^2 + CL + d)^{-1}$$

in selecting sample from Gamma distribution under proportional allocation and equal allocation. Here the values of b, c and d are calculated from variance ratio when $L = 1, 2$ and 3. The minimization of this variance for the above cost function gives

$$L = 2(C - C_0)/3C_1$$

$$n = (C - C_0)/3C_2.$$

13.8 Effects of Error Due to Estimation of Stratum Size

It is already noted that in stratified random sampling the sample stratum size n_h depends on population stratum size N_h ($h = 1, 2, \dots, L$), specially if sample is selected under proportional allocation and Neyman allocation. The problem in selecting sample arises if N_h is not known or if N_h is replaced from the knowledge of some previous survey results. The previous survey information on N_h may not be valid for current survey. In such a case $W_h = \frac{N_h}{N}$ is not used properly to have the estimate, rather its estimate ω_h is used, where

$$\bar{y}_{st} = \sum_{h=1}^L \omega_h \bar{y}_h.$$

Here ω_h is not the exact value of N_h/N ; $h = 1, 2, \dots, L$.

Due to the use of ω_h instead of W_h the following error arises in the sample estimates :

- (i) The sample estimate is not unbiased. Since the estimator is biased, the mean square error of estimate instead of variance is used to estimate the accuracy of the estimator. But the accuracy of the estimator should be estimated on the basis of the mean of the estimator.
- (ii) The amount of bias is not reduced even the sample size is increased. As a result the stratified random sampling for a fixed sample size is less precise than the simple random sampling.
- (iii) The standard error of \bar{y}_{st} provides the downward biased estimator of error of \bar{y}_{st} .

Let us investigate the mean square error of \bar{y}_{st} , when \bar{y}_{st} is biased. Let the biased estimator of population mean is

$$\bar{y}_{st} = \sum_{h=1}^L \omega_h \bar{y}_h, \text{ where } \omega_h \text{ is the estimator of } W_h = \frac{N_h}{N}.$$

Then
$$E(\bar{y}_{st}) = \sum_{h=1}^L \omega_h \bar{Y}_h.$$

Against
$$E(\bar{y}_{st})_U = \sum_{h=1}^L W_h \bar{Y}_h, \text{ where } (\bar{y}_{st})_U = \text{unbiased estimator.}$$

Therefore, the bias of estimator is

$$= \sum_h (\omega_h - W_h) \bar{Y}_h.$$

This bias is free of sample size. Now, the MSE (\bar{y}_{st}) is

$$\text{MSE}(\bar{y}_{st}) = \sum \frac{\omega_h^2 S_h^2}{n_h} (1 - f_h) + \left[\sum (\omega_h - W_h) \bar{Y}_h \right]^2.$$

But the variance of $(\bar{y}_{st})_U$ is

$$V(\bar{y}_{st})_v = \sum \frac{W_h^2 S_h^2}{n_h} (1 - f_h).$$

It is observed that $V(\bar{y}_{st})_v$ is less than $V(\bar{y}_{st})$.

13.9 Determination of Sample Size

The determination of sample size has been discussed in section 11.11. In this section, we shall discuss the determination of sample size in stratified random sampling. Let us discuss the method in dealing with (a) continuous variable and (b) qualitative variable or discrete variable.

Determination of sample size in case of continuous variable : Let n_h, \bar{y}_h and s_h be the sample size, sample mean and sample standard deviation of h -th stratum. Let S_h be the estimator of S_h , where S_h is the population standard deviation of h -th stratum. Let us consider $n_h = \omega_h n$, where ω_h is to be assumed. Then the variance of \bar{y}_{st} is

$$V = V(\bar{y}_{st}) = \frac{1}{n} \sum \frac{W_h s_h^2}{\omega_h} - \frac{1}{N} \sum W_h s_h^2, \text{ where } W_h = \frac{N_h}{N}.$$

We have
$$n = \frac{\sum_h \frac{W_h s_h^2}{\omega_h}}{V + \frac{1}{N} \sum W_h s_h^2}.$$

If f.p.c. is neglected, then the preliminary value of n is, say, n_0 , where

$$n_0 = \frac{1}{V} \sum_h \frac{W_h s_h^2}{\omega_h}.$$

If n_0/N is not negligible, then

$$n = \frac{n_0}{1 + \frac{1}{NV} \sum W_h s_h^2}.$$

The value of n depends on allocation of sample sizes also. Let us consider that

$$n_h \propto N_h s_h.$$

Then
$$n = \frac{(\sum W_h s_h)^2}{V + \frac{1}{N} \sum W_h s_h^2}.$$

Again, $n_h \propto N_h$, then $n_0 = \frac{\sum W_h s_h^2}{V}$ and $n = \frac{n_0}{1 + \frac{n_0}{N}}$.

In practice, the value of V is not known. Let the error in estimator be d and it is pre-determined. Then $V = (d/z)^2$, where z is the value of standard normal variate.

Determination of sample size in case of qualitative variable : If the variable under study is qualitative in nature, we usually estimate the population proportion, where population proportion of a characteristic is P . The estimate of P is p_{st} , where $p_{st} = \sum W_h p_h$ and $p_h = \frac{a_h}{n_h}$. Here a_h is the total number of sample units in h -th stratum possessing a characteristic. The sample variance of the characteristic of h -th stratum is

$$s_h^2 = \frac{np_h q_h}{n_h - 1}.$$

Then
$$V(p_{st}) = \sum_{h=1}^L \frac{W_h P_h Q_h}{n_h} (1 - f_h).$$

If $n_h \propto N_h$, $V(p_{st})_{prop} = \frac{1-f}{n} \sum_{h=1}^L W_h P_h Q_h$.

The estimate of $V(p_{st})_{prop}$ is given by

$$v(p_{st})_{prop} = \frac{1-f}{n} \sum \frac{W_h p_h q_h}{n_h - 1}.$$

Again, if $n_h \propto N_h s_h$ or, if $n_h \propto \sqrt{\frac{N_h}{N_h - 1} P_h Q_h}$,

then
$$V(p_{st})_{opt(N)} = \frac{1}{n} \left(\sum_h W_h \sqrt{\frac{N_h P_h Q_h}{N_h - 1}} \right)^2 - \frac{1}{N^2} \sum \frac{N_h^2 P_h Q_h}{N_h - 1}.$$

The estimate of this variance is

$$v(p_{st})_{opt(N)} = \frac{1}{n} \left(\sum_h W_h \sqrt{\frac{n_h p_h q_h}{n_h - 1}} \right)^2 - \frac{1}{N^2} \sum_h \frac{N_h n_h p_h q_h}{n_h - 1}.$$

The $V(p_{st})$ will be minimum, if

$$n_h \approx \frac{N_h \sqrt{P_h Q_h}}{\sum (N_h \sqrt{P_h Q_h})} n \quad \text{or,} \quad n \approx \frac{n_h \sum N_h \sqrt{P_h Q_h}}{N_h \sqrt{P_h Q_h}}.$$

Let $V(p_{st})_{prop} = V$ and let the assumed value of V be $n_0 = \frac{\sum W_h p_h q_h}{V}$.

Then $n = \frac{n_0}{1 + \frac{n_0}{N}}$.

If $V(p_{st})_{opt(N)} = V$, then $n_0 = \frac{(\sum W_h \sqrt{p_h q_h})^2}{V}$

and $n = \frac{n_0}{1 + \frac{1}{NV} \sum W_h p_h q_h}$.

However, the value of V is not known to us and it cannot be estimated before the survey. But, we can assume that the estimator should have error d (say). Then $V = (d/z)^2$ can be calculated, where z is the value of normal variate.

Example 13.6 : To estimate the milk production per cow in an area, the area has been divided into strata according to the number of cows in stratum. The stratum size and stratum sample variance of milk production are given below :

Stratum SL. No.	Stratum Size, N_h	s_h^2 of milk production	Total milk production
1	115	1162	20042 kg
2	85	1436	
3	150	562	
4	200	875	
Total	550		

Determine the sample size to estimate the total milk production in the area. It is assumed that the C.V. of total milk production is 5%.

Solution : Let \hat{Y}_{st} be the total milk production in the area.

Then $s.e.(\hat{Y}_{st}) = 0.05 \times 20042 = 1002.1$.

$\therefore V(\hat{Y}_{st}) = V = [s.e.(\hat{Y}_{st})]^2 = (1002.1)^2 = 1004204.41$.

$$n_0 = \frac{N}{V} \sum N_h s_h^2 = \frac{550}{1004204.41} \times 514990 = 282.$$

Now, for proportional allocation,

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{282}{1 + \frac{282}{550}} = \frac{282}{1.5127} = 186.$$

13.10 Effect of Deviation from Optimum Allocation

According to Neyman allocation the sample size n_h is given by

$$n_h = \frac{n N_h S_h}{\sum N_h S_h}, \quad h = 1, 2, \dots, L$$

and the variance of \bar{y}_{st} is given by

$$V(\bar{y}_{st})_{opt(N)} = \frac{1}{n} \sum_h (W_h S_h)^2 - \frac{1}{N} \sum_h W_h S_h^2 = \frac{1}{n N^2} \left(\sum N_h S_h \right)^2 - \frac{1}{N^2} \sum N_h S_h^2.$$

In practice, the value of S_h is not known, whereas we use it to calculate n_h . However, the estimate s_h of S_h can be used. But due to this replacement the value of n_h is not exactly found out. Let the value of n_h be n'_h , if it is calculated using s_h . Therefore, we have

$$V(\bar{y}_{st}) = \sum \frac{W_h^2 S_h^2}{n'_h} - \frac{1}{N} \sum W_h S_h^2.$$

Now, the difference in the value of $V(\bar{y}_{st})$ due to the use of n'_h instead of n_h is

$$V(\bar{y}_{st}) - V(\bar{y}_{st})_{\text{opt}(N)} = \sum \frac{W_h^2 S_h^2}{n'_h} - \frac{1}{n} \left(\sum W_h S_h \right)^2.$$

Replacing $W_h S_h$ in right side in terms of n_h , we get

$$V(\bar{y}_{st}) - V(\bar{y}_{st})_{\text{opt}(N)} = \frac{(\sum W_h S_h)^2}{n^2} \left[\sum \frac{n_h^2}{n_h} - n \right] = \frac{(\sum W_h S_h)^2}{n^2} \sum \frac{(n'_h - n_h)^2}{n'_h}.$$

At this stage, if f.p.c. is neglected,

$$V(\bar{y}_{st})_{\text{opt}(N)} = \frac{1}{n^2} \left(\sum W_h S_h \right)^2.$$

Therefore, the relative increase in the variance due to the deviation of optimum allocation is given by

$$\frac{V(\bar{y}_{st}) - V(\bar{y}_{st})_{\text{opt}(N)}}{V(\bar{y}_{st})_{\text{opt}(N)}} = \frac{1}{n} \sum_h \frac{(n'_h - n_h)^2}{n'_h}.$$

Example 13.7 : Using the data of Example 13.1 investigate the impact of deviation of optimum allocation.

Solution : Given $N_1 = 121$, $N_2 = 149$, $N_3 = 75$, $N = 345$; $S_1^2 = 0.0235$, $S_2^2 = 0.0394$, $S_3^2 = 0.6501$. According to Neyman allocation $n_1 = 8$, $n_2 = 14$, $n_3 = 28$; $s_1^2 = 0.01071$, $s_2^2 = 0.2086$, $s_3^2 = 0.2903$. The variance of \bar{y}_{st} under Neyman allocation is $v(\bar{y}_{st}) = 0.0020045$. If n_h is calculated using s_h^2 instead of S_h^2 , we get

$$n'_1 = \frac{n N_1 s_1}{\sum N_h s_h} = \frac{50 \times 121 \times 0.1035}{120.9818} = 5, \quad n'_2 = \frac{n N_2 s_2}{\sum N_h s_h} = \frac{50 \times 149 \times 0.4567}{120.9818} = 28.$$

$$n'_3 = \frac{n N_3 s_3}{\sum N_h s_h} = \frac{50 \times 75 \times 0.5388}{120.9818} = 17.$$

Now, the $V(\bar{y}_{st})$ using n'_h is

$$\begin{aligned} V(\bar{y}_{st}) &= \sum_h \frac{W_h^2 S_h^2}{n'_h} - \frac{1}{N} \sum W_h S_h^2 = \frac{1}{N^2} \sum \frac{N_h^2 S_h^2}{n'_h} - \frac{1}{N^2} \sum N_h S_h^2 \\ &= \frac{1}{(345)^2} [68.8127 + 31.2399 + 215.1066 - 57.4716] \\ &= 0.0021649. \end{aligned}$$

Again, variance of \bar{y}_{st} under Neyman allocation is

$$\begin{aligned} V(\bar{y}_{st})_{\text{opt}(N)} &= \frac{1}{n N^2} \left(\sum N_h S_h \right)^2 - \frac{1}{N^2} \sum N_h S_h^2 = \frac{1}{(345)^2} \left[\frac{(\sum N_h S_h)^2}{n} - \sum N_h S_h^2 \right] \\ &= \frac{1}{(345)^2} \left[\frac{11793.13216}{50} - 57.4716 \right] = 0.0014988. \end{aligned}$$

Hence, the relative increase in the variance due to deviation of optimum allocation is

$$\frac{V(\bar{y}_{st}) - V(\bar{y}_{st})_{\text{opt}(N)}}{V(\bar{y}_{st})_{\text{opt}(N)}} = \frac{0.0021649 - 0.0014988}{0.0014988} = 44.44\%.$$

13.11 Stratified Sampling with Varying Probabilities

We have discussed stratified random sampling where sample units from h -th stratum are selected under simple random sampling scheme without replacement. In practice, the sample units may be selected with replacement and also these are selected with varying probabilities. Let us now discuss the sample selection with varying probabilities.

Let P_{hi} be the probability of selection of i -th unit from h -th stratum; $i = 1, 2, \dots, N_h$; $h = 1, 2, \dots, L$ so that

$$\sum_{i=1}^{N_h} P_{hi} = 1.$$

Let y_{hi} be the value of i -th unit of h -th stratum. Then, we can define

$$Z_{hi} = \frac{y_{hi}}{N_h P_{hi}}.$$

Then
$$\bar{Z}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi} = \frac{1}{n_h N_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{P_{hi}}.$$

This \bar{Z}_h is the unbiased estimator of population mean.

$$V(\bar{Z}_h) = \frac{\sigma_h^2}{n_h}, \quad \text{where } \sigma_h^2 = \sum_{i=1}^{N_h} P_{hi} (Z_{hi} - \bar{Z}_h)^2$$

Here
$$\bar{Z}_h = \sum_{i=1}^{N_h} P_{hi} Z_{hi} = \bar{Y}_h.$$

Again,
$$\bar{Z}_w = \sum_{h=1}^L \frac{N_h}{N} \bar{Z}_h = \sum_{h=1}^L W_h \bar{Z}_h = \bar{Y}.$$

The variance of \bar{Z}_h is

$$V(\bar{Z}_h) = \frac{1}{n_h} \sum_{i=1}^{N_h} P_{hi} \left(\frac{y_{hi}}{N_h P_{hi}} - \bar{Y}_h \right)^2 = \frac{1}{n_h} \left[\frac{1}{N_h^2} \sum_{i=1}^{N_h} \frac{y_{hi}^2}{P_{hi}} - \bar{Y}_h^2 \right].$$

The unbiased estimator of \bar{Y} is

$$\begin{aligned} \bar{Z}_w &= \frac{1}{N} \sum_{h=1}^L N_h \bar{Z}_h = \sum_{h=1}^L W_h \bar{Z}_h = \frac{1}{N} \sum_{h=1}^L \frac{N_h}{n_h N_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{P_{hi}} \\ &= \frac{1}{N} \sum_h \frac{1}{n_h} \sum_i \frac{y_{hi}}{P_{hi}}. \end{aligned}$$

Here
$$\begin{aligned} E(\bar{Z}_w) &= \sum_h W_h E(\bar{Z}_h) = \sum_h W_h \frac{1}{n_h N_h} E \sum_{i=1}^{n_h} \frac{y_{hi}}{P_{hi}} \\ &= \sum_h W_h \bar{Y}_h = \bar{Y}. \end{aligned}$$

The variance of \bar{Z}_w is

$$\begin{aligned} V(\bar{Z}_w) &= V \left[\sum_{h=1}^L W_h \bar{Z}_h \right] = \sum_{h=1}^L W_h^2 V(\bar{Z}_h) = \sum W_h^2 \frac{\sigma_h^2}{n_h} \\ &= \frac{1}{N^2} \sum_{h=1}^L \frac{N_h^2}{n_h} \sum_{i=1}^{N_h} P_{hi} \left(\frac{y_{hi}}{N_h P_{hi}} - \bar{Y}_h \right)^2 \\ &= \frac{1}{N^2} \sum \frac{1}{n_h} \left[\sum_{i=1}^{N_h} y_{hi}^2 / P_{hi} - N_h^2 \bar{Y}_h^2 \right]. \end{aligned}$$

The unbiased estimator of $V(\bar{Z}_w)$ is

$$v(\bar{Z}_w) = \sum_{h=1}^L W_h^2 \frac{s_h^2}{n_h}, \text{ where } s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Z_{hi} - \bar{Z}_h)^2.$$

13.12 Post-stratification

The stratified sampling is precise than simple random sampling if strata sizes and variances of the variable under study for different strata are known. In practice, the sizes of strata may not be known. Stratum variance may be replaced by its estimate from any related previous survey. But the estimated stratum size does not serve the purpose well. As an example of unknown stratum size, let us mention the case of sampling to estimate the fertility of ever married couples. The level of fertility varies with the variation of female education and occupation. But, in practice, it is difficult to know the number of ever married females of child bearing ages who have different levels of education or different levels of occupation. In such a case if strata is formed according to the level of education or occupation, it is difficult to get the value of N_h , even it is difficult to identify their strata unless a survey is conducted. The problem is solved by stratification after the completion of survey work. Hansen, Hurwitz and Madow (1953) have discussed the sampling technique of post-stratification. Williams (1962) has proposed the formula to get the variance of the estimator in such sampling.

Let there be N units in a population. Consider that n units are selected from the population by simple random sampling and these units are divided into L strata according to the values of the variable under study. Assume that the population size of h -th stratum is N_h and it is known. Then, we can find $\bar{y}_w = \sum W_h \bar{y}_h$ instead of $\bar{y} = \frac{1}{n} \sum \sum y_{hi}$, where $W_h = N_h/N$ and \bar{y}_h is the sample mean of h -th stratum. The variance of \bar{y}_w is given by

$$V(\bar{y}_w) = \sum_{h=1}^L \frac{W_h S_h^2}{n_h} - \frac{1}{N} \sum W_h S_h^2,$$

where n_h is the size of sample from h -th stratum. It is assumed that no n_h is zero. If in a repeated sampling n_h is found zero, the sampling units of two or more can be added.

Stephen (1945) has showed that if n_h is large enough, $V(\bar{y}_w)$ is not increased too much, where

$$E[v(\bar{y}_w)] = \frac{1-f}{n} \sum W_h S_h^2 + \frac{1}{n^2} (1 - W_h) S_h^2.$$

Chapter 14

Systematic Sampling

14.1 Introduction

In random sampling the sampling units are selected at random using random numbers. Each unit is selected at random. In some cases the first unit is selected at random and other units are selected in a systematic way, where the subsequent units are selected at a pre-specified distance from the first unit. Since the units, except the first unit, are selected in a systematic way, the technique of selecting sample units is called systematic sampling.

Let there be N units in a population, where the units are identified by serial number 1 to N . Let us consider that $N = nk$ and we need to select n units in the sample. Assume that k is an integer. Then one unit from first 1 to k units is selected randomly and let the serial number of this selected unit is 'i'. Now, the subsequent units for systematic sample will be at a distance of k units from the preceding selected units. Thus, if first selected unit is 'i', the subsequent selected units bearing serial numbers are $i + k, i + 2k, i + 3k, \dots, i + (n - 1)k$. Since the first unit is selected at random and other units are selected in a systematic way, the sampling technique is called *systematic sampling*. It is also called *Pseudo Random Sampling*, since first unit is selected randomly.

14.2 Method of Selecting Systematic Sample

There are mainly two different techniques which are adopted in selecting systematic sample. These are : (a) Linear Systematic Sampling (b) Circular Systematic Sampling.

Linear Systematic Sampling : Let there be $N = nk$ units in a population. Assume that the population units are linearly arranged and each unit is identified by a serial number from 1 to N . The problem is to select a random sample of n observations. Let us select a random number which will lie in the limit 1 to k . Let this number be i ($i \leq k$). Then, according to systematic sampling scheme the selected units in the sample bear the serial number $i, i + k, i + 2k, \dots, i + (n - 1)k$.

The Observations of k Systematic Samples

Sl.No.	1	2	...	i	...	k
1	y_1	y_2		y_i		y_k
2	y_{1+k}	y_{2+k}		y_{i+k}		y_{k+k}
3	y_{1+2k}	y_{2+2k}		y_{i+2k}		y_{k+2k}
⋮	⋮	⋮		⋮		⋮
	$y_{1+(j-1)k}$	$y_{2+(j-1)k}$		$y_{i+(j-1)k}$		$y_{k+(j-1)k}$
⋮	⋮	⋮		⋮		⋮
n	$y_{1+(n-1)k}$	$y_{2+(n-1)k}$		$y_{i+(n-1)k}$		$y_{k+(n-1)k}$
Mean	\bar{y}_1	\bar{y}_2		\bar{y}_i		\bar{y}_k

The process of selection by systematic sampling scheme provides k samples. Let y_j be the value of the variable under study of j th population unit ($j = 1, 2, \dots, N$). Then the observations of k samples will be as in the previous page.

The k samples as shown above is available if $N = nk$. In each sample there are n observations. But, if $N \neq nk$, at least one sample is not of size n . The sample size of one sample may be of size $(n - 1)$ or $(n + 1)$. Let us explain this by an example. Let there be $N = 24$ observations and we need a sample of size $n = 5$. Then the sample observations of different samples will be as follows :

	First Case		Second Case
Sl.No.	Sample observations	Sl.No.	
1	$y_1, y_6, y_{11}, y_{16}, y_{21}$	1	$y_1, y_6, y_{11}, y_{16}, y_{21}, y_{26}$
2	$y_2, y_7, y_{12}, y_{17}, y_{22}$	2	$y_2, y_7, y_{12}, y_{17}, y_{22}, \text{---}$
3	$y_3, y_8, y_{13}, y_{18}, y_{23}$	3	$y_3, y_8, y_{13}, y_{18}, y_{23}, \text{---}$
4	$y_4, y_9, y_{14}, y_{19}, y_{24}$	4	$y_4, y_9, y_{14}, y_{15}, y_{24}, \text{---}$
5	$y_5, y_{10}, y_{15}, y_{20}, \text{---}$	5	$y_5, y_{10}, y_{15}, y_{20}, y_{25}, \text{---}$

It is seen in the first case that the fifth sample is of size $n = 4$ (< 5). Again, if $N = 26$ (second case), the first sample has $n = 6$ (> 5) observations. However, this problem does not arise always even if $N \neq nk$.

If $N = nk$, k is considered an integer near to N/n and a random number is selected which exists between numbers 1 to k . The every k th unit is to be selected in the sample. Thus, if $N = 24$, k is to be considered as 5 ($\because 24/5 \approx 5$) and a random number is to be selected from the numbers 1 to 5. If the selected random number is 4 (4th sample of the first case), then the 4th, 9th, 14th, 19th and 24th units are included in the sample.

To select units by systematic sampling scheme, the population units are divided into n groups each of k units. A unit is selected first by a random process from first k units of the first group and then the other units are included in the sample at k th distance from the preceding selected unit. However, sometimes the first unit is not selected at random, rather first unit is selected from the centre of the list of k units. Thus, if k is even, $\frac{k}{2}$ th unit or $\frac{k+2}{2}$ th unit is selected as the first unit and if k is odd, $\frac{k+1}{2}$ th unit is selected as the first unit. Other units are selected at k th distance from the preceding selected unit. Thus, if $N = 24$ and if we need a sample of size $n = 5$, k is to be taken as 5 and the first selected unit is $\frac{5+1}{2}$ th or 3rd unit. The selected sample contains the observations $y_3, y_8, y_{13}, y_{18}, y_{23}$. The latter method of selection of systematic sample is proposed by Madow (1953).

Circular Systematic Sampling : It has already been mentioned that if $N \neq nk$, the sample size in selecting sampling units is changed [n is less or more than by 1]. To avoid this problem Lahiri (1952) has proposed a method in which first unit is selected randomly from all units in the population and the subsequent units are selected at a distance k from the preceding selected unit. For example, let us again consider the selection of $n = 5$ units in a sample from $N = 24$ units. Here $k \approx 5$. Let the first selected unit be 12. Then the sample observations are $y_{12}, y_{17}, y_{22}, y_3, y_8$.

In general, there are 5 samples. The sample observations are :

Sample No.	Sample Observations				
1	y_1	y_6	y_{11}	y_{16}	y_{21}
2	y_2	y_7	y_{12}	y_{17}	y_{22}
3	y_3	y_8	y_{13}	y_{18}	y_{23}
4	y_4	y_9	y_{14}	y_{19}	y_{24}
5	y_5	y_{10}	y_{15}	y_{20}	y_1

If the first selected unit is i th unit, then the other units in the sample will be of number

$$i + jk, \quad \text{if } i + jk \leq N \quad [j = 0, 1, \dots, (n - 1)]$$

$$i + jk - N, \quad \text{if } i + jk > N.$$

The probability of inclusion of any unit in such a sampling is $1/N$.

It is known that, if a single sample is selected in this sampling scheme, the estimate of sampling variance is not unbiased. To avoid this problem, Singh and Singh (1977) have proposed another method of sampling plan. Let there be N units in a population. We need to select a sample of n ($n < N$) units. Let u ($\leq n$) and d be two pre-determined integers. These two integers are to be selected in such a way that

- (i) In each sample different sampling units are included.
- (ii) The probability of inclusion of every pairs of unit will not be zero.

Now, let us select a random number from the numbers 1 to N . Let this selected number be r . Starting from this r th unit u units are to be selected one by one. Then $n - u = v$ units are to be selected at d distances so that $d \leq u$ and $u + vd \leq N$. If the sampling units are selected in the above mentioned way, the n units will be included in the sample in two steps or in many steps. Consider that to select n units in the sample, there needs p steps, where

$$p \geq \frac{\{\log \log \frac{N}{2} - \log \log \frac{n}{2}\}}{\log 2}.$$

14.3 Advantages and Disadvantages of Systematic Sampling

Advantages

- (i) The main advantage of this method is that the sampling units are selected easily. There is no chance of error to be creeped in selecting sampling units. This sampling technique is easily applied in the field of agricultural survey. Moreover, the technique is suitable if frame is not available. This is the main advantage of systematic sampling over simple random sampling.
- (ii) Since every k th unit is selected in the sample, there is no chance of exclusion of any special unit of population in the sample. If the units at a distance k are not correlated, the estimate of population characteristic becomes more precise. Since population units are divided into n strata of sizes k and one unit is selected from each stratum, the efficiency of systematic sampling is similar to that of stratified sampling when one unit is selected from each stratum.
- (iii) Systematic sampling is similar to cluster sampling. If there are k clusters and a cluster of n units are selected in the sample, then cluster sample is equivalent to systematic sample.

Disadvantages

However, there are some disadvantages of systematic sampling. The estimate under this scheme does not provide unbiased estimate if there is periodicity among the population units. Moreover, if a single sample is selected, it is difficult to get estimate of sampling variance.

14.4 Method of Estimation in Systematic Sampling

Let there be N units a population such that $N = nk$. Let us consider that a sample of n units is selected from the population. Let the value of j th observation of i th sample be y_{ij} ($i = 1, 2, \dots, k; j = 1, 2, \dots, n$) and the mean of i th sample be

$$\bar{y}_i = \frac{1}{n} \sum_j y_{ij}.$$

This mean is systematic sample mean and is denoted by

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

The population mean $\bar{Y} = \frac{1}{N} \sum \sum y_{ij} = \frac{1}{nk} \sum \sum y_{ij}$.

Theorem : From a population of size $N = nk$ if a systematic sample of size n is drawn, then systematic sample mean \bar{y}_{sy} is an unbiased estimator of population mean \bar{Y} and the variance of \bar{y}_{sy} is given by

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

where $S^2 = \frac{1}{N-1} \sum_i \sum_j (y_{ij} - \bar{Y})^2$ and $S_{wsy}^2 = \frac{1}{k(n-1)} \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$.

Proof : The systematic sample mean \bar{y}_{sy} is

$$\bar{y}_{sy} = \frac{1}{n} \sum_j y_{ij} = \bar{y}_i.$$

$$E(\bar{y}_{sy}) = E(\bar{y}_i) = \frac{1}{k} \sum_{i=1}^k \bar{y}_i = \frac{1}{nk} \sum \sum y_{ij} = \bar{Y}.$$

Hence, systematic sample mean is an unbiased estimator of population mean.

The variance of the systematic sample mean is

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum (\bar{y}_i - \bar{Y})^2.$$

We have $(N-1)S^2 = \sum \sum (y_{ij} - \bar{Y})^2 = n \sum (\bar{y}_i - \bar{Y})^2 + \sum \sum (y_{ij} - \bar{y}_i)^2$.

According to the technique of analysis of variance the d.f. of $\sum \sum (\bar{y}_i - \bar{Y})^2$ is $k(n-1)$. Since

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum (\bar{y}_i - \bar{Y})^2, \text{ we can write :}$$

$$(N-1)S^2 = nkV(\bar{y}_{sy}) + \sum \sum (y_{ij} - \bar{y}_i)^2.$$

$$\begin{aligned} \therefore V(\bar{y}_{sy}) &= \frac{(N-1)S^2}{nk} - \frac{1}{nk} \sum \sum (y_{ij} - \bar{y}_{i.})^2 = \frac{(N-1)S^2}{N} - \frac{1}{N} \sum \sum (y_{ij} - \bar{y}_{i.})^2 \\ &= \frac{(N-1)S^2}{N} - \frac{k(n-1)}{N} S_{wsy}^2. \end{aligned}$$

Corollary : $V(\bar{y}_{sy}) = \frac{k-1}{k} S_e^2$, where $S_e^2 = \frac{1}{k-1} \sum (\bar{y}_{i.} - \bar{Y})^2$. Here S_e^2 is the variance of sample mean and this is obtained from i -th sample mean.

The above formula of $V(\bar{y}_{sy})$ does not provide any information that the variance of systematic sample mean decreases with the increase in sample size. Therefore, systematic sampling technique is to be applied with care.

$$\text{Again, } V(\bar{y}_{sy}) = \frac{(N-1)S^2}{N} - \frac{k(n-1)}{N} S_{wsy}^2.$$

It is observed that $V(\bar{y}_{sy})$ will be minimum if S_{wsy}^2 becomes more. This S_{wsy}^2 becomes more if the sampling units are more heterogeneous. This is possible if the units within a group are more homogeneous and units of different groups are heterogeneous.

Corollary : The variance of the systematic sample mean is also written as

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1)\rho_w],$$

where
$$\rho_w = \frac{E(y_{ij} - \bar{Y})(y_{iu} - \bar{Y})}{E(y_{ij} - \bar{Y})^2}, \quad j \neq u$$

$$= \frac{2}{(n-1)(N-1)S^2} \sum_i^k \sum_{j < u}^n (y_{ij} - \bar{Y})(y_{iu} - \bar{Y}).$$

Also
$$V(\bar{y}_{sy}) = \frac{S_{wst}^2}{n} \left(\frac{N-n}{N} \right) [1 + (n-1)\rho_{wst}],$$

where
$$S_{wst}^2 = \frac{1}{n(k-1)} \sum_i^k \sum_j^n (y_{ij} - \bar{y}_{.j})^2$$

and
$$\rho_{wst} = \frac{E(y_{ij} - \bar{y}_{.j})(y_{iu} - \bar{y}_{.u})}{E(y_{ij} - \bar{y}_{.j})^2}$$

$$= \frac{2}{n(n-1)(k-1)} \sum_i^k \sum_{j < u}^n \frac{(y_{ij} - \bar{y}_{.j})(y_{iu} - \bar{y}_{.u})}{S_{wst}^2}.$$

Here S_{wst}^2 is the variance of the observations within the groups (strata).

We have
$$V(\bar{y}_{sy}) = \frac{1}{k} \sum (\bar{y}_{i.} - \bar{Y})^2$$

$$\begin{aligned} n^2 k V(\bar{y}_{sy}) &= n^2 \sum (\bar{y}_{i.} - \bar{Y})^2 = \sum_i^k [(y_{i1} - \bar{Y}) + (y_{i2} - \bar{Y}) + \dots + (y_{in} - \bar{Y})]^2 \\ &= \sum_i^k \sum_j^n (y_{ij} - \bar{Y})^2 + 2 \sum_j \sum_{j < u} (y_{ij} - \bar{Y})(y_{iu} - \bar{Y}) \\ &= (N-1)S^2 + (n-1)(N-1)S^2 \rho_w. \end{aligned}$$

$$V(\bar{y}_{sy}) = \frac{(N-1)S^2}{n^2 k} [1 + (n-1)\rho_w] = \frac{N-1}{nN} S^2 [1 + (n-1)\rho_w].$$

Here ρ_w is the intra-class correlation coefficient of observations within a sample. Thus, it may be concluded that, if the intra-class correlation coefficient of observations within a systematic sample increases, the $V(\bar{y}_{sy})$ is also increased. This means that the efficiency of systematic sampling will be increased if the classes are so formed that the correlation of observations between classes are minimum.

Let us now verify the formula $V(\bar{y}_{sy}) = \frac{S_{wst}^2}{n} \left(\frac{N-n}{N} \right) [1 + (n-1)\rho_{wst}]$.

Consider that \bar{y}_j is the mean of j -th class ($j = 1, 2, \dots, n$). The variance of the observations of j -th class included in different systematic sample is

$$S_j^2 = \frac{1}{k-1} \sum (y_{ij} - \bar{y}_j)^2.$$

It has already been mentioned that systematic sampling may be considered as stratified sampling, where one unit is selected from each stratum of size k . Then

$$V(\bar{y}_{st}) = \frac{k-1}{nN} \sum_{j=1}^n S_j^2 = \frac{k-1}{N} S_{wst}^2.$$

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

$$\begin{aligned} \text{Again, } V(\bar{y}_{sy}) &= \frac{1}{k} \sum (\bar{y}_i - \bar{Y})^2 = \frac{1}{k} \sum_i \left[\frac{1}{n} \sum_j y_{ij} - \frac{1}{n} \sum_{j=1}^n \bar{y}_j \right]^2 \\ &= \frac{1}{n^2 k} \sum_i \left[\sum_j y_{ij} - \bar{y}_j \right]^2 \\ &= \frac{1}{n^2 k} \left[\sum_i \sum_j (y_{ij} - \bar{y}_j)^2 + \sum_i \sum_{j < u} (y_{ij} - \bar{y}_j)(y_{iu} - \bar{y}_u) \right] \\ &= \frac{1}{n^2 k} [n(k-1)S_{wst}^2 + n(n-1)(k-1)\rho_{wst}S_{wst}^2] \\ &= \frac{(k-1)S_{wst}^2}{N} [1 + (n-1)\rho_{wst}] = \frac{N-n}{Nn} S_{wst}^2 [1 + (n-1)\rho_{wst}]. \end{aligned}$$

Here ρ_{wst} is the non-circular serial correlation coefficient.

Thus, the efficiency of systematic sampling depends on ρ_{wst} . If $\rho_{wst} = 0$, systematic sampling and stratified sampling are equally efficient. The efficiency of systematic sampling will be decreasing, if ρ_{wst} increases gradually.

Estimation of sampling variance of systematic sample mean : The systematic sample may be considered as simple random sample of size $n = 1$. The sample of one unit does not provide unbiased estimate of variance. However, the biased estimator if $V(\bar{y}_{sy})$ may be found out.

Let us consider that k independent systematic samples are selected and each sample is of size n . Then the estimator of variance of systematic sample mean is

$$v(\bar{y}_{sy}) = \frac{N-n}{Nn} s_{we}^2, \text{ where } s_{we}^2 = \frac{1}{n-1} \left[\sum_j y_{ij}^2 - n\bar{y}_i^2 \right]$$

Comparison of systematic sampling, simple random sampling and stratified sampling : Let there be N units in a population and a simple random sample of n units be selected from this population. Then

$$V(\bar{y}) = \frac{N-n}{Nn} S^2, \quad \text{where } \bar{y} = \frac{1}{n} \sum y_i, \quad S^2 = \frac{1}{N-1} \sum (y_i - \bar{Y})^2.$$

If a systematic sample of size n is drawn from this population, then

$$V(\bar{y}_{sy}) = \frac{N-1}{Nn} S^2 [1 + (n-1)\rho_w],$$

where ρ_w is the intra-class correlation coefficient. It is already mentioned that $V(\bar{y}_{sy}) > V(\bar{y})$, if ρ increases. Again,

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

If systematic sampling is more efficient than simple random sampling, then

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

or, $k(n-1)S_{wsy}^2 > \left[(N-1) - \frac{N-n}{h} \right] S^2$

or, $k(n-1)S_{wsy}^2 > k(n-1)S^2$

or, $S_{wsy}^2 > S^2.$

Thus systematic sampling is more efficient than simple random sampling, if $S_{wsy}^2 > S^2$.

Since $V(\bar{y}_{sy}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1)\rho_w]$, ρ_w cannot be less than $-1/(n-1)$. Because $V(\bar{y}_{sy})$ cannot be negative. Therefore, the lowest value of ρ_w is $-1/(n-1)$. Now, $V(\bar{y}_{sy})$ as given above and $V(\bar{y})$ can be compared. Hence, the relative efficiency of systematic sampling compared to simple random sampling is

$$\frac{V(\bar{y})}{V(\bar{y}_{sy})} = \frac{N-n}{(N-1)} [1 + \rho_w(n-1)]^{-1}.$$

Thus, both sampling scheme will be of equal efficiency, if $\rho_w = -1/(N-1)$. If $\rho_w > -1/(N-1)$, simple random sampling is more efficient than systematic sampling. But, if $\rho_w < -1/(N-1)$, systematic sampling will be more efficient.

It has already been mentioned that systematic sampling is equivalent to stratified sampling, where one unit is selected from each stratum. In such a case,

$$V(\bar{y}_{st}) = \frac{k-1}{nk} S_{wst}^2.$$

Again, $V(\bar{y}_{sy}) = \frac{k-1}{nk} S_{wst}^2 [1 + (n-1)\rho_{wst}].$

Comparing above two variances, it can be mentioned that if $\rho_{wst} = 0$, both sampling schemes are equally efficient. The relative efficiency of systematic sampling compared to stratified sampling is

$$\frac{V(\bar{y}_{st})}{V(\bar{y}_{sy})} = [1 + (n-1)\rho_{wst}]^{-1}.$$

If $\rho_{wst} > 0$, the systematic sampling becomes less efficient compared to stratified sampling.

Example 14.1 : A farmer sells milk everyday. The amounts of milk sold (kg) per day are shown below :

Day	Milk sold	Day	Milk sold	Day	Milk sold	Day	Milk sold	Day	Milk sold	Day	Milk sold
1	30.5	41	28.0	81	29.0	121	20.0	161	30.0	201	25.0
2	24.6	42	28.5	82	29.0	122	25.6	162	27.0	202	25.5
3	32.4	43	28.5	83	29.0	123	30.0	163	26.0	203	24.0
4	22.6	44	28.5	84	30.0	124	30.0	164	25.0	204	28.0
5	20.5	45	30.0	85	30.0	125	31.2	165	25.0	205	28.5
6	28.9	46	30.0	86	30.0	126	32.3	166	22.0	206	29.0
7	30.4	47	30.0	87	29.0	127	34.0	167	26.0	207	29.0
8	30.2	48	30.0	88	27.0	128	30.0	168	28.2	208	29.0
9	30.6	49	30.0	89	27.5	129	32.0	169	20.5	209	30.0
10	26.6	50	30.0	90	28.0	130	32.0	170	20.0	210	29.5
11	28.7	51	25.5	91	28.0	131	32.0	171	21.2	211	25.0
12	25.0	52	26.5	92	28.0	132	30.0	172	28.0	212	25.0
13	26.0	53	26.0	93	28.0	133	30.0	173	28.0	213	25.0
14	25.0	54	26.0	94	29.0	134	30.0	174	30.0	214	26.0
15	25.0	55	26.0	95	29.0	135	30.0	175	30.0	215	22.0
16	25.0	56	25.5	96	29.0	136	30.5	176	30.2	216	21.2
17	25.0	57	25.0	97	30.0	137	30.5	177	30.5	217	22.5
18	26.0	58	26.0	98	30.0	138	31.0	178	33.0	218	23.0
19	25.0	59	27.0	99	30.0	139	30.0	179	29.0	219	28.0
20	30.0	60	28.0	100	30.0	140	30.0	180	30.0	220	22.5
21	28.0	61	28.0	101	30.0	141	28.0	181	30.5	221	24.6
22	28.0	62	28.0	102	30.0	142	27.0	182	31.2	222	25.0
23	28.0	63	28.5	103	25.0	143	27.5	183	34.0	223	30.0
24	28.0	64	28.5	104	25.0	144	27.5	184	35.0	224	30.0
25	29.0	65	28.5	105	25.0	145	27.5	185	34.0	225	30.0
26	31.0	66	28.5	106	25.0	146	28.0	186	30.2	226	30.0
27	32.0	67	32.0	107	26.0	147	29.0	187	30.5	227	28.2
28	30.0	68	32.0	108	26.0	148	29.0	188	30.5	228	28.0
29	28.0	69	32.0	109	26.0	149	29.5	189	30.0	229	29.0
30	29.0	70	31.0	110	26.0	150	29.0	190	30.0	230	29.0
31	27.5	71	30.0	111	26.0	151	28.0	191	30.0	231	29.0
32	27.5	72	30.0	112	27.0	152	28.0	192	30.0	232	32.5
33	27.5	73	30.0	113	27.0	153	28.0	193	26.0	233	32.0
34	25.0	74	28.5	114	27.0	154	28.0	194	25.5	234	32.0
35	25.5	75	29.0	115	27.0	155	28.0	195	24.0	235	30.0
36	27.5	76	31.5	116	27.0	156	28.5	196	24.0	236	29.0
37	27.8	77	31.5	117	24.0	157	30.2	197	24.0	237	29.0
38	28.0	78	31.0	118	24.5	158	30.0	198	20.0	238	29.0
39	28.0	79	28.2	119	25.5	159	30.5	199	22.2	239	30.5
40	28.0	80	28.5	120	20.0	160	30.5	200	25.0	240	31.5

Day	Milk sold	Day	Milk sold	Day	Milk sold	Day	Milk sold	Day	Milk sold	Day	Milk sold
241	31.5	262	32.0	283	28.0	304	27.0	325	27.5	346	30.2
242	31.0	263	30.0	284	28.0	305	30.0	326	27.8	347	30.4
243	30.0	264	30.0	285	28.0	306	28.0	327	29.0	348	30.4
244	30.0	265	28.0	286	34.0	307	28.5	328	29.0	349	24.0
245	28.0	266	28.5	287	33.0	308	28.5	329	30.0	350	26.2
246	28.0	267	28.0	288	32.0	309	29.0	330	24.5	351	25.0
247	24.0	268	28.0	289	32.0	310	29.0	331	25.5	352	25.0
248	24.0	269	28.0	290	30.0	311	30.0	332	24.5	353	32.0
249	26.0	270	29.5	291	26.0	312	30.0	333	23.0	354	32.5
250	26.5	271	29.5	292	26.6	313	30.0	334	27.0	355	30.2
251	26.5	272	27.0	293	26.5	314	31.5	335	24.0	356	30.0
252	27.0	273	27.0	294	27.0	315	32.5	336	24.0	357	31.6
253	27.0	274	26.0	295	27.0	316	33.0	337	24.0	358	25.5
254	27.0	275	26.2	296	28.2	317	33.0	338	24.5	359	25.8
255	28.0	276	26.4	297	25.0	318	28.0	339	25.5	360	25.0
256	29.5	277	25.0	298	25.0	319	29.0	340	25.0	361	24.0
257	30.0	278	25.0	299	24.0	320	29.0	341	26.0	362	24.5
258	30.0	279	24.0	300	22.0	321	30.0	342	26.2	363	20.2
259	30.0	280	22.0	301	20.0	322	30.0	343	28.2	364	23.0
260	29.5	281	20.0	302	22.0	323	26.0	344	28.6	365	23.5
261	30.0	282	20.0	303	27.0	324	26.5	345	29.2		

- (i) Draw a systematic sample of size $n = 30$.
- (ii) Estimate the total amount of milk sold throughout the year.
- (iii) Estimate the variance of the estimated total milk production.
- (iv) Estimate the grain in precision of systematic sampling compared to simple random sampling.

Solution : (i) We have $N = 365, n = 30, k \approx 12 (k \approx \frac{N}{n})$. Since k is not an integer, we need to select sample observations by circular systematic sampling. The first unit is selected using a "Random Number" selected from the numbers 1 to 365. This number is (using random number table shown in appendix) 161. Therefore, the selected days are :

161, 173, 185, 197, 209, 221, 233, 245, 257, 269, 281, 293, 305, 317, 329, 341, 353, 365, 012, 024, 036, 048, 060, 072, 084, 096, 108, 120, 132, 144.

The Selected Sample Observations

SL. No.	Milk sold	SL. No.	Milk sold	SL. No.	Milk sold	SL. No.	Milk sold	SL. No.	Milk sold
1	30.0	7	32.0	13	30.0	19	25.0	25	30.0
2	28.0	8	28.0	14	33.0	20	28.0	26	29.0
3	34.0	9	30.0	15	30.0	21	27.5	27	26.0
4	24.0	10	28.0	16	26.0	22	30.0	28	20.0
5	30.0	11	20.0	17	32.0	23	28.0	29	30.0
6	24.6	12	26.5	18	23.5	24	30.0	30	27.5

(ii) The estimate of total milk sold throughout the year is given by

$$\begin{aligned}\hat{Y}_{sy} &= N\bar{y}_{sy}, \quad \text{where } \bar{y}_{sy} = \frac{1}{n} \sum_{j=1}^n y_{ij}, \quad i = 1 \\ &= \frac{840.6}{30} = 28.02 \text{ kg} \\ &= 365 \times 28.02 = 10227.3 \text{ kg.}\end{aligned}$$

(iii) The estimate of variance of the estimated total milk production is

$$v(\hat{Y}_{sy}) = \frac{N(N-n)}{n} s_{we}^2,$$

$$\text{where } s_{we}^2 = \frac{1}{n-1} \left[\sum y_{ij}^2 - n\bar{y}_i^2 \right],$$

$$\bar{y}_i = \frac{1}{n} \sum_{j=1}^n y_{ij} = 28.02$$

$$= \frac{1}{30-1} [23879.16 - 30(28.02)^2] = 11.2258.$$

$$\therefore v(\hat{Y}_{sy}) = \frac{365(365-30)}{30} \times 11.2258 = 45754.46.$$

(iv) The gain in precision of systematic sampling over simple random sampling is

$$\frac{v(\hat{Y}) - v(\hat{Y}_{sy})}{v(\hat{Y}_{sy})}.$$

$$\text{Here } v(\hat{Y}) = \frac{N(N-n)}{n} s^2,$$

$$\text{where } s^2 = \frac{1}{n-1} \left[\sum_{j=1}^n y_{ij}^2 - \frac{(\sum_{j=1}^n y_{ij})^2}{n} \right], \quad i = 1.$$

$$\therefore s^2 = \frac{1}{30-1} \left[23879.16 - \frac{(840.6)^2}{30} \right] = 11.2258.$$

$$v(\hat{Y}) = \frac{365(365-30)}{30} \times 11.2258 = 45754.46.$$

It is observed that $v(\hat{Y}) = v(\hat{Y}_{sy})$. However, $v(\hat{Y}_{sy})$ is not an unbiased estimator of $V(\hat{Y}_{sy})$ and hence, this $v(\hat{Y}_{sy})$ should not be used to compare with $v(\hat{Y})$.

Another biased estimator of $V(\hat{Y}_{sy})$ is

$$\begin{aligned}v(\hat{Y}_{sy}) &= \frac{N(N-n)}{2n(n-1)} \left[\sum_{j=1}^{n-1} (y_{ij} - y_{ij+1})^2 \right], \quad i = 1 \\ &= \frac{365(365-30)}{2 \times 30(30-1)} (710.67) = 49940.90.\end{aligned}$$

Hence, there is no net gain in precision of systematic sampling compared to simple random sampling, since $v(\hat{Y}_{sy}) > v(\hat{Y})$. This comparison is also not perfect as $v(\hat{Y}_{sy})$ is not unbiased.

Example 14.2 : The following data represent the amount of fish (in kg) sold in a market in different days in the month of April 2015.

Day	Amount of fish	Day	Amount of fish	Day	Amount of fish	Day	Amount of fish	Day	Amount of fish
1	800.5	7	600.0	13	700.6	19	660.2	25	895.0
2	970.2	8	715.8	14	785.7	20	775.0	26	1008.0
3	612.5	9	865.2	15	975.0	21	842.0	27	975.0
4	995.7	10	1000.0	16	670.0	22	632.8	28	1062.5
5	870.5	11	1014.5	17	500.2	23	780.6	29	770.6
6	900.0	12	880.7	18	842.8	24	990.0	30	715.5

- (i) Draw all possible systematic samples of days after each 5 days to estimate the average fish sold in the market.
- (ii) Estimate the variance of the estimate of average fish sold per day.
- (iii) Compare systematic sample with simple random sample and stratified sample.

Solution : (i) Here $N = 30$, $k = 5$, $n = 6$. Since $k = 5$, we have to select a random number from numbers 1 to 5. Using the 'Random Number Table' given in Appendix, the selected random number is 5. Hence, all possible systematic samples are as follows :

Sample numbers	Sampling units						Total y_{i0}	Mean \bar{y}_{i0}
1	870.5	1000.0	975.0	775.0	895.0	715.5	5231.0	871.83
2	800.5	900.0	1014.5	670.0	842.0	1008.0	5235.0	872.50
3	970.2	600.0	880.7	500.2	632.8	975.0	4558.9	759.82
4	612.5	715.8	700.6	842.8	780.6	1062.5	4714.8	785.80
5	995.7	865.2	785.7	660.2	990.0	770.6	5067.4	844.57
Total y_{0j}	4249.4	4081.0	4356.5	3448.2	4140.4	4531.6	24807.1	

The estimate of average fish sold per day is

$$\bar{y}_{sy} = \frac{1}{n} \sum y_{ij} = \frac{5231.0}{6} = 871.83.$$

This average is based on the first sample. The estimates on the basis of other samples are shown in the table ($\bar{y}_{i.}$, $i = 1, 2, 3, 4, 5$).

(ii) The estimate of variance of the estimate of average fish is $v(\bar{y}_{sy}) = \frac{N-n}{Nn} s_{wc}^2$.

Here $s_{wc}^2 = \frac{1}{n-1} \left[\sum_{j=1}^n y_{ij}^2 - n\bar{y}_{i.}^2 \right] \quad i = 1,$

since estimate of average fish sold is obtained from the first sample.

$$s_{wc}^2 = \frac{1}{6-1} [4621985.5 - 6(871.83)^2] = 12292.04.$$

$$v(\bar{y}_{sy}) = \frac{30-6}{30 \times 6} 12292.04 = 1638.94.$$

(iii) The variance of systematic sample mean is

$$V(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k}{N} (n-1) S_{w_{sy}}^2,$$

$$\text{where } S^2 = \frac{1}{N-1} \left[\sum \sum y_{ij}^2 - \frac{(\sum \sum y_{ij})^2}{N} \right] = \frac{1}{30-1} \left[21148128.53 - \frac{(24807.1)^2}{30} \right]$$

$$= 21898.44.$$

$$S_{w_{sy}}^2 = \frac{1}{k(n-1)} \sum \sum (y_{ij} - \bar{y}_i)^2$$

$$= \frac{1}{k(n-1)} \left[\left\{ \sum y_{1j}^2 - \frac{(\sum y_{1j})^2}{n} \right\} + \left\{ \sum y_{2j}^2 - \frac{(\sum y_{2j})^2}{n} \right\} \right.$$

$$\left. + \left\{ \sum y_{3j}^2 - \frac{(\sum y_{3j})^2}{n} \right\} + \left\{ \sum y_{4j}^2 - \frac{(\sum y_{4j})^2}{n} \right\} + \left\{ \sum y_{5j}^2 - \frac{(\sum y_{5j})^2}{n} \right\} \right]$$

$$= \frac{1}{5(6-1)} \left[\left\{ 4621985.5 - \frac{(5231.0)^2}{6} \right\} + \left\{ 4653938.5 - \frac{(5235.0)^2}{6} \right\} + \left\{ 3678181.41 \right. \right.$$

$$\left. - \frac{(4558.9)^2}{6} \right\} + \left\{ 3826920.7 - \frac{(4714.8)^2}{6} \right\} + \left\{ 4367102.42 - \frac{(5067.4)^2}{6} \right\} \right]$$

$$= \frac{1}{25} [571455.6949] = 22858.2278.$$

$$V(\bar{y}_{sy}) = \frac{30-1}{30} \times 21898.44 - \frac{5(6-1)}{30} \times 22858.2278 = 2119.969.$$

$$V(\bar{y}) = \frac{N-n}{Nn} S^2 = \frac{30-6}{30 \times 6} \times 21898.44 = 2919.792.$$

The gain in precision of systematic sampling compared to simple random sampling is

$$\frac{V(\bar{y}) - V(\bar{y}_{sy})}{V(\bar{y}_{sy})} \times 100\% = \frac{2919.792 - 2119.969}{2119.969} \times 100\% = 37.73\%$$

$$\text{Again, } V(\bar{y}_{st}) = \frac{k-1}{nN} \sum_{j=1}^n S_j^2, \text{ where } S_j^2 = \left[\sum_{i=1}^k y_{ij}^2 - \frac{(y_{.j})^2}{k} \right] / k - 1 = \frac{k-1}{N} S_{w_{st}}^2.$$

SL. No.	$\sum_i y_{ij}^2$	$y_{.j}^2/k$	$S_j^2 = \frac{1}{k-1} \left[\sum_i y_{ij}^2 - \frac{(y_{.j})^2}{k} \right]$
1	3706433.28	3611480.072	23738.302
2	3430940.68	3330912.2	25007.12
3	3863632.59	3795818.45	16953.535
4	2445900.92	2378016.648	16971.068
5	3499861.20	3428582.432	17819.692
6	4201359.86	4107079.712	23570.037
Total			124059.754

$$V(\bar{y}_{st}) = \frac{5-1}{6 \times 30} 124059.754 = 2756.883.$$

Therefore, the gain in precision of systematic sampling compared to stratified sampling is

$$\frac{V(\bar{y}_{st}) - V(\bar{y}_{sy})}{V(\bar{y}_{sy})} \times 100\% = \frac{2756.883 - 2119.969}{2119.969} \times 100\% = 30.04\%.$$

14.5 Systematic Sampling when Population Units are in Random Order

If the frame is formed with random serial number, then there is no possibility of any linear trend in the population observations; the population observations which are located nearby are not correlated and no strata can be formed with those nearby units. For example, let us consider the survey to estimate the monthly expenditure of some students of a college where students are listed according to the first letter of their name and students are selected in a systematic way after every 5 students. In such a case the variable monthly expenditure of students will not be correlated with the serial number of the frame or there will not be any trend with the increase in serial number and expenditure. There will be less chance of correlation in amount of expenditure of a group of 5 or 10 students who are nearby in the frame. The students cannot be grouped into strata. The systematic sampling in such a situation is equivalent to simple random sampling. However, $V(\bar{y}_{sy})$ will not always be equal to $V(\bar{y})$, since $V(\bar{y}_{sy})$ depends on the value of k .

Let there be N units in a finite population. These N units can be arranged in $N!$ ways. Now, if any one arrangement is selected at random and if systematic sample is selected from that randomly selected arrangement of population units, then

$$E V(\bar{y}_{sy}) = V(\bar{y}).$$

However, if population units are not selected from $N!$ arrangements, $EV(\bar{y}_{sy}) \neq V(\bar{y})$. The $V(\bar{y})$ is fixed for every arrangements of $N!$.

Theorem : Let the observations of a super population be y_1, y_2, \dots, y_N such that

$$E(y_i) = \mu, \quad E(y_i - \mu)(y_j - \mu) = 0, \quad \text{if } i \neq j \\ = \sigma_i^2, \quad \text{if } i = j.$$

Then, if systematic sample of size n is drawn from such a population

$$E V(\bar{y}_{sy}) = EV(\bar{y}).$$

Proof: We know that, if a simple random sample of size n is drawn from a finite population of size N , then

$$V(\bar{y}) = \frac{N-n}{Nn} S^2 = \frac{N-n}{Nn} \cdot \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2.$$

But
$$\sum_{i=1}^N (y_i - \bar{Y})^2 = \sum_{i=1}^N [(y_i - \mu) - (\bar{Y} - \mu)]^2 = \sum_{i=1}^N (y_i - \mu)^2 - N(\bar{Y} - \mu)^2.$$

Again,
$$E(\bar{Y} - \mu)^2 = \frac{1}{N^2} \sum_{i=1}^N \sigma_i^2 \quad [\because y_i \text{ and } y_j \text{ (} i \neq j \text{) are not correlated}]$$

Then
$$E V(\bar{y}) = \frac{N-n}{Nn} \frac{1}{N-1} \left[\sum_{i=1}^N E(y_i - \mu)^2 - NE(\bar{Y} - \mu)^2 \right]$$

$$\begin{aligned}
 &= \frac{N-n}{Nn} \frac{1}{N-1} \left[\sum \sigma_i^2 - \frac{N}{N^2} \sum \sigma_i^2 \right] \\
 &= \frac{N-n}{nN^2} \sum \sigma_i^2.
 \end{aligned}$$

Let \bar{y}_u is the sample mean of u -th systematic sample. Then

$$\begin{aligned}
 V(\bar{y}_{sy}) &= \frac{1}{k} \sum_u^k (\bar{y}_u - \bar{Y})^2 = \frac{1}{k} \sum_{u=1}^k [(\bar{y}_u - \mu) - (\bar{Y} - \mu)]^2 \\
 &= \frac{1}{k} \left[\sum (\bar{y}_u - \mu)^2 - k(\bar{Y} - \mu)^2 \right].
 \end{aligned}$$

Here \bar{y}_u is the sample mean of uncorrelated sample. Therefore, we can write :

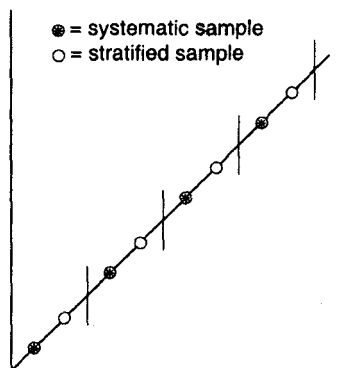
$$\begin{aligned}
 E[V(\bar{y}_{sy})] &= \frac{1}{k} \left[\frac{\sum \sigma_i^2}{n^2} - \frac{k \sum \sigma_i^2}{N^2} \right] \\
 &= \frac{N-n}{nN^2} \sum \sigma_i^2 = E[V(\bar{y})].
 \end{aligned}$$

14.6 . Systematic Sampling in Populations with Linear Trend

Let there be N units in a population and there is linear trend in the observations of population units. The observations can be represented by a model :

$$y_i = \alpha + \beta i, \quad i = 1, 2, \dots, N.$$

It is observed that the value of y_i is increased according to the model given above, where α and β are constants. The values of y_i for different values of i can be shown graphically as follows :



From the graph it is clear that if in a stratum systematic sample observation is smaller, it is smaller in every stratum. But it is not true, if there is one unit in a stratum and in that case, $V(\bar{y}_{sy}) > V(\bar{y}_{st})$. This is true since there is a chance of less variation within the observations of a stratum.

For linear trend in population observations,

$$\bar{Y} = \frac{1}{N} \sum (\alpha + \beta i) = \alpha + \beta \frac{(N+1)}{2}.$$

$$\begin{aligned} \text{Also } S^2 &= \frac{1}{N-1} \sum (y_i - \bar{Y})^2 = \frac{1}{N-1} \sum \left[\alpha + \beta i - \alpha - \beta \frac{(N+1)}{2} \right]^2 \\ &= \frac{\beta^2}{N-1} \sum \left(i - \frac{N+1}{2} \right)^2 = \frac{N(N+1)\beta^2}{12} = \frac{nk(nk+1)\beta^2}{12}. \end{aligned}$$

$$\text{Then } V(\bar{y}) = \frac{N-n}{Nn} S^2 = \frac{N-n}{Nn} \frac{nk(nk+1)\beta^2}{12} = \frac{(k-1)(nk+1)\beta^2}{12}.$$

In a similar way, we have

$$V(\bar{y}_{st}) = \frac{N-n}{Nn} S_w^2 = \frac{k-1}{nk} \frac{k(k+1)}{12} \beta^2 = \frac{(k-1)(k+1)\beta^2}{12n} = \frac{(k^2-1)\beta^2}{12n}.$$

Again, in case of systematic sampling,

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{i=1}^k (\bar{y}_i - \bar{Y})^2 = \frac{1}{k} \frac{k(k+1)(k-1)\beta^2}{12} = \frac{(k^2-1)\beta^2}{12}.$$

$$\text{Therefore, } V(\bar{y}_{st}) = \frac{k^2-1}{12n} \leq V(\bar{y}_{sy}) = \frac{k^2-1}{12} \leq \frac{(k-1)(N+1)}{12}.$$

Now, if $n=1$, $V(\bar{y}_{st}) = V(\bar{y}_{sy})$.

It is observed that the variance of stratified sample is $\frac{1}{n}$ th portion of variance of systematic sample. This is true for simple random sample also. We have

$$\begin{aligned} V(\bar{y}_{st}) : V(\bar{y}_{sy}) : V(\bar{y}) &:: \frac{k+1}{n} : k+1 : nk+1 \\ &\approx \frac{1}{n} : 1 : n = 1 : n : n^2 \end{aligned}$$

However, the efficiency of systematic sample can be increased using Yates (1948) correction or using modified method suggested by Singh et al. (1968).

Chapter 15

Ratio and Regression Estimator

15.1 Ratio Estimator

The objective of the sample survey is to estimate the parameter of the population under consideration. The estimator so far discussed is simple average of sample observations. However, the estimator can be defined using some auxiliary information which are related to the sample observations, if these are available from each sampling unit. Such types of estimators where the values of the related variable are used to estimate the parameter of the main variable are (a) Ratio Estimator, and (b) Regression Estimator. In this chapter we shall discuss ratio estimator. The ratio estimator can even be defined if total value of the auxiliary variable instead of all observations is known.

Let there be N units in a population. Assume that n units are selected under simple random sampling scheme without replacement (SRSWOR). Let y_i be the value of the variable under study recorded from i -th unit in the population ($i = 1, 2, \dots, N$) and x_i be the value of a related variable which is related to y_i . Now,

$$Y = \sum_{i=1}^N y_i = \text{total of variable } y, \quad \bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i.$$

$$X = \sum_{i=1}^N x_i = \text{total of variable } x, \quad \bar{X} = \frac{1}{N} \sum_{i=1}^N x_i.$$

$$r_i = \frac{y_i}{x_i} = \text{ratio of the value of } y \text{ and } x \text{ variable for } i\text{-th unit.}$$

$$\bar{r} = \frac{1}{N} \sum_{i=1}^N r_i = \frac{1}{N} \sum_{i=1}^N \frac{y_i}{x_i}, \quad R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}.$$

$$\hat{r} = \frac{1}{n} \sum_{i=1}^n r_i, \quad \hat{R} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}},$$

$$\text{where } y = \sum_{i=1}^n y_i, \quad x = \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Also, consider that the population correlation coefficient of two variables is ρ . Here $\hat{R}, \hat{\bar{Y}}_R$ and \hat{Y}_R are respectively the estimator of R, \bar{Y} and Y .

Theorem : In simple random sampling if sample of size n is selected, then the sample observations provides ratio estimator of population mean, population total and population ratio as $\hat{\bar{Y}}_R, \hat{Y}_R$ and \hat{R} respectively. If the sample size is large, the variances of these estimators are

$$V(\hat{\bar{Y}}_R) \approx \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N \sum_{i=1}^N (y_i - Rx_i)^2$$

$$V(\hat{Y}_R) \approx \frac{N^2(1-f)}{n(N-1)} \sum_{i=1}^N (y_i - Rx_i)^2$$

and
$$V(\hat{R}) \approx \frac{1-f}{n\bar{X}^2} \frac{1}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2,$$

where $\hat{Y}_R = \frac{\bar{y}}{\bar{x}}\bar{X}$, $\hat{Y}_R = \frac{y}{x}X = \frac{\bar{y}}{\bar{x}}$, $\hat{R} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}}$, $f = \frac{n}{N}$.

Proof: Let us first check whether \hat{R} is a biased or unbiased estimator of R . We can write,

$$\hat{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}}. \text{ We can also write :}$$

$$\frac{1}{\bar{x}} = \frac{1}{\bar{x} + (\bar{x} - \bar{X})} = \frac{1}{\bar{x}} \left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right)^{-1} \approx \frac{1}{\bar{X}} \left(1 - \frac{\bar{x} - \bar{X}}{\bar{X}}\right).$$

[neglecting the higher powers of $(\bar{x} - \bar{X})/\bar{X}$].

Hence,
$$\hat{R} - R \approx \frac{\bar{y} - R\bar{x}}{\bar{X}} \left(1 - \frac{\bar{x} - \bar{X}}{\bar{X}}\right).$$

Now
$$E(\bar{y} - R\bar{x}) = E(\bar{y}) - R(\bar{x}) = \bar{Y} - R\bar{X} = 0.$$

Therefore,
$$E(\hat{R} - R) \approx -E\frac{(\bar{y} - R\bar{x})(\bar{x} - \bar{X})}{\bar{X}^2}.$$

But
$$E\bar{y}(\bar{x} - \bar{X}) = E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) = \frac{1-f}{n} \rho S_x S_y,$$

where
$$S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2, \quad S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2,$$

$$\rho = \frac{E(\bar{x} - \bar{X})(\bar{y} - \bar{Y})}{S_x S_y}.$$

Again,
$$E\bar{x}(\bar{x} - \bar{X}) = E(\bar{x} - \bar{X})(\bar{x} - \bar{X}) = \frac{1-f}{n} S_x^2.$$

Hence,
$$\begin{aligned} E(\hat{R} - R) &\approx \frac{1-f}{n\bar{X}^2} [RS_x^2 - \rho S_y S_x] = \frac{1-f}{n} \left[\frac{S_x^2}{\bar{X}^2} - \frac{S_{yx}}{R\bar{X}^2} \right] R \\ &= \frac{1-f}{n} [C_x^2 - C_{xy}], \quad C_x^2 = \frac{S_x^2}{\bar{X}^2}, \quad C_{xy} = \frac{S_{yx}}{\bar{X}\bar{Y}}. \end{aligned}$$

Hence, \hat{R} is not an unbiased estimator of R . Therefore, other estimators involving \hat{R} are also not unbiased.

Let us now find the variance of \hat{R} , where

$$\begin{aligned} V(\hat{R}) &= E\{\hat{R} - E(\hat{R})\}^2 \approx E(\hat{R} - R)^2 \quad [\because E(\hat{R}) \approx R] \\ &= E\left[\frac{\bar{y}}{\bar{x}} - R\right]^2 = E\left(\frac{\bar{y} - R\bar{x}}{\bar{x}}\right)^2 \approx \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2. \end{aligned}$$

Let $u_i = y_i - Rx_i \quad (i = 1, 2, \dots, N).$

Then $\bar{u} = \bar{y} - R\bar{x}$ and $\bar{U} = \bar{Y} - R\bar{X} = 0$.

$$\therefore V(\hat{R}) \approx \frac{1}{\bar{X}^2} E(\bar{u} - \bar{U})^2 = \frac{1}{\bar{X}^2} V(\bar{u}) = \frac{1}{\bar{X}^2} \frac{1-f}{n} S_u^2.$$

But
$$S_u^2 = \frac{1}{N-1} \sum_{i=1}^N (u_i - \bar{U})^2 = \frac{1}{N-1} \sum u_i^2 = \frac{1}{N-1} \sum (y_i - Rx_i)^2.$$

Hence,
$$V(\hat{R}) \approx \frac{1-f}{n\bar{X}^2} \frac{1}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2.$$

We know $\hat{Y}_R = \hat{R}\bar{X}$.

$$V(\hat{Y}_R) = \bar{X}^2 V(\hat{R}) \approx \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2.$$

Also $\hat{Y}_R = \hat{R}X = \hat{R}N\bar{X}$.

$$\therefore V(\hat{Y}_R) = N^2 \bar{X}^2 V(\hat{R}) \approx \frac{N^2(1-f)}{n} \frac{1}{N-1} \sum (y_i - Rx_i)^2.$$

It is observed that the ratio estimator is obtained using the values of y -variable and x -variable. But the values of these variables may vary from sample to sample and hence, the distribution of ratio estimator may be different. However, the limiting distribution of ratio estimator follows normal distribution, if sample size becomes large. This indicates that the ratio estimator is consistent though it is biased. If the sample size is not large enough, the distribution of ratio estimator is positively skewed.

Corollary : In simple random sampling the variance of ratio estimator \hat{R} of R is

$$V(\hat{R}) \approx \frac{1-f}{n\bar{X}^2} [S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y] = \frac{(1-f)R^2}{n} [C_y^2 + C_x^2 - 2\rho C_x C_y],$$

where $C_x = \frac{S_x}{\bar{X}}$ and $C_y = \frac{S_y}{\bar{Y}}$.

Proof: We have

$$\begin{aligned} V(\hat{R}) &\approx \frac{1-f}{n\bar{X}^2} \frac{1}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2 \\ &= \frac{1-f}{n\bar{X}^2} \frac{1}{N-1} \sum [(y_i - \bar{Y}) - R(x_i - \bar{X})]^2, \quad \because \bar{Y} - R\bar{X} = 0 \\ &= \frac{1-f}{n\bar{X}^2} \left[\frac{1}{N-1} \sum (y_i - \bar{Y})^2 + \frac{R^2}{N-1} \sum (x_i - \bar{X})^2 \right. \\ &\quad \left. - \frac{2R}{N-1} \sum (x_i - \bar{X})(y_i - \bar{Y}) \right] \\ &= \frac{1-f}{n\bar{X}^2} [S_y^2 + R^2 S_x^2 - 2RS_{yx}] \\ &= \frac{1-f}{n\bar{X}^2} [S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y], \quad \rho = \frac{S_{yx}}{S_x S_y} \end{aligned}$$

$$= \frac{(1-f)R^2}{n} \left[\frac{S_y^2}{Y^2} + \frac{S_x^2}{X^2} - 2\rho \frac{S_x}{X} \frac{S_y}{Y} \right], \quad \because R = \frac{\bar{Y}}{\bar{X}}$$

$$= \frac{R^2(1-f)}{n} [C_y^2 + C_x^2 - 2\rho C_x C_y].$$

Corollary : In simple random sampling the variance of the ratio estimator $\hat{\bar{Y}}_R$ of \bar{Y} is

$$V(\hat{\bar{Y}}_R) \approx \frac{1-f}{n} [S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y] = \frac{\bar{Y}^2(1-f)}{n} [C_y^2 + C_x^2 - 2\rho C_x C_y].$$

Corollary : In simple random sampling the variance of the ratio estimator \hat{Y}_R of \bar{Y} is

$$V(\hat{Y}_R) \approx \frac{(1-f)N^2}{n} [S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y]$$

$$= \frac{(1-f)\bar{Y}^2}{n} [C_y^2 + C_x^2 - 2\rho C_x C_y].$$

From the above results it is observed that the variances of \hat{R} , $\hat{\bar{Y}}_R$ and \hat{Y}_R are multiple of a fixed quantity

$$\frac{1-f}{n} [C_y^2 + C_x^2 - 2\rho C_x C_y].$$

This quantity is the square of coefficient of variation. Thus, we can write :

$$(\text{C.V.})^2 = \frac{V(\hat{Y}_R)}{\bar{Y}^2} = \frac{1-f}{n} [C_y^2 + C_x^2 - 2\rho C_x C_y].$$

This $(\text{C.V.})^2$ is same for all estimators. This $(\text{C.V.})^2$ is called relative variance of the estimator [Hansen., Hurwitz and Madow].

Corollary : If $C_y = C_x = C$, then $V\left(\frac{\hat{R}}{R}\right) = \frac{1-f}{n} 2C^2(1-\rho)$.

Proof : We have $V(\hat{R}) \approx \frac{1-f}{n} R^2 [C_y^2 + C_x^2 - 2\rho C_x C_y]$.

$$\therefore V\left(\frac{\hat{R}}{R}\right) = \frac{1-f}{n} [C_y^2 + C_x^2 - 2\rho C_x C_y] = \frac{1-f}{n} (2C^2 - 2\rho C^2), \quad \because C_y = C_x = C$$

$$= \frac{1-f}{n} 2C^2(1-\rho).$$

Corollary : In simple random sampling the ratio of bias of \hat{R} to the standard error (S.E.) of \hat{R} is

$$\frac{\text{bias}}{\text{S.E.}} = \frac{\frac{1-f}{n\bar{X}^2} (RS_x^2 - \rho S_y S_x)}{\bar{X} \left[\frac{1-f}{n} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x) \right]^{1/2}} = \sqrt{\frac{1-f}{n\bar{X}^2}} \frac{S_x (RS_x - \rho S_y)}{(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x)^{1/2}}$$

$$= \text{C.V.}(\bar{x}) \frac{RS_x - \rho S_y}{(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x)^{1/2}}, \quad \text{where } \text{C.V.}(\bar{x}) = \frac{S_x \sqrt{1-f}}{\bar{X} \sqrt{n}}.$$

Also, we have $\frac{\text{bias}}{\text{S.E.}} = \text{Cov}(\bar{x}) \frac{(R^2 S_x^2 + \rho^2 S_y^2 - 2R\rho S_x S_y)^{1/2}}{(R^2 S_x^2 + S_y^2 - 2R\rho S_x S_y)^{1/2}}$.

$$\therefore \frac{\text{bias}}{\text{S.E.}} \leq \text{Cov}(\bar{x}), \quad \because \rho^2 \leq 1.$$

Kish, Namboodri and Pillai (1962) have shown that in maximum cases the above ratio is less than 0.03.

Hartley and Ross (1954) have proposed a formula to find the bias. They have utilized the covariance of \hat{R} and \bar{x} in case of simple random sampling to find the bias of \hat{R} . This covariance is

$$\text{Cov}(\hat{R}, \bar{x}) = E\left(\frac{\bar{y}}{\bar{x}}, \bar{x}\right) - E\left(\frac{\bar{y}}{\bar{x}}\right) E(\bar{x}) = \bar{Y} - \bar{X}E(\hat{R}).$$

$$\text{We have } E(\hat{R}) = \frac{\bar{Y}}{\bar{X}} - \frac{\text{Cov}(\hat{R}, \bar{x})}{\bar{X}} = R - \frac{1}{\bar{X}} \text{Cov}(\hat{R}, \bar{x}).$$

Therefore, the bias of \hat{R} is

$$E(\hat{R}) - R = -\frac{1}{\bar{X}} \text{Cov}(\hat{R}, \bar{x}).$$

The above bias formula is exact since the approximate value of $\left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right)^{-1}$ is not used to find bias. We have

$$|\text{bias}(\hat{R})| = \frac{\text{Cov}(\hat{R}, \bar{x})}{\bar{X}} = \frac{\rho_{\hat{R}, \bar{x}} \sigma_{\hat{R}} \sigma_{\bar{x}}}{\bar{X}} \leq \frac{\sigma_{\hat{R}} \sigma_{\bar{x}}}{\bar{X}} \quad [\because \rho_{\hat{R}, \bar{x}} \leq 1]$$

$$\therefore \frac{\text{bias}(\hat{R})}{\sigma_{\hat{R}}} \leq \text{Cov}(\hat{R}, \bar{x}).$$

The above ratio is true in finding \hat{Y}_R and $\hat{\bar{Y}}_R$. From the above result Hartley and Ross have concluded that the bias of ratio estimator is negligible if $\text{Cov}(\hat{R}, \bar{x}) < 0.1$.

15.2 Estimation of Variance of Ratio Estimator from Sample

$$\text{We have } V(\hat{Y}_R) \approx \frac{1-f}{n} \frac{N^2}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2, \quad f = \frac{n}{N}.$$

$$\text{To estimate } V(\hat{Y}_R) \text{ we need to estimate } S_u^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2.$$

$$\text{Cuchran (1977) has shown that } s_u^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2, \quad \hat{R} = \frac{\bar{y}}{\bar{x}}$$

can be used as an estimator of S_u^2 . This estimator is biased of order $1/n$. Hence, the estimator of $V(\hat{Y}_R)$ is

$$\begin{aligned} v(\hat{Y}_R) &= \frac{N^2(1-f)}{n} \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \\ &= \frac{N^2(1-f)}{n} \frac{1}{n-1} \left[\sum y_i^2 + \hat{R}^2 \sum x_i^2 - 2\hat{R} \sum x_i y_i \right] \\ &= \frac{N^2(1-f)}{n} [s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{yx}], \end{aligned}$$

where $s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$, $s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$, $s_{yx} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$.

In a similar way the estimator of $V(\hat{R})$ obtained as

$$v_1(\hat{R}) = \frac{1-f}{n\bar{X}^2} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}), \text{ if } \bar{X} \text{ is known}$$

and $v_2(\hat{R}) = \frac{1-f}{n\bar{x}^2} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}), \text{ if } \bar{X} \text{ is not known.}$

The estimator of $V(\hat{Y}_R)$ is

$$v(\hat{Y}_R) = \frac{1-f}{n} [s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}].$$

Using the above estimators of $V(\hat{R})$ and $V(\hat{Y}_R)$ the confidence limits of \bar{R} and \bar{Y} are found out, where the $(1 - \alpha)100\%$ confidence limits are

$$\hat{R} \pm Z_{\frac{\alpha}{2}} \sqrt{v(\hat{R})} \quad \text{and} \quad \hat{Y}_R \pm Z_{\frac{\alpha}{2}} \sqrt{v(\hat{Y}_R)}.$$

Here $Z_{\frac{\alpha}{2}}$ is the value of normal variate for probability α . However, if the distribution of \hat{R} is skewed, the confidence limit of R is

$$\hat{R} \pm Z_{\frac{\alpha}{2}} \sqrt{C_{\bar{y}\bar{y}} + C_{\bar{x}\bar{x}} - 2C_{\bar{y}\bar{x}}},$$

where $C_{\bar{y}\bar{y}} = \frac{1-f}{n} \frac{s_y^2}{\bar{y}^2}$, $C_{\bar{x}\bar{x}} = \frac{1-f}{n} \frac{s_x^2}{\bar{x}^2}$, $C_{\bar{y}\bar{x}} = \frac{1-f}{n} \frac{s_{yx}}{\bar{x}\bar{y}}$.

Example 15.1 : To estimate the jute production in an area 30 farmers living in the area are randomly selected. The jute grower farmers in the area are 500. The information on jute production (y in 100 kg) and the amount of land cultivated (x in acre) are recorded from the selected farmers. Find ratio estimate of total jute production in the area. Also find 95% confidence limit of the total jute production. Calculate the relative precision of your estimator compared to simple estimator. The total land used in the area for jute cultivation is approximately 450.00 acres.

Sl. No.	Jute production y (in 100 kg)	Land cultivated x (in acre)	Sl. No.	Jute production y (in 100 kg)	Land cultivated x (in acre)
1	2.2	0.5	11	4.5	1.0
2	2.6	0.8	12	3.0	0.5
3	4.0	1.0	13	2.8	0.6
4	6.5	2.0	14	2.4	0.6
5	4.6	1.2	15	3.5	0.8
6	6.0	2.2	16	4.5	1.4
7	2.5	0.5	17	8.0	1.5
8	2.5	0.6	18	10.0	2.0
9	3.0	0.9	19	12.5	2.2
10	7.0	1.5	20	8.5	1.8

Sl. No.	Jute production y (in 100 kg)	Land cultivated x (in acre)	Sl. No.	Jute production y (in 100 kg)	Land cultivated x (in acre)
21	9.0	1.5	26	3.5	0.8
22	3.5	0.7	27	9.5	2.4
23	3.0	0.5	28	12.0	2.5
24	2.0	0.4	29	10.5	1.5
25	1.5	0.2	30	4.8	1.0

Solution : We have $N = 500$, $n = 30$, $X = 450.00$, $\sum y = 159.9$, $\sum x = 35.1$, $\bar{y} = 5.33$, $\bar{x} = 1.17$ $\hat{R} = \bar{y}/\bar{x} = 4.56$. Hence, the estimate of total jute production is

$$\hat{Y}_R = \hat{R}X = 4.56 \times 450.00 = 2052.00.$$

The estimate of variance of \hat{Y}_R is

$$v(\hat{Y}_R) = \frac{N^2(1-f)}{n} [s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}], \quad f = \frac{30}{500} = 0.06.$$

$$s_y^2 = \frac{1}{n-1} \left[\sum y^2 - \frac{(\sum y)^2}{n} \right] = \frac{1}{30-1} \left[1153.65 - \frac{(159.9)^2}{30} \right] = 10.3925.$$

$$s_x^2 = \frac{1}{n-1} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right] = \frac{1}{30-1} \left[53.83 - \frac{(35.1)^2}{30} \right] = 0.4401.$$

$$s_{yx} = \frac{1}{n-1} \left[\sum xy - \frac{\sum x \sum y}{n} \right] = \frac{1}{30-1} \left[243.37 - \frac{159.9 \times 35.1}{30} \right] = 1.9409.$$

$$\begin{aligned} \therefore v(\hat{Y}_R) &= \frac{(500)^2(1-0.06)}{30} [10.3925 + (4.56)^2 0.4401 - 2 \times 4.56 \times 1.9409] \\ &= 17466.24955. \end{aligned}$$

$$\therefore \text{s.e.}(\hat{Y}_R) = \sqrt{v(\hat{Y}_R)} = 132.16.$$

The 95% confidence limits of Y are

$$\begin{aligned} \hat{Y}_{RL} &= \hat{Y}_R - Z_{0.025} \text{s.e.}(Y_R) \quad \text{and} \quad \hat{Y}_{RU} = \hat{Y}_R + Z_{0.025} \text{s.e.}(\hat{Y}_R) \\ &= 2052.00 - 1.96 \times 132.16 \quad \quad \quad = 2052.00 + 1.96 \times 132.6 \\ &= 1792.97. \quad \quad \quad \quad \quad \quad \quad = 2311.90. \end{aligned}$$

The estimate of variance of total jute production on the basis of simple estimate [estimate based on simple random sampling] is

$$v(\hat{Y}) = \frac{N^2(1-f)}{n} s_y^2 = \frac{(500)^2(1-0.06)}{30} \times 10.3925 = 81407.92.$$

Therefore, the relative precision of ratio estimator compared to simple estimator is

$$\text{R.P.} = \frac{v(\hat{Y}) - v(\hat{Y}_R)}{v(\hat{Y}_R)} = \frac{81407.92 - 17466.25}{17466.25} = 3.661 \text{ or } 366.1\%.$$

The estimate of average jute production per farmer is

$$\hat{Y}_R = \hat{R} \bar{X} = 4.56 \times 0.9 = 4.104 \text{ (100 kg) } [\because \bar{X} = 0.9].$$

The estimate of variance of \hat{Y}_R is

$$\begin{aligned} v(\hat{Y}_R) &= \frac{1-f}{n} [s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}] \\ &= \frac{(1-0.06)}{30} [10.3925 + (4.56)^2 0.4401 - 2 \times 4.56 \times 1.9409] \\ &= 0.0577. \end{aligned}$$

15.3 Comparison of Ratio Estimator and Simple Estimator

The simple estimator of population mean also known as sample mean per unit is \bar{y} , where $\bar{y} = \frac{1}{n} \sum y$. The estimator of \bar{Y} in case of ratio estimator is \hat{Y}_R . This ratio estimator will be efficient than \bar{y} , if $V(\hat{Y}_R) < V(\bar{y})$.

Theorem : In simple random sampling if sample size is large, the $V(\hat{Y}_R)$ is less than $V(\bar{y})$ when

$$\rho > \frac{1}{2} \frac{S_x}{\bar{X}} / \frac{S_y}{\bar{Y}} = \frac{1}{2} \frac{C.V.(x)}{C.V.(y)}.$$

Proof: We have $V(\bar{y}) = \frac{1-f}{n} s_y^2$ and $V(\hat{Y}_R) = \frac{1-f}{n} [S_y^2 + R^2 S_x^2 - 2RS_{yx}]$.

If $V(\hat{Y}_R)$ is less than $V(\bar{y})$, we can write :

$$\begin{aligned} S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x &< S_y^2 \\ 2\rho S_y S_x &> R S_x^2 \\ \rho > \frac{1}{2} R \frac{S_x}{S_y} &= \frac{1}{2} \frac{\bar{Y}}{\bar{X}} \frac{S_x}{S_y} = \frac{1}{2} \frac{S_x}{\bar{X}} / \frac{S_y}{\bar{Y}} = \frac{1}{2} \frac{C.V.(x)}{C.V.(y)}. \end{aligned}$$

15.4 Unbiased Ratio Estimator and its Variance

It has already been shown that the ratio estimator in case of simple random sampling is not unbiased. Let us investigate the sampling method which may provide unbiased ratio estimator. One of the method of sampling is to select the sampling units with varying probabilities. Lahiri (1951) has proposed such a method. According to Lahiri the sampling units are to be selected on the basis of probability proportional to total size of sampling units (ppts). In this scheme, the first unit is selected with probability proportional to $\sum_{i=1}^n x_i$ and the remaining $(n-1)$ units are to be selected with equal probability. Midzuno (1951) has proposed to select first unit with probability proportional to x_1 and the remaining $(n-1)$ units are to be selected from $(N-1)$ units with equal probability without replacement.

Let there be N units in a finite population. We need a sample of size n . Let the first unit in the sample be selected with probability p_i ($i = 1, 2, \dots, N$), where $\sum_{i=1}^N p_i = 1$. The remaining $(n-1)$ units are selected with equal probability without replacement. Consider that i -th unit

is selected first. The probability of selection of i -th unit and the remaining $(n-1)$ units under simple random sampling scheme is

$$p_i / \binom{N-1}{n-1}, \quad i = 1, 2, \dots, N.$$

Now, if $p_i = x_i/X$, then the probability of selection of a sample of size n is

$$P(S) = \frac{\sum_{i=1}^n p_i}{\binom{N-1}{n-1}} = \frac{\sum_{i=1}^n x_i}{\binom{N-1}{n-1} X}.$$

Here S is the sample space [$S \equiv S(y_1, y_2, \dots, y_n)$].

The ratio estimator of population total under the sampling scheme mentioned above is

$$\hat{Y}_R = \hat{R}X = \frac{\bar{y}}{\bar{x}}X.$$

Theorem : If sampling units are selected with varying probabilities and without replacement, then the ratio estimator of population total is unbiased.

Proof : The ratio estimator of population total is $\hat{Y}_R = \frac{\bar{y}}{\bar{x}}X = \frac{y}{x}X$.

$$\begin{aligned} \text{Now, } E(\hat{Y}_R) &= E\left[\frac{y}{x}X\right] = X E\left(\frac{y}{x}\right) = X \sum_{i=1}^N \frac{y}{x} \frac{x}{\binom{N-1}{n-1} X} \\ &= \frac{1}{\binom{N-1}{n-1}} \sum_1 y = Y. \end{aligned}$$

Here \sum_1 indicates the total of y_i included in all samples. Hence, ratio estimator is unbiased.

In a similar way, we can show that the ratio estimator of \bar{Y} is $\hat{\bar{Y}}_R$ and it is also unbiased.

The variance of \hat{Y}_R is [Raj (1964)]

$$V(\hat{Y}_R) = \frac{X}{\binom{N-1}{n-1}} \sum_1 \frac{y^2}{x} - Y^2.$$

The estimator of this variance is

$$v(\hat{Y}_R) = \hat{Y}_R^2 - \frac{N\bar{X}}{x} \left[\bar{y}^2 - \left(\frac{1}{n} - \frac{1}{N} \right) s_y^2 \right].$$

Unbiased ratio estimator in case of sampling with replacement : Let $y_i = \bar{Y} + \epsilon_i$.

Now, if we have a sample of n observations, then

$$\bar{y} = \bar{Y} + \bar{\epsilon}, \quad \text{where } E(\bar{\epsilon}) = 0 \text{ and } E(\bar{\epsilon}^2) = \frac{N-n}{Nn} S_y^2.$$

Again, let $x_i = \bar{X} + \epsilon_1$. Then $\bar{x} = \bar{X} + \bar{\epsilon}_1$. Here also

$$E(\bar{\epsilon}_1) = 0, \quad E(\bar{\epsilon}_1^2) = \frac{N-n}{Nn} S_x^2.$$

Now, if sampling units are selected with varying probabilities, we can define Z_i for i -th unit, where

$$Z_i = \frac{y_i}{Np_i} = \bar{Y} + \epsilon_i.$$

Also, we can define $v_i = \frac{x_i}{Np_i} = \bar{X} + \epsilon_{1i}$

The ratio of Z_i to v_i is $r_i = \frac{Z_i}{v_i} = \frac{y_i}{x_i}$.

Then $\hat{R} = \frac{\bar{Z}}{\bar{v}}$.

We have $E(Z_i) = \bar{Y}$, $E(v_i) = \bar{X}$, $E(\bar{Z}) = \bar{Y}$, $E(\bar{v}) = \bar{X}$.

$$E(\bar{\epsilon}^2) = \frac{\sigma_z^2}{n} = \frac{1}{n} \sum_{i=1}^N p_i (Z_i - \bar{Y})^2$$

$$E(\bar{\epsilon}_1^2) = \frac{\sigma_v^2}{n} = \frac{1}{n} \sum_{i=1}^N p_i (v_i - \bar{X})^2$$

and
$$E(\bar{\epsilon} \bar{\epsilon}_1) = - \left\{ \sum_{i=1}^N p_i (Z_i - \bar{Y})(v_i - \bar{X}) \right\} = -\frac{1}{n} \rho \sigma_z \sigma_v.$$

We can express \hat{R} in terms of $\bar{\epsilon}$ and $\bar{\epsilon}_1$, where

$$\hat{R} = \frac{\bar{Y} \left(1 + \frac{\bar{\epsilon}}{\bar{Y}} \right)}{\bar{X} \left(1 + \frac{\bar{\epsilon}_1}{\bar{X}} \right)}.$$

Here
$$E(\hat{R}) = R + R \left[\frac{E(\bar{\epsilon}_1^2)}{\bar{X}^2} - \frac{E(\bar{\epsilon} \bar{\epsilon}_1)}{\bar{Y} \bar{X}} \right].$$

Now, putting the values of $E(\bar{\epsilon}_1^2)$ and $E(\bar{\epsilon} \bar{\epsilon}_1)$, in $E(\hat{R})$, we get

$$E(\hat{R}) = R \left[1 + \frac{1}{n} \left(\frac{\sigma_v^2}{\bar{X}^2} - \frac{\rho \sigma_z \sigma_v}{\bar{Y} \bar{X}} \right) \right].$$

Therefore, the variance of \hat{R} is

$$V(\hat{R}) = \frac{R^2}{n} \left(\frac{\sigma_z^2}{\bar{Y}^2} + \frac{\sigma_v^2}{\bar{X}^2} - \frac{2\rho \sigma_z \sigma_v}{\bar{Y} \bar{X}} \right).$$

The estimator of this variance is

$$v(\hat{R}) = \frac{1}{n} \frac{1}{\bar{v}^2} \frac{1}{n-1} \sum_{i=1}^n (Z_i - \hat{R}v_i)^2.$$

The estimator of \bar{Y} is $\hat{\bar{Y}} = \hat{R} \bar{X}$ and the variance of this estimator is

$$V(\hat{\bar{Y}}) = \frac{1}{n} \sum_{i=1}^N p_i (Z_i - Rv_i)^2 = \frac{1}{nN^2} \sum_{i=1}^n \frac{1}{p_i} (Y_i - Rx_i)^2.$$

The estimator of this variance is

$$v(\hat{Y}) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n (Z_i - \hat{R}v_i)^2.$$

If the probability of selection of i -th unit is proportional to x_i , i.e.,

$$p_i = x_i/N\bar{X}.$$

Then $Z_i = \frac{y_i}{Np_i} = \frac{y_i}{x_i} \bar{X} = r_i \bar{X}; v_i = \frac{x_i}{Np_i} = \bar{X}.$

We have $\bar{Z} = \bar{r} \bar{X}, \bar{v} = \bar{X}, \hat{R} = \bar{r}, R = \frac{1}{N} \sum_{i=1}^N r_i, \hat{Y}_R = \bar{X} \bar{r} = \bar{Z}.$

Therefore, $E(\hat{Y}_R) = E(\bar{z}) = \bar{Y}$ and $V(\hat{Y}_R) = \frac{\bar{X}^2}{n} \sum_{i=1}^N p_i \left(r_i - \frac{1}{N} \sum_{i=1}^N r_i \right)^2.$

The estimator of this variance is

$$v(\hat{Y}_R) = \frac{\bar{X}^2}{n} \frac{1}{n-1} \sum_{i=1}^n (r_i - \bar{r})^2.$$

Example 15.2 : A dairy project is conducted in 25 villages in a district. The number of cattle and the number of cows in two consecutive years in the villages are shown below :

SL. No. of village	No. of cattle in previous year a_i	No. of cows in previous year x_i	No. of cows in following year, y_i
01	807	210	245
02	408	115	109
03	642	180	195
04	538	148	162
05	312	90	90
06	416	100	120
07	363	75	82
08	162	52	63
09	612	163	160
10	540	178	180
11	343	107	125
12	275	82	95
13	178	63	70
14	578	162	175
15	607	175	180
16	361	122	131
17	195	48	52
18	212	50	51
19	473	138	132
20	308	147	142

SL. No. of village	No. of cattle in previous year a_i	No. of cows in previous year x_i	No. of cows in following year, y_i
21	213	62	58
22	414	152	147
23	162	75	68
24	470	168	180
25	285	62	70
Total	$A = 9874$	$X = 2924$	

(i) Select a random sample of 40% villages with probability proportional to total number of cattle in the villages and estimate the number of cows in the following year. Also estimate the variance of your estimator.

(ii) Compare your estimator with simple estimator under PPS scheme.

Solution : Here $N = 25$, $n = 10$, $X = 2924$, $A = 9874$.

Following Lahiri's rule and using random number table given in Appendix the pairs of selected random numbers are :

(16, 161), (11, 111), (10, 100), (19, 190), (17, 175), (09, 095), (01, 018), (24, 242), (25, 252), (20, 203).

Therefore, the serial numbers of selected villages are 16, 11, 10, 19, 17, 09, 01, 24, 25, 20.

The information of sample observations are given below :

SL. No.	a_i	x_i	y_i	$p_i = \frac{a_i}{A}$	$Z_i = \frac{y_i}{Np_i}$	$v_i = \frac{x_i}{Np_i}$	$r_i = \frac{y_i}{x_i}$
16	361	122	131	0.03656	143.33	133.48	1.0738
11	343	107	125	0.03474	143.93	123.20	1.1682
10	540	178	180	0.05469	131.65	130.19	1.0112
19	473	138	132	0.04790	110.23	115.24	0.9565
17	195	48	52	0.01975	105.32	97.21	1.0833
09	612	163	160	0.06198	103.26	105.19	0.9816
01	807	210	245	0.08173	119.91	102.78	1.1667
24	470	168	180	0.04760	151.26	141.18	1.0714
25	285	62	70	0.02886	97.02	85.93	1.1290
20	308	147	142	0.03119	182.11	188.52	0.9660
Total		1343	1417		1288.02	1222.92	10.6077

Here $\hat{R} = \frac{\sum Z_i}{\sum v_i} = \frac{1288.02}{1222.92} = 1.0532$.

The total number of cows in the study area is

$\hat{Y}_R = \hat{R}X = 1.0532 \times 2924 = 3080$.

The estimates of variance of \hat{Y}_R is

$$v(\hat{Y}_R) = \frac{N^2}{n(n-1)} \sum_{i=1}^n (Z_i - \hat{R}v_i)^2 = \frac{N^2}{n(n-1)} \left[\sum Z_i^2 - 2\hat{R} \sum Z_i v_i + \hat{R}^2 \sum v_i^2 \right]$$

$$\begin{aligned}
&= \frac{(25)^2}{10(10-1)} [172331.5674 - 2 \times 1.0532 \times 164153.9021 + (1.0532)^2 \times 157158.8404] \\
&= \frac{625 \times 585.3181}{90} = 4064.7092.
\end{aligned}$$

(ii) The simple estimator of cows is

$$\hat{Y} = N\bar{z} = 25 \times \frac{1288.02}{10} = 3220.05$$

$$\begin{aligned}
v(\hat{Y}) &= \frac{N^2}{n(n-1)} \left[\sum Z_i^2 - \frac{(\sum Z_i)^2}{n} \right] = \frac{(25)^2}{10(10-1)} \left[172331.5674 - \frac{(1288.02)^2}{10} \right] \\
&= 44666.7733.
\end{aligned}$$

The relative precision of ratio estimator compared to simple estimator is

$$\frac{v(\hat{Y}) - v(\hat{Y}_R)}{v(\hat{Y}_R)} \times 100 = \frac{44666.7733 - 4064.7092}{4064.7092} \times 100 = 998.89\%.$$

The sample is selected with probability proportional to number of cattles ($p_i = a_i/A$). The sample can also be selected with $p_i = x_i/X$. In that case,

$$\hat{Y}_R = N\bar{Z}, \quad \bar{Z} = \bar{r}\bar{X}, \quad \text{where } \bar{r} = \frac{1}{n} \sum r_i = 1.06077.$$

$$\therefore \bar{Z} = 1.06077 \times 116.96 = 124.0676$$

$$\text{and } \hat{Y} = 25 \times 124.0676 = 3101.69 \approx 3102.$$

The estimate of variance of \hat{Y}_R is

$$\begin{aligned}
v(\hat{Y}_R) &= \frac{\bar{X}^2 N^2}{n(n-1)} \sum (r_i - \bar{r})^2 = \frac{\bar{X}^2 N^2}{n(n-1)} \left[\sum r_i^2 - \frac{(\sum r_i)^2}{n} \right] \\
&= \frac{(124.0676)^2 (25)^2}{10(10-1)} \left[11.3091 - \frac{(10.6077)^2}{10} \right] \\
&= 6068.3931.
\end{aligned}$$

15.5 Unbiased Ratio Type Estimator

It has already been observed that the estimators of population total and population mean are unbiased if these estimators are found out using the value of $\bar{r} = \frac{1}{n} \sum r_i = \frac{1}{n} \sum_{i=1}^n y_i/x_i$.

However, the estimator is unbiased if sampling units are selected under PPS sampling scheme. Let us now consider the estimator using \bar{r} when sample is selected under simple random sampling scheme. Hartley and Ross (1954) have proposed ratio estimator using \bar{r} . They also proposed to modify the bias in the ratio estimator.

$$\text{Let } \bar{r} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i} = \frac{1}{n} \sum_{i=1}^n r_i.$$

$$\begin{aligned}
\text{Now, } \frac{1}{N} \sum_{i=1}^N r_i (x_i - \bar{X}) &= \frac{1}{N} \sum_{i=1}^N \frac{y_i}{x_i} x_i - \frac{1}{N} \sum_{i=1}^N \frac{y_i}{x_i} \bar{X} \\
&= \bar{Y} - \bar{X} E(r_i) = \bar{X} [R - E(r_i)].
\end{aligned}$$

But in simple random sampling $E(r_i) = E(\bar{r})$.

Therefore, bias $(\bar{r}) = E(\bar{r}) - R = -\frac{1}{N\bar{X}} \sum_{i=1}^N r_i(x_i - \bar{X})$.

Hence, to get the unbiased estimator of R we need to find unbiased estimator of

$$\frac{1}{N\bar{X}} \sum_{i=1}^N r_i(x_i - \bar{X}).$$

We know that, if (y_i, x_i) is the joint value of variable observed from i -th unit ($i = 1, 2, \dots, N$) and if \bar{y}, \bar{x} are the simple random sample mean of n observations of the variable y and x , respectively, then

$$E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}).$$

Using this result, we can say that the estimator

$$\frac{1}{n-1} \sum_{i=1}^n r_i(x_i - \bar{x}) = \frac{n}{n-1} (\bar{y} - \bar{r}\bar{x})$$

is an unbiased estimator of $\frac{1}{N-1} \sum_{i=1}^N r_i(x_i - \bar{X})$.

Therefore, using this value in the formula of bias of \bar{r} , we get the ratio estimator [Hartley and Ross]

$$\hat{R}_{HR} = \bar{r} + \frac{n(N-1)}{(n-1)\bar{X}N} (\bar{y} - \bar{r}\bar{x}).$$

The ratio estimator of population total is

$$\hat{Y}_{HR} = \left[\bar{r} + \frac{n(N-1)}{(n-1)\bar{X}N} (\bar{y} - \bar{r}\bar{x}) \right] X = \bar{r}X + \frac{n(N-1)}{(n-1)} (\bar{y} - \bar{r}\bar{x}).$$

Also, we have, the ratio estimator of population mean, where

$$\hat{\bar{Y}}_{HR} = \hat{R}_{HR}\hat{\bar{X}} = \left[\bar{r} + \frac{n(N-1)}{(n-1)N\bar{X}} (\bar{y} - \bar{r}\bar{x}) \right] \bar{X} = \bar{r}\bar{X} + \frac{n(N-1)}{N(n-1)} (\bar{y} - \bar{r}\bar{x}).$$

For large sample size the variance of $\hat{\bar{Y}}_{HR}$ is

$$V(\hat{\bar{Y}}_{HR}) \approx \frac{1}{n} [S_y^2 + R_1^2 S_x^2 - 2R_1 S_{yx}], \text{ where } R_1 = \frac{1}{N} \sum_{i=1}^N \frac{y_i}{x_i}.$$

For large sample size, we have

$$V(\hat{\bar{Y}}_R) \approx \frac{1}{n} [S_y^2 + R^2 S_x^2 - 2R S_{yx}], \text{ where } R = \bar{Y}/\bar{X}.$$

Now, comparing $V(\hat{\bar{Y}}_R)$ and $V(\hat{\bar{Y}}_{HR})$, we have

$$V(\hat{\bar{Y}}_R) - V(\hat{\bar{Y}}_{HR}) = \frac{S_x^2}{n} [(R - B)^2 - (R_1 - B)^2], \text{ where } B = \frac{S_{yx}}{S_x^2}.$$

Thus, it is observed that, for n large, \hat{Y}_{HR} will be more precise than \hat{Y}_R , if and only if B is more close to R_1 compared to R . Both estimator will be equally precise, if $R_1 = R$. Hence, we can write :

$$V(\hat{R}_{HR}) \approx \frac{S_y^2 + R_1^2 S_x^2 - 2R_1 S_{yx}}{n\bar{X}^2}.$$

The unbiased estimator of variance of \hat{Y}_{HR} is

$$v(\hat{Y}_{HR}) \approx \frac{1}{n} [s_y^2 + \bar{r}^2 s_x^2 - 2\bar{r} s_{yx}], \text{ where } \bar{r} = \frac{1}{n} \sum_{i=1}^n y_i/x_i.$$

Example 15.3 : Using the data of previous example, find unbiased ratio type estimator of total cows and also find estimate of variance of your estimator.

Solution : We have $N = 25$, $n = 10$, $\bar{r} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i} = \frac{10.6077}{10} = 1.06077$.

The estimate of total cows (unbiased-type ratio estimator) is

$$\begin{aligned} \hat{Y}_{HR} &= \bar{r}X + \frac{n(N-1)}{(n-1)}(\bar{y} - \bar{r}\bar{x}) \\ &= 1.06077 \times 2924 + \frac{10(25-1)}{10-1}(141.7 - 1.06077 \times 134.3) \\ &= 3081.4. \end{aligned}$$

$$\text{We have } s_y^2 = \frac{1}{n-1} \left[\sum y^2 - \frac{(\sum y)^2}{n} \right] = \frac{1}{10-1} \left[228403 - \frac{(1417)^2}{10} \right] = 3068.2333$$

$$s_x^2 = \frac{1}{n-1} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right] = \frac{1}{10-1} \left[203711 - \frac{(1343)^2}{10} \right] = 2594.0111$$

$$s_{yx} = \frac{1}{n-1} \left[\sum xy - \frac{\sum x \sum y}{n} \right] = \frac{1}{10-1} \left[215093 - \frac{1343 \times 1417}{10} \right] = 2754.4333.$$

The unbiased estimator of variance of \hat{Y}_{HR} is

$$\begin{aligned} v(\hat{Y}_{HR}) &\approx \frac{N^2}{n} [s_y^2 + \bar{r}^2 s_x^2 - 2\bar{r} s_{yx}] \\ &= \frac{(25)^2}{10} [3068.2333 + (1.06077)^2 2594.0111 - 2 \times 1.06077 \times 2754.4333] \\ &= 8966.23. \end{aligned}$$

15.6 Conditions Under which Ratio Estimator is a Best Linear Unbiased Estimator

The ratio estimator is not best linear unbiased for every population. Brewer (1963) and Royall (1970) have discussed the situation when ratio estimator becomes best linear unbiased for finite population. According to them, $(y_i, x_i), i = 1, 2, \dots, N$, are N pairs of values which are assumed to be observed from a random sample drawn from a super population. The super population observations are related by $y_i = \beta x_i + \epsilon_i$. Assume that x_i and ϵ_i are independent and $x_i > 0$. For a class in which x_i is fixed, it is assumed that $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \lambda x_i$, where $x_i (i = 1, 2, \dots, N)$ are known value.

Theorem : If the population observations are related by $y_i = \beta x_i + \epsilon_i$, the ratio estimator $\hat{Y}_R = \hat{R}X$ for random or non-random sample is best liner unbiased estimator when the sample is selected on the basis of x_i .

Proof : We have $y_i = \beta x_i + \epsilon_i$.

The condition is $E(\epsilon_i/x_i) = 0$ and $V(\epsilon_i/x_i) = \lambda x_i$.

Now, under repeated sampling, we can write :

$$Y = \beta X + \sum_{i=1}^N \epsilon_i, \text{ where } Y = \sum_{i=1}^N y_i, X = \sum_{i=1}^N x_i.$$

Then $E(Y) = \beta X$.

Let $\hat{Y} = \sum_{i=1}^n l_i y_i$. Then the ratio estimator of population total is $\hat{Y} = \beta \sum_{i=1}^n l_i x_i + \sum_{i=1}^n l_i \epsilon_i$.

Now, if n values of x_i in a sample of n units are fixed, then

$$E(\hat{Y}) = \beta \sum_{i=1}^n l_i x_i \text{ and } V(\hat{Y}) = \lambda \sum_{i=1}^n l_i^2 x_i.$$

It is observed that, if $\sum_{i=1}^n l_i x_i \equiv X$, then \hat{Y} is unbiased in terms of model. Now, under the condition $\sum_{i=1}^n l_i x_i \equiv X$, the variance of \hat{Y} can be minimized using Lagrange's multiplier. We can use

$$Zl_i x_i = Cx_i, \text{ where } l_i = \frac{X}{n\bar{x}} = \text{constant}.$$

Hence, $\hat{Y} = n\bar{y}X/n\bar{x} = \hat{R}X = \hat{Y}_R$.

This \hat{Y} is best linear unbiased estimator. The variance of \hat{Y}_R is $V(\hat{Y}_R) = \frac{\lambda(X - n\bar{x})X}{n\bar{x}}$.

The estimator of λ is $\hat{\lambda} = \frac{1}{n-1} \sum_{i=1}^n \frac{1}{x_i} (y_i - \hat{R}x_i)^2$.

The unbiased estimator of $V(\hat{Y}_R)$ is $v(\hat{Y}_R) = \frac{\hat{\lambda}(X - n\bar{x})X}{n\bar{x}}$.

15.7 Ratio Estimator when \bar{X} is not Known

It is observed that the estimator of population mean depends on \bar{X} and the ratio estimator of population total depends on X . In practice, the value of X may not be known. Therefore, the value of X is to be estimated. For this, double sampling technique is used, where a random sample of size n_1 is selected to estimate the value of X . Finally, a random sample of size n ($n < n_1$) is selected to estimate the parameter of the distribution of study variable. The ratio estimator in such a situation of double sampling is

$$\hat{Y}_{RD} = \frac{\bar{y}}{\bar{x}} \bar{x}_1, \text{ where } \bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i.$$

$\bar{y} = \frac{1}{n} \sum y_i, \bar{x} = \frac{1}{n} \sum x_i$. However, this estimator \hat{Y}_{RD} is not unbiased. Sukhatme and Sukhatme (1970) have derived the relative bias of this estimator. Let us investigate the bias of the estimator. Let $\bar{y} = \bar{Y} + \epsilon, \bar{x} = \bar{X} + \bar{\epsilon}_1$ and $\bar{x}_1 = \bar{X} + \epsilon_2$.

Let us assume that $E(\epsilon) = E(\epsilon_1) = E(\epsilon_2) = 0$. Then

$$\hat{Y}_{RD} = \frac{(\bar{Y} + \epsilon)(\bar{X} + \epsilon_2)}{(\bar{X} + \epsilon_1)} = \bar{Y} \left[1 + \frac{\epsilon}{\bar{Y}} \right] \left[1 + \frac{\epsilon_2}{\bar{X}} \right] \left[1 + \frac{\epsilon_1}{\bar{X}} \right]^{-1}.$$

It is assumed that $\frac{\epsilon_1}{\bar{X}} < 1$, so that $\left[1 + \frac{\epsilon_1}{\bar{X}} \right]^{-1}$ can be expanded.

Now, neglecting the terms with power more than two in expanding the term, we get

$$\hat{Y}_{RD} = \bar{Y} \left[1 + \frac{\epsilon}{\bar{Y}} + \frac{\epsilon_2}{\bar{X}} - \frac{\epsilon_1}{\bar{X}} + \frac{\epsilon\epsilon_2}{\bar{X}\bar{Y}} - \frac{\epsilon\epsilon_1}{\bar{X}\bar{Y}} - \frac{\epsilon_1\epsilon_2}{\bar{X}^2} + \frac{\epsilon_1^2}{\bar{X}^2} \right].$$

But $E[\epsilon\epsilon_2] = \text{Cov}(\bar{y}, \bar{x}_1) = \text{Cov}[E(\bar{y}/n_1), E(\bar{x}_1/n_1)] + E[\text{Cov}(\bar{y}, \bar{x}_1/n_1)]$

$$= \text{Cov}(\bar{y}_1, \bar{x}_1), \quad \text{where } \bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i = \frac{N-n_1}{n_1 N} S_{yx}.$$

Also, we have $E(\epsilon\epsilon_1) = \text{Cov}(\bar{y}, \bar{x}) = \frac{N-n}{nN} S_{yx}$.

$$E(\epsilon_1\epsilon_2) = V(\bar{x}_1) = \frac{N-n_1}{n_1 N} S_x^2 \quad \text{and} \quad E(\epsilon_2^2) = V(\bar{x}) = \frac{N-n}{nN} S_x^2.$$

Now, to find the $E(\hat{Y}_{RD})$ we can replace the above values of covariances. On simplification, we get

$$E[\hat{Y}_{RD}] \approx \left[1 + \left(\frac{1}{n} - \frac{1}{n_1} \right) (C_x^2 - \rho C_x C_y) \right], \quad \text{where } C_x = \frac{S_x}{\bar{X}}, \quad C_y = \frac{S_y}{\bar{Y}}.$$

Therefore, the bias of \hat{Y}_{RD} is $\left(\frac{1}{n} - \frac{1}{n_1} \right) (C_x^2 - \rho C_x C_y)$.

This bias is negligible if the sample size (n) in the second stage is large enough. Moreover, if the regression of y on x is linear and the regression line passes through the origin, then for the approximation up to the first degree the \hat{Y}_{RD} becomes unbiased. In such a case, the variance of the estimator is

$$\begin{aligned} V(\hat{Y}_{RD}) &= E[\hat{Y}_{RD} - \bar{Y}]^2 = \bar{Y}^2 E \left[\frac{\epsilon}{\bar{Y}} + \frac{\epsilon_2}{\bar{X}} - \frac{\epsilon_1}{\bar{X}} \right]^2 \\ &= \bar{Y}^2 E \left[\frac{\epsilon^2}{\bar{Y}^2} + \frac{\epsilon_2^2}{\bar{X}^2} + \frac{\epsilon_1^2}{\bar{X}^2} + \frac{2\epsilon\epsilon_2}{\bar{Y}\bar{X}} - \frac{2\epsilon\epsilon_1}{\bar{Y}\bar{X}} - \frac{2\epsilon_1\epsilon_2}{\bar{X}^2} \right] \\ &= \left(\frac{1}{n} - \frac{1}{n_1} \right) (S_y^2 + R^2 S_x^2 - 2RS_{yx}) + \left(\frac{1}{n_1} - \frac{1}{N} \right) S_y^2. \end{aligned}$$

The variance formula indicates that, if $R^2 S_x^2 - 2RS_{yx} < 0$ or if $2\rho S_y > RS_x$ or, $\rho \frac{C_y}{C_x} > \frac{1}{2}$, then \hat{Y}_{RD} is more precise than \bar{y} .

In practice, the use of \hat{Y}_{RD} depends on the cost of survey at the second stage. Sukhatme and Sukhatme (1970) have also considered the estimator assuming a cost function for the second stage sampling. They have assumed the linear relationship of x and y variables and mentioned that the precision of the estimator under double sampling Scheme will be increased sufficiently,

if $\beta \geq \frac{1}{2}$. If $\beta < \frac{1}{2}$, the precision of the estimator is not increased sufficiently when α is not small enough. Here

$$\alpha = \frac{S_{yx}}{S_y}, \beta = \frac{C_2}{C_1}, S_{y \cdot x} = \frac{1}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2.$$

The cost function is $C_0 = C_1n + C_2n_1$.

Example 15.4 : To estimate the number of dead children in a district in a year, a survey is conducted in the district. For the purpose of the survey 100 villages are randomly selected from the villages in the district. The number of married couples (within the age limit 14-49 years; x) and the number of dead children (y) below age 4 years are recorded from each village. The total number of villages in the district is 1250. The number of couples in the district is not known. Estimate the total number of dead children in the survey year. Also estimate the variance of year estimator.

SL. No. of village	x	y	SL. No. of village	x	y	SL. No. of village	x	y	SL. No. of village	x	y
001	112	5	026	103	2	051	105	3	076	111	2
002	178	2	027	66	1	052	119	5	077	172	10
003	108	4	028	78	0	053	168	4	078	83	4
004	52	2	029	42	2	054	142	3	079	84	3
005	170	3	030	63	1	055	147	4	080	118	2
006	98	2	031	71	2	056	99	3	081	123	3
007	123	1	032	70	3	057	143	5	082	140	3
008	107	4	033	105	2	058	72	2	083	145	6
009	133	2	034	108	1	059	63	1	084	74	2
010	95	0	035	112	2	060	68	0	085	55	1
011	66	1	036	92	1	061	77	2	086	138	4
012	77	2	037	98	2	062	42	2	087	82	2
013	85	2	038	125	2	063	45	2	088	77	3
014	98	3	039	100	2	064	112	3	089	128	4
015	160	2	040	78	1	065	77	0	090	156	5
016	175	4	041	95	2	066	78	2	091	178	4
017	180	5	042	162	7	067	117	2	092	144	3
018	162	3	043	180	6	068	184	7	093	168	7
019	112	2	044	108	3	069	195	12	094	182	10
020	73	1	045	99	2	070	122	4	095	92	0
021	70	1	046	88	2	071	128	3	096	98	3
022	85	0	047	73	2	072	99	2	097	160	5
023	42	1	048	66	2	073	144	4	098	168	4
024	99	2	049	75	1	074	97	2	099	101	3
025	111	3	050	67	1	075	52	0	100	102	2
Total	2771	57		2324	52		2695	77		3079	95

Solution : We have $N = 1250$, $n_1 = 100$, $x_1 = 10869$, x is not known. To estimate the total number of dead children a simple random sample of 50 villages from the first sample is selected. The information of double sample are given below :

SL. No. of village	x	y	SL. No. of village	x	y	SL. No. of village	x	y	SL. No. of village	x	y
014	98	3	090	156	5	004	52	2	035	112	2
061	77	2	027	66	1	097	160	5	052	119	5
081	123	3	015	160	2	047	73	2	092	144	3
037	98	2	075	52	0	018	162	3	068	184	7
053	168	4	095	92	0	051	105	3	082	140	3
011	66	1	016	175	4	033	105	2	083	145	6
046	88	2	058	72	2	019	112	2	084	74	2
042	162	7	072	99	2	079	84	3	024	99	2
021	70	1	062	42	2	029	42	2	002	178	2
064	112	3	077	172	10	059	63	1	092	144	3
100	102	2	061	77	2	069	195	12	020	73	1
040	78	1	032	70	3	063	45	2			
067	117	2	078	83	4	041	95	2			
Total	1359	33		1316	37		1293	41		1412	36

$$n = 50, \bar{x}_1 = \frac{1}{n_1} \sum x = \frac{10869}{100} = 108.69, \bar{x} = \frac{1}{n} \sum x = \frac{5380}{50} = 107.6,$$

$$\bar{y} = \frac{1}{n} \sum y = \frac{147}{50} = 2.94,$$

$$s_y^2 = \frac{1}{n-1} \left[\sum y^2 - \frac{(\sum y)^2}{n} \right] = \frac{1}{50-1} \left[681 - \frac{(147)^2}{50} \right] = 5.078,$$

$$s_x^2 = \frac{1}{n-1} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right] = \frac{1}{50-1} \left[665874 - \frac{(5380)^2}{50} \right] = 1775.224,$$

$$s_{yx} = \frac{1}{n-1} \left[\sum xy - \frac{\sum x \sum y}{n} \right] = \frac{1}{50-1} \left[19031 - \frac{5380 \times 147}{50} \right] = 65.588.$$

The estimate of total dead children in the district is

$$\hat{Y}_{RD} = N \frac{\bar{y}}{\bar{x}} \bar{x}_1 = 1250 \times \frac{2.94}{107.6} \times 108.69 = 3712.$$

Here $\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{2.94}{107.6} = 0.027.$

The estimate of variance of \hat{Y}_{RD} is

$$\begin{aligned} v(\hat{Y}_{RD}) &= N^2 \left[\frac{n_1 - n}{nn_1} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}) + \frac{N - n_1}{Nn_1} s_y^2 \right] \\ &= (1250)^2 \left[\frac{100 - 50}{50 \times 100} (5.078 + (0.027)^2 1775.224 - 2 \times 0.027 \times 65.588) \right. \\ &\quad \left. + \frac{1250 - 100}{1250 \times 100} \times 5.075 \right] \\ &= (1250)^2 [0.0283038 + 0.0467176] = 117220.9375. \end{aligned}$$

The simple estimate of total dead children using second sample is

$$\hat{Y} = N\bar{y} = 1250 \times 2.94 = 3675.$$

The estimated variance of this estimator is

$$v(\hat{Y}) = N^2 \frac{N-n}{nN} s_y^2 = (1250)^2 \frac{1250-50}{50 \times 1250} \times 5.078 = 152340.$$

The relative precision of ratio estimator compared to simple estimator is

$$\frac{v(\hat{Y}) - v(\hat{Y}_{RD})}{v(\hat{Y}_{RD})} \times 100\% = \frac{152340 - 117220.9375}{117220.9375} \times 100 = 29.96\%.$$

15.8 Difference Estimator

It has already been mentioned that if the study variable y and the auxiliary variable x are related by the relation

$$y = \beta x + \epsilon,$$

[i.e., y and x are linearly related and the regression line passes through the origin] then the ratio estimator is best linear unbiased estimator. But the relation of y and x may not exist for all populations. Let us assume that y and x linearly related but the regression line does not pass through the origin, i.e.,

$$y = \alpha + \beta x + \epsilon, \quad \text{where } E(\epsilon) = 0.$$

Then $E(y) = \alpha + \beta x$. In such a situation the difference $y - \beta x$ can be used to find an estimator of population parameter. The estimator which is found out using the difference $y - \beta x$ is called difference estimator.

The difference estimator of population mean is

$$\hat{Y}_D = (\bar{y} - \beta \bar{x}) + \beta \bar{X},$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i$ and β is a constant.

Theorem : In case of simple random sampling without replacement the difference estimator of population mean is unbiased and the variance of this estimator is

$$V(\hat{Y}_D) = \frac{1-f}{n} [S_y^2 + \beta^2 S_x^2 - 2\beta S_{yx}].$$

Proof : We have $\hat{Y}_D = (\bar{y} - \beta \bar{x}) + \beta \bar{X}$

$$E[\hat{Y}_D] = E(\bar{y} - \beta \bar{x}) + E(\beta \bar{X}) = E(\bar{y}) - \beta E(\bar{x}) + \beta \bar{X} = \bar{Y} - \beta \bar{X} + \beta \bar{X} = \bar{Y}.$$

Hence, \hat{Y}_D is unbiased estimator of \bar{Y} .

$$\begin{aligned} V(\hat{Y}_D) &= E[\hat{Y}_D - \bar{Y}]^2 = E[(\bar{y} - \beta \bar{x}) + \beta \bar{X} - \bar{Y}]^2 = E[(\bar{y} - \bar{Y}) - \beta(\bar{x} - \bar{X})]^2 \\ &= E(\bar{y} - \bar{Y})^2 + \beta^2 E(\bar{x} - \bar{X})^2 - 2\beta E(\bar{x} - \bar{X})(\bar{y} - \bar{Y}) = \frac{1-f}{n} [S_y^2 + \beta^2 S_x^2 - 2\beta S_{yx}]. \end{aligned}$$

Here β is a known value. Des Raj (1968) has shown that

$$\beta = \frac{S_{yx}}{S_x^2} = \text{population regression coefficient.}$$

For known β , the variance of \hat{Y}_D is minimum, where the minimum variance is

$$V(\hat{Y}_D)_{\min} = \frac{1-f}{n}(1-\rho^2)S_y^2,$$

where ρ is the population correlation coefficient.

For known value of β the unbiased estimator of variance of \bar{Y}_D is

$$v(\hat{Y}_D) = \frac{1-f}{n}[s_y^2 + \beta^2 s_x^2 - 2\beta s_{yx}].$$

The relative efficiency of difference estimator depends on the value of β . If x_i values ($i = 1, 2, \dots, N$) are the y_i values recorded in any previous survey, then the value of β can be assumed 1. In that case, \hat{Y}_D becomes unbiased and its variance becomes minimum. However, if $\beta = \frac{y}{x}$, then \hat{Y}_D and \hat{Y}_R are the similar estimates. Des Raj has shown that, if β lies within zero and twice of the population regression coefficient, then \hat{Y}_D is more efficient than \bar{y} , otherwise \bar{y} is more efficient.

Example 15.5 : In a district there are 1250 villages. One hundred villages are randomly selected to estimate the number of married couples who have adopted family planning program. The survey is conducted in two consecutive years. The number of adopter couples in the district during first year of survey is estimated as $X = 90000$. The number of adopter couples in the first year (x) and in the second year (y) in the selected villages are shown below :

SL. No.	x	y	SL. No.	x	y	SL. No.	x	y	SL. No.	x	y
1	70	72	29	60	72	57	50	52	85	110	115
2	110	120	30	65	46	58	68	70	86	97	92
3	65	68	31	36	30	59	71	73	87	71	65
4	30	33	32	55	62	60	121	124	88	92	101
5	102	110	33	85	98	61	108	92	89	95	97
6	60	68	34	92	97	62	95	98	90	42	38
7	71	74	35	52	28	63	88	101	91	28	32
8	65	62	36	45	49	64	77	72	92	26	30
9	72	78	37	64	60	65	62	70	93	44	52
10	67	70	38	12	28	66	42	48	94	53	62
11	40	45	39	42	45	67	32	31	95	72	88
12	30	32	40	28	40	68	44	44	96	128	126
13	92	97	41	49	42	69	28	28	97	162	160
14	52	28	42	36	38	70	42	47	98	178	182
15	100	112	43	53	55	71	52	50	99	66	70
16	85	90	44	48	38	72	55	62	100	72	78
17	108	115	45	120	128	73	111	109			
18	96	90	46	88	92	74	99	95			
19	62	54	47	97	95	75	82	78			
20	42	45	48	51	61	76	97	96			
21	43	44	49	55	65	77	62	61			
22	60	72	50	37	48	78	58	55			
23	95	94	51	41	44	79	77	72			
24	108	108	52	45	48	80	50	55			
25	112	109	53	121	128	81	48	47			
26	108	112	54	178	195	82	23	31			
27	90	85	55	66	72	83	15	18			
28	120	138	56	92	90	84	87	90			
Total	2093	2190		1813	1894		1844	1869		1336	1388

Estimate the number of adopter couples in the following year. Also estimate the standard error of your estimator. Compare your estimator with simple estimator.

Solution : $N = 1250$, $n = 100$, $X = 90000$, $x = 7086$, $y = 7341$. $\bar{x} = 70.86$, $\bar{y} = 73.41$. Here y_i values are the values of x_i in the following year. Hence, $\beta = 1$ can be assumed. Under this assumption the difference estimate of total number of adopter couples in the district is

$$\hat{Y}_D = N\hat{Y}_D = N[\bar{y} - \beta\bar{x} + \beta\bar{X}] = 1250[73.41 - 1 \times 70.86 + 72] = 93188.$$

$$s_x^2 = \frac{1}{n-1} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right] = \frac{1}{100-1} \left[612085 - \frac{(7086)^2}{100} \right] = 1110.8186.$$

$$s_y^2 = \frac{1}{n-1} \left[\sum y^2 - \frac{(\sum y)^2}{n} \right] = \frac{1}{100-1} \left[654605 - \frac{(7341)^2}{100} \right] = 1168.709.$$

$$s_{yx} = \frac{1}{n-1} \left[\sum xy - \frac{\sum x \sum y}{n} \right] = \frac{1}{100-1} \left[630719 - \frac{7086 \times 7341}{100} \right] = 1116.523.$$

The estimate of variance of \hat{Y}_D is

$$\begin{aligned} v(\hat{Y}_D) &= \frac{N^2(1-f)}{n} [s_y^2 + \beta^2 s_x^2 - 2\beta s_{yx}] \\ &= \frac{(1250)^2(1-0.08)}{100} [1168.709 + (1)^2 1110.8186 - 2 \times 1 \times 1116.523] \\ &= 668173. \end{aligned}$$

$$\text{s.e.}(\hat{Y}_D) = \sqrt{v(\hat{Y}_D)} = 817.42.$$

The simple estimate of total adopter couples is

$$\hat{Y} = N\bar{y} = 1250 \times 73.41 = 91763.$$

$$v(\hat{Y}) = \frac{N^2(1-f)}{n} s_y^2 = 16800191.88.$$

The relative precision of \hat{Y}_D compared to \hat{Y} is $\frac{v(\hat{Y}) - v(\hat{Y}_D)}{v(\hat{Y}_D)} 100\% = 2414.3\%$.

15.9 Ratio Estimator in Case of Stratified Sampling

Let the population units N be divided into L strata such that the size of h -th stratum ($h = 1, 2, \dots, L$) is N_h ($N = \sum_h N_h$). Let y_{hi} be the value of study variable y recorded from i -th unit in h -th stratum ($i = 1, 2, \dots, N_h$). The corresponding value of an auxiliary variable x is x_{hi} . The problem is to find a ratio estimator of population total Y .

There are two methods to estimate the population parameter. These are (a) separate ratio estimator, (b) combined ratio estimator.

(a) **Separate Ratio Estimator :** Let us discuss the separate ratio estimator of population total. The estimator is

$$\hat{Y}_{RS} = \sum_{h=1}^L \frac{y_h}{x_h} X_h = \sum_{h=1}^L \frac{\bar{y}_h}{\bar{x}_h} X_h,$$

where X_h = total of values of x_{hi} of h -th stratum,

x_h = total of values of x_{hi} of h -th stratum in sample,

y_h = total of values of y_{hi} of h -th stratum in sample.

If the sample is selected under SRSWOR from each stratum, then the variance of \hat{Y}_{RS} is

$$V(\hat{Y}_{RS}) = \sum_{h=1}^L \frac{N_h^2(1-f_h)}{n_h} [S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h S_{hyx}],$$

where $R_h = \frac{Y_h}{X_h}$, $f_h = \frac{n_h}{N_h}$,

Y_h = total of value of y_{hi} of h -th stratum,

n_h = sample of size from h -th stratum,

$$n = \sum_{h=1}^L n_h.$$

Since sample is selected under SRSWOR from h -th stratum, the variance of \hat{Y}_{HR} is

$$V(\hat{Y}_{HR}) = \frac{N_h^2(1-f_h)}{n_h} [S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h S_{hyx}],$$

where $R_h = \frac{Y_h}{X_h}$, $S_{hy}^2 = \frac{1}{N_h - 1} \sum (y_{hi} - \bar{Y}_h)^2$, $S_{hx}^2 = \frac{1}{N_h - 1} \sum (x_{hi} - \bar{X}_h)^2$,

$$S_{hyx} = \frac{1}{N_h - 1} \sum (x_{hi} - \bar{X}_h)(y_{hi} - \bar{Y}_h).$$

The separate ratio estimator is

$$\hat{Y}_{RS} = \sum \frac{\bar{y}_h}{\bar{x}_h} X_h = \sum \hat{Y}_{hR}, \text{ where } \hat{Y}_{hR} = \frac{\bar{y}_h}{\bar{x}_h} X_h.$$

Since $\hat{Y}_{RS} = \sum_h \hat{Y}_{hR}$, $V(\hat{Y}_{RS}) = \sum_h V(\hat{Y}_{hR})$,

[∵ ratio estimators from different strata are independent].

$$\therefore V(\hat{Y}_{RS}) = \sum_h \frac{N_h^2(1-f_h)}{n_h} [S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h S_{hyx}].$$

The estimator of $V(\hat{Y}_{RS})$ is

$$v(\hat{Y}_{RS}) = \sum_h \frac{N_h^2(1-f_h)}{n_h} [s_{hy}^2 + \hat{R}_h^2 s_{hx}^2 - 2\hat{R}_h s_{hyx}], \text{ where } \hat{R}_h = \frac{y_h}{x_h} = \frac{\bar{y}_h}{\bar{x}_h}.$$

Since ratio estimator is biased, the \hat{Y}_{hR} is biased. Therefore, we can write :

$$\frac{|\text{bias}(\hat{Y}_{hR})|}{\sigma_{\hat{Y}_{hR}}} \leq \text{C.V.}(\bar{x}_h).$$

If the bias of each stratum is of same direction, the bias of \hat{Y}_{RS} is approximately L times the bias of \hat{Y}_{hR} . But the standard error of \hat{Y}_{RS} is \sqrt{L} times of standard error of \hat{Y}_{hR} . Hence,

$$\frac{|\text{bias} \hat{Y}_{RS}|}{\sigma_{\hat{Y}_{hR}}} \text{ is of order } \sqrt{L} \text{ C.V.}(\bar{x}_h).$$

(b) **Combined Ratio Estimator** : In case of separate ratio estimator, it is mentioned that large size sample is to be selected from each stratum. But it is not always possible to select large sample from all strata in all sample survey. In such a case combined ratio estimator can be found out. Hansen, Hurwitz and Gurney (1946) have mentioned such combined estimator. According to them, if n is big and if SRS is selected from all strata, then the combined estimator of population total is \hat{Y}_{RC} , where

$$\hat{Y}_{RC} = \frac{\hat{Y}_{st}}{\hat{X}_{st}} X = \frac{\bar{y}_{st}}{\bar{x}_{st}} X.$$

Here $\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h$, $\bar{x}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{x}_h$

and $X = \sum_{h=1}^L \sum_{i=1}^{N_h} x_{hi}$ and the value of X is assumed to be known.

For this estimator the stratum total X_h may not be known, whereas in separate estimator X_h must be known.

The bias of \hat{Y}_{RC} can be obtained using the formula suggested by Hartley and Ross (1954), where

$$\text{Cov} (\hat{R}_c, \bar{x}_{st}) = E \left(\frac{\bar{y}_{st}}{\bar{x}_{st}}, \bar{x}_{st} \right) - E(\hat{R}_c)E(\bar{x}_{st}).$$

Here $\hat{R}_c = \frac{\bar{y}_{st}}{\bar{x}_{st}}$.

$\therefore \text{Cov} (\hat{R}_c, \bar{x}_{st}) = \bar{Y}_1 - \bar{X}E(\hat{R}_c)$

$\therefore E(\hat{R}_c) = R - \frac{1}{\bar{X}} \text{Cov} (\hat{R}_c, \bar{x}_{st})$

$\therefore \text{Bias} (\hat{R}_c) = E(\hat{R}_c) - R = -\frac{1}{\bar{X}} \text{Cov} (\hat{R}_c, \bar{x}_{st}) = -\frac{\rho_{\hat{R}_c, \bar{x}_{st}} \sigma_{\bar{x}_{st}} \sigma_{\hat{R}_c}}{\bar{X}}$

$\therefore \frac{\text{Bias} (\hat{R}_c)}{\sigma_{\hat{R}_c}} = \frac{\rho_{\hat{R}_c, \bar{x}_{st}} \sigma_{\bar{x}_{st}}}{\bar{X}} < \text{C.V.} (\bar{x}_{st}).$

It is clear that, if $\text{C.V.} (\bar{x}_{st}) < 0.1$, then the bias of \hat{R}_c and \hat{Y}_{RC} will be less than their standard error.

Theorem : If SRSWOR scheme is applied to select sample from all strata and if n is big, then the variance of \hat{Y}_{RC} is given by

$$V(\hat{Y}_{RC}) = \sum_{h=1}^L \frac{N_h^2(1-f_h)}{n_h} [S_{hy}^2 + R^2 S_{hx}^2 - 2RS_{h_{yx}}].$$

Proof : It is known that $\hat{Y}_{RC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} X$ and $Y = N\bar{Y} - NR\bar{X}$.

Then $\hat{Y}_{RC} - Y = \frac{\bar{y}_{st}}{\bar{x}_{st}} X - NR\bar{X} = \frac{N\bar{X}}{\bar{x}_{st}} (\bar{y}_{st} - R\bar{x}_{st}) = N(\bar{y}_{st} - R\bar{x}_{st})$.

Let $u_{hi} = y_{hi} - Rx_{hi}$ ($i = 1, 2, \dots, N_h$)

$\bar{U} = \bar{Y} - R\bar{X} = 0$ and $\bar{u}_{st} = \bar{y}_{st} - R\bar{x}_{st}$, $N\bar{u}_{st} = N(\bar{y}_{st} - R\bar{x}_{st})$.

Therefore, $\hat{Y}_{RC} - Y = N\bar{u}_{st}$, $\bar{u}_{st} = \frac{1}{N} \sum N_h \bar{u}_h$, $\bar{u}_h = \frac{1}{n_h} \sum_i u_{hi}$.

$$\begin{aligned} V(\hat{Y}_{RC}) &= E(\hat{Y}_{RC} - Y)^2 = E[N\bar{u}_{st} - N\bar{U}]^2 = N^2 V(\bar{u}_{st}) \\ &= \sum_{h=1}^L \frac{N_h(N_h - n_h)}{n_h} S_{hu}^2, \end{aligned}$$

where
$$\begin{aligned} S_{hu}^2 &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (u_{hi} - \bar{U})^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} [(y_{hi} - \bar{Y}_h) - R(x_{hi} - \bar{X}_h)]^2 \\ &= S_{hy}^2 + R^2 S_{hx}^2 - 2RS_{hyx}. \end{aligned}$$

$$\begin{aligned} \therefore V(\hat{Y}_{RC}) &= \sum_{i=1}^L \frac{N_h(N_h - n_h)}{n_h} [S_{hy}^2 + R^2 S_{hx}^2 - 2RS_{hyx}] \\ &= \sum_{h=1}^{N_h} \frac{N_h^2(1 - f_h)}{n_h} [S_{hy}^2 + R^2 S_{hx}^2 - 2RS_{hyx}]. \end{aligned}$$

The unbiased estimator of this variance is

$$v(\hat{Y}_{RC}) = \sum_{h=1}^h \frac{N_h^2(1 - f_h)}{n_h} [s_{hy}^2 + \hat{R}^2 s_{hx}^2 - 2\hat{R}s_{hyx}].$$

Now, we can compare the separate and combined ratio estimator. We have

$$\begin{aligned} V(\hat{Y}_{RC}) - V(\hat{Y}_{RS}) &= \sum_{h=1}^{N_h} \frac{N_h^2(1 - f_h)}{n_h} [(R^2 - R_h^2)S_{hx}^2 - 2(R - R_h)S_{hyx}] \\ &= \sum \frac{N_h^2(1 - f_h)}{n_h} [(R^2 - R_h^2)S_{yx}^2 + 2(R_h - R)(\rho_h S_{hy} S_{hx} - R_h^2 S_{hx}^2)]. \end{aligned}$$

The last term in right-hand side of the above result is normally small. The value will be zero, if y_{hi} and x_{hi} of h -th stratum are linearly related and if the straight line of y_{hi} on x_{hi} passes through the origin. Thus, it can be said that, if R_h of h -th stratum is not constant, then the separate ratio estimator is more efficient than the combined ratio estimator. But the bias of combined ratio estimator is less than that of separate ratio estimator.

Thus, to use any of the estimator the following rules can be followed :

- (1) The combined estimator is to be preferred if the value of R_h does not vary too much and if small size sample is selected from all strata. Separate ratio estimator is preferred if strata can be rearranged and if large size sample is selected from all strata.
- (2) If X_h is known, then separate ratio estimator is preferred.

15.10 Optimum Allocation in Case of Ratio Estimator

Let the cost function be

$$C = C_0 + \sum_{h=1}^L C_h n_h.$$

The value of n_h is to be selected on the basis of the given cost function so that $V(\hat{Y}_{RS})$ is minimum.

$$\begin{aligned} \text{Let } \phi &= V(\hat{Y}_{RS}) + \lambda[C_0 + \sum_h C_h n_h - C] \\ &= \sum_{h=1}^L \frac{N_h^2(1-f_h)}{n_h} [S_{hy}^2 + R_h^2 S_{Hx}^2 - 2R_h S_{hyx}] + \lambda[C_0 + \sum C_h n_h - C] \\ &= \sum_{h=1}^L \frac{N_h^2 \left(1 - \frac{n_h}{N_h}\right)}{n_h} S_{hd}^2 + \lambda[C_0 + \sum C_h n_h - C]. \end{aligned}$$

$$\text{Here } S_{hd}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} d_{hi}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - R_h x_{hi})^2.$$

$$\text{Now } \frac{\partial \phi}{\partial n_h} = -\frac{N_h^2 S_{hd}^2}{n_h} + \lambda C_h.$$

Putting $\frac{\partial \phi}{\partial n_h} = 0$, we get

$$\lambda C_h = \frac{N_h^2 S_{hd}^2}{n_h^2} \quad \text{or, } n_h = \frac{N_h S_{hd}}{\sqrt{\lambda C_h}}.$$

Also, we get

$$\sqrt{\lambda} = \frac{1}{n} \sum_{h=1}^L \frac{N_h S_{hd}}{\sqrt{C_h}} \quad \text{or, } n_h = \frac{1}{n} \frac{N_h S_{hd} / \sqrt{C_h}}{\sum \frac{N_h S_{hd}}{\sqrt{C_h}}} \quad \text{or, } n_h \propto \frac{N_h S_{hd}}{\sqrt{C_h}}.$$

Cochran (1977) has mentioned that, when ratio estimator is best linear unbiased, S_{dh} is proportional to $\sqrt{\bar{X}_h}$. Then

$$n_h \propto N_h \sqrt{\bar{X}_h} / \sqrt{C_h}.$$

In some cases $d_{hi} \propto \bar{X}_h^2$, then $n_h \propto N_h \bar{X}_h / \sqrt{C_h}$.

15.11 Ratio Estimator to Estimate the Parameter Related to Qualitative Variable

Let $y_i = 1$, if i -th unit possesses the characteristic under study and $y_i = 0$, otherwise ($i = 1, 2, \dots, N$). Total number of units in the population possessing the characteristic is $\sum_{i=1}^N y_i = A$. Therefore, $(N - A)$ units do not possess the characteristic. The population proportion of the units possessing the characteristic is

$$P = \frac{A}{N} = \frac{1}{N} \sum_{i=1}^N y_i = \bar{Y}.$$

Also, we have

$$S_y^2 = \frac{1}{N-1} \left[\sum_{i=1}^N y_i^2 - \frac{(\sum_{i=1}^N y_i)^2}{N} \right] = \frac{1}{N-1} \left[A - \frac{A^2}{N} \right] = \frac{NPQ}{N-1},$$

where $Q = 1 - P = 1 - \frac{A}{N}$.

Let $x_i = 1$, if i -th unit does not possess the characteristic under study, and $x_i = 0$, otherwise.

Then
$$\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{N-A}{N} = 1 - P = Q.$$

$$S_x^2 = \frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \frac{(\sum x_i)^2}{N} \right] = \frac{1}{N-1} \left[N-A - \frac{(N-A)^2}{N} \right] = \frac{NPQ}{N-1}.$$

$$S_{xy} = \frac{1}{N-1} \left[\sum_{i=1}^N x_i y_i - \frac{\sum x_i \sum y_i}{N} \right] = \frac{1}{N-1} \left[Q - \frac{A(N-A)}{N} \right] = -\frac{NPQ}{N-1}.$$

The population ratio is

$$R = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i} = \frac{A}{N-A} = \frac{P}{Q}.$$

Let the estimate of R be

$$\hat{R} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{a}{n-a} = \frac{p}{q}, \quad \text{where } p = \frac{a}{n}, \quad q = 1 - p.$$

Therefore,
$$V(\hat{R}) = \frac{1-f}{nQ^2} \left[\frac{NPQ}{N-1} + \frac{R^2 NPQ}{N-1} + \frac{2RN PQ}{N-1} \right] = \frac{NP(1-f)}{n(N-1)Q} [1 + R^2 - 2R]$$

$$= \frac{NP(1-f)}{n(N-1)Q} (1-R)^2.$$

Here
$$E(\hat{R}) = R \left[1 + \frac{1-f}{n} \left(\frac{S_x^2}{\bar{X}^2} - \frac{S_{yx}}{R\bar{X}^2} \right) \right] = R \left[1 + \frac{1-f}{n} (C_x^2 - \rho C_y C_x) \right]$$

$$= \frac{P}{Q} \left[1 + \frac{N-n}{N-1} \frac{1}{nQ} \right].$$

Thus, it may be concluded that, if N is large enough, the relative bias of \hat{R} is $\frac{1}{nQ}$. The bias is reduced, if n increases.

15.12 Regression Estimator

The ratio estimator is defined using the information of auxiliary variable. Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be n pairs of values recorded from an investigation on randomly selected n sampling units from a population of size N . Consider that the study variable is Y and its auxiliary variable is X . Since the auxiliary variable X is one which is measured from the sampling units in different occasions or it is a related variable to the study variable Y , there may be a linear relationship of these two variables. Let us assume such a linear relationship of Y and X as

$$y = \alpha + \beta x + \epsilon, \quad (1)$$

such that $E(y) = \alpha + \beta x$ ($\because E(\epsilon) = 0$).

Using this relationship of X and Y , we can estimate the population mean \bar{Y} . Such an estimator of \bar{Y} using the regression estimator b of β is known as *regression estimator*.

To define ratio estimator, it is assumed that there is a linear relationship of X and Y . It is observed that the ratio estimator is best if the regression line of y on x passes through the origin. But, the regression line may not pass through the origin. In such a situation the estimator is derived using the regression model (1). In the regression model β is the regression parameter. We can use difference estimator and $V(\bar{y}_D)$ becomes minimum, if β is known. Here \bar{y}_D is the difference estimator of \bar{Y} .

In practice, β is not known. The sample estimator of β is given by

$$b = \hat{\beta} = \frac{\sum^n (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}.$$

Then the regression estimator of population mean \bar{Y} is given by

$$\bar{y}_{lr} = \bar{y} - b(\bar{x} - \bar{X}),$$

where \bar{y}_{lr} is the linear regression estimator of \bar{Y} . The linear regression estimator of population total is $Y_{lr} = N\bar{y}_{lr}$.

The regression estimator \bar{y}_{lr} is a consistent estimator since $n \rightarrow \infty, \bar{x} \rightarrow \bar{X}$ and in such a situation $\bar{y}_{lr} \rightarrow \bar{Y}$. But \bar{y}_{lr} is not an unbiased estimator of \bar{Y} . The value of \bar{y}_{lr} takes different shapes depending on value of b . Let $b = 0$, then $\bar{y}_{lr} = \bar{y}$, which takes the form of simple estimator and this estimator is consistent and unbiased. If $b = 1$, then $\bar{y}_{lr} = \bar{y} + (\bar{X} - \bar{x})$. This estimator is also consistent and unbiased. Again, let us consider that $b = \bar{y}/\bar{x}$. Then

$$\bar{y}_{lr} = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \bar{y} + \frac{\bar{y}}{\bar{x}}\bar{X} - \bar{y} = R\bar{X},$$

where $R = \bar{y}/\bar{x}$. The estimator \bar{y}_{lr} now transforms to ratio estimator. This estimator is a biased one. Let us now investigate the bias of general regression estimator.

15.13 Bias of Regression Estimator

The regression estimator is $\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$. Assume that

$$\bar{x} = \bar{X} + \epsilon_1, \quad s_{xy} = S_{xy} + \epsilon_2, \quad s_x^2 = S_x^2 + \epsilon_3.$$

$$E(\epsilon_1) = E(\epsilon_2) = E(\epsilon_3) = 0.$$

Then
$$\bar{y}_{lr} = \bar{y} + \frac{S_{xy} + \epsilon_2}{S_x^2 + \epsilon_3} [\bar{X} - \bar{X} + \epsilon_1]$$

$$= \bar{y} - \beta \epsilon_1 \left[1 + \frac{\epsilon_2}{S_{xy}} \right] \left[1 + \frac{\epsilon_3}{S_x^2} \right]^{-1}, \quad \text{where } \beta = \frac{S_{xy}}{S_x^2}.$$

Let $\left| \frac{\epsilon_3}{S_x^2} \right| < 1$.

This restriction indicates that the expansion of $\left(1 + \frac{\epsilon_3}{S_x^2} \right)^{-1}$ is possible.

Neglecting the terms involving powers more than equal to 2 in ϵ_3 , we get

$$\bar{y}_{lr} \approx \bar{y} - \beta \left[\frac{\epsilon_1 \epsilon_2}{S_{xy}} - \frac{\epsilon_1 \epsilon_3}{S_x^2} \right].$$

$$\therefore E[\bar{y}_{lr}] \approx E(\bar{y}) - \beta \left[\frac{E(\epsilon_1 \epsilon_2)}{S_{xy}} - \frac{E(\epsilon_1 \epsilon_3)}{S_x^2} \right] = \bar{Y} - \beta \left[\frac{\text{Cov}(\bar{x}, s_{xy})}{S_{xy}} - \frac{\text{Cov}(\bar{x}, s_x^2)}{S_x^2} \right].$$

$$\therefore E(\bar{y}_{lr}) - \bar{Y} = -\beta \left[\frac{\text{Cov}(\bar{x}, s_{xy})}{S_{xy}} - \frac{\text{Cov}(\bar{x}, s_x^2)}{S_x^2} \right].$$

Hence, the regression estimator is biased one and the bias is given by

$$\text{Bias}(\bar{y}_{lr}) = -\beta \left[\frac{\text{Cov}(\bar{x}, s_{xy})}{S_{xy}} - \frac{\text{Cov}(\bar{x}, s_x^2)}{S_x^2} \right].$$

Sukhatme and Sukhatme (1970) have shown that

$$\text{Bias}(\bar{y}_{lr}) = -\frac{N-n}{Nn} \beta \left[\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{20}}{S_x^2} \right],$$

where $\mu_{21} = E(\bar{x} - \bar{X})^2(\bar{y} - \bar{Y})$ and $\mu_{20} = E(\bar{x} - \bar{X})^2$.

The amount of bias of regression estimator can also be shown alternatively. For this, let

$$\bar{y} = \bar{Y}(1 + e), \quad \bar{x} = \bar{X}(1 + e_1) \quad \text{and} \quad b = \beta(1 + e_2)$$

with assumption $E(e) = E(e_1) = E(e_2) = 0$. Now, putting the values of \bar{y} , \bar{x} and \bar{X} in regression estimator, we get

$$\bar{y}_{lr} = \bar{Y} + e\bar{Y} + \beta\bar{X}e_1 - \beta\bar{X}e_1e_2.$$

$$\therefore E(\bar{y}_{lr}) = \bar{Y} - \beta\bar{X}E(e_1e_2) = \bar{Y} - E(\bar{x} - \bar{X})(b - \beta) = \bar{Y} - \text{Cov}(\bar{x}, b).$$

$$\therefore \text{Bias}(\bar{y}_{lr}) = -\text{Cov}(\bar{x}, b).$$

This bias will be negligible, if sample size becomes large. Moreover, for bivariate normal distribution $\text{Cov}(\bar{x}, b) = 0$ and in that case, \bar{y}_{lr} becomes unbiased estimator of \bar{Y} .

The latter bias in \bar{y}_{lr} is estimated by

$$\text{Bias}(\bar{y}_{lr}) \approx \frac{\sum_{i=1}^k (\bar{x}_i - \bar{y}_i)(b_i - b)}{k-1},$$

where k is the number of sub-samples selected from sample of size n . The i -th sub-sample provides \bar{x}_i , \bar{y}_i and b_i ($i = 1, 2, \dots, k$). Using this estimator of bias, one can find unbiased regression estimator of population mean and hence, population total.

15.14 Variance of Regression Estimator

It has already been mentioned that the regression estimator becomes unbiased if b is any known constant, say b_0 . Then $\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x})$. The variance of this estimator is given by

$$V(\bar{y}_{lr}) = \frac{1-f}{n} [S_y^2 - 2b_0S_{yx} + b_0^2S_x^2],$$

$$\text{where } S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2, \quad S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2.$$

$$S_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})(y_i - \bar{Y}).$$

Proof : Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be n pairs of values observed from n sample points selected at random from a population of size N , where x_i values are the values of some auxiliary variable related to y_i values ($i = 1, 2, \dots, n$). Let b_0 be a known value. Then the linear regression estimator of population mean \bar{Y} is

$$\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x}).$$

Let us define $u_i = y_i - b_0(x_i - \bar{X})$, $i = 1, 2, \dots, n$.

Then $\bar{u} = \bar{y} - b_0(\bar{x} - \bar{X}) = \bar{y} + b_0(\bar{X} - \bar{x}) = \bar{y}_{lr}$.

$$E(\bar{y}_{lr}) = E(\bar{y}) + b_0E(\bar{x} - \bar{X}) = \bar{Y}.$$

Hence, \bar{y}_{lr} is unbiased estimator of \bar{Y} .

Now, $V(\bar{y}_{lr}) = \frac{1-f}{n} S_u^2$ (\because u_i values are observed from simple random sampling)

$$\text{Here } S_u^2 = \frac{1}{N-1} \sum_{i=1}^N (u_i - \bar{U})^2,$$

$$\begin{aligned} \bar{U} &= \frac{1}{N} \sum_{i=1}^N u_i = \frac{1}{N} \sum_{i=1}^N [y_i - b_0(x_i - \bar{X})] \\ &= \frac{1}{N-1} \sum_{i=1}^N [y_i - b_0(x_i - \bar{X}) - \bar{Y} - b_0(\bar{X} - \bar{X})]^2 \\ &= \frac{1}{N-1} \sum (y_i - \bar{Y})^2 + \frac{b_0^2}{N-1} \sum (x_i - \bar{X})^2 - \frac{2b_0}{N-1} \sum (y_i - \bar{Y})(x_i - \bar{X}) \\ &= S_y^2 + b_0^2 S_x^2 - 2b_0 S_{yx}. \end{aligned}$$

$$\therefore V(\bar{y}_{lr}) = \frac{1-f}{n} [S_y^2 + b_0^2 S_x^2 - 2b_0 S_{yx}].$$

Corollary : In simple random sampling if \bar{y}_{lr} is the regression estimator of population mean, then the estimator of variance of \bar{y}_{lr} is given by

$$v(\bar{y}_{lr}) = \frac{1-f}{n} [s_y^2 + b_0^2 s_x^2 - 2b_0 s_{yx}],$$

where $\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x})$. Here b_0 is a known constant.

Proof : Since $u_i = y_i - b_0(x_i - \bar{X})$, $i = 1, 2, \dots, n$ are observed from simple random sample of size n from a population of size N ,

$$s_u^2 = \frac{1}{n-1} \sum_{i=1}^n (u_i - \bar{u})^2 \text{ is an unbiased estimator of } S_u^2 = \frac{1}{N-1} \sum_{i=1}^N (u_i - \bar{U})^2.$$

$$\text{Here } s_u^2 = \frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{y}) - b_0(x_i - \bar{x})]^2 = s_y^2 + b_0^2 s_x^2 - 2b_0 s_{yx}$$

$$E(s_u^2) = S_u^2 = E[s_y^2 + b_0^2 s_x^2 - 2b_0 s_{yx}] = S_y^2 + b_0^2 S_x^2 - 2b_0 S_{yx}.$$

$$\therefore E[v(\bar{y}_{lr})] = V(\bar{y}_{lr}) = \frac{1-f}{n} [S_y^2 + b_0^2 S_x^2 - 2b_0 S_{yx}].$$

We have discussed the formula for $V(\bar{y}_{lr})$, when β is known. In practice, β is not known. It is estimated from sample observations. In such a case we need to estimate β so that $V(\bar{y}_{lr})$ becomes minimum.

Theorem : In simple random sampling the variance of regression estimator of population mean becomes minimum, if

$$b_0 = \frac{S_{yx}}{S_x^2} = \frac{\sum_{i=1}^N (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^N (x_i - \bar{X})^2}, \quad \text{where } \bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x}).$$

Proof : Given $b_0 = \frac{S_{yx}}{S_x^2} = B =$ population regression coefficient of y on x . If $b_0 \neq B$, let $b_0 = B + d$, where d is a positive constant.

$$\begin{aligned} \text{Then } V(\bar{y}_{lr}) &= \frac{1-f}{n} [S_y^2 + b_0^2 S_x^2 - 2b_0 S_{yx}] \\ &= \frac{1-f}{n} [S_y^2 + (B+d)^2 S_x^2 - 2(B+d) S_{yx}] \\ &= \frac{1-f}{n} \left[S_y^2 + \left(\frac{S_{yx}}{S_x^2} + d \right)^2 S_x^2 - 2 \left(\frac{S_{yx}}{S_x^2} + d \right) S_{yx} \right] \\ &= \frac{1-f}{n} \left[S_y^2 + S_x^2 \left(\frac{S_{yx}^2}{S_x^4} + d^2 + 2d \frac{S_{yx}}{S_x^2} \right) - \frac{2S_{yx}^2}{S_x^2} - 2d S_{yx} \right] \\ &= \frac{1-f}{n} \left[S_y^2 + d^2 S_x^2 - \frac{S_{yx}^2}{S_x^2} \right]. \end{aligned}$$

At this stage $V(\bar{y}_{lr})$ will be minimum, if $d = 0$. Then

$$V(\bar{y}_{lr}) = \frac{1-f}{n} \left[S_y^2 - \frac{S_{yx}^2}{S_x^2} \right].$$

Thus, $V(\bar{y}_{lr})$ becomes minimum if $b_0 = B = \frac{S_{yx}}{S_x^2}$. The minimum value of $V(\bar{y}_{lr})$ is

$$V(\bar{y}_{lr})_{\min} = \frac{1-f}{n} \left[S_y^2 - \frac{S_{yx}^2}{S_x^2} \right] = \frac{1-f}{n} S_y^2 [1 - \rho^2],$$

$$\therefore \rho = \frac{S_{xy}}{S_x S_y}.$$

Corollary : The estimator of $V(\bar{y}_{lr})$ for large sample size is

$$v(\bar{y}_{lr}) = \frac{1-f}{n} (1 - r^2) s_y^2,$$

$$\text{where } s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2, \quad r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

The linear regression estimator $\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x})$ is considered assuming population regression coefficient known. In practice, the population regression coefficient B is not known and it is estimated from sample observations, where the sample estimator is

$$b = \hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

The regression estimator of population mean is then $\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$.

Let us now investigate $V(\bar{y}_{lr})$

We have already shown that

$$\bar{y}_{lr} = \bar{y} - \beta\epsilon_1 \left[1 + \frac{\epsilon_2}{S_{xy}} \right] \left[1 + \frac{\epsilon_3}{S_x^2} \right]^{-1},$$

where $\bar{x} = \bar{X} + \epsilon_1$, $s_{xy} = S_{xy} + \epsilon_2$, $s_x^2 = S_x^2 + \epsilon_3$.

$$\therefore \bar{y}_{lr} \approx \bar{y} - \beta\epsilon_1 - \beta \left[\frac{\epsilon_1\epsilon_2}{S_{xy}} - \frac{\epsilon_1\epsilon_3}{S_x^2} \right].$$

$$\text{But } E(\bar{y}_{lr}) \approx \bar{Y} - \frac{N-n}{nN} \beta \left[\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{20}}{S_x^2} \right],$$

where $\mu_{21} = E(\bar{x} - \bar{X})^2(\bar{y} - \bar{Y})$, $\mu_{20} = E(\bar{x} - \bar{X})^2$.

$$\begin{aligned} \text{Therefore, } V(\bar{y}_{lr}) &= E[\bar{y}_{lr} - E(\bar{y}_{lr})]^2 \approx E[\bar{y} - \bar{Y} - \beta\epsilon_1]^2 \\ &= E[(\bar{y} - \bar{Y})^2 + \beta^2\epsilon_1^2 - 2\beta\epsilon_1(\bar{y} - \bar{Y})] \\ &= E(\bar{y} - \bar{Y})^2 + \beta^2 E(\epsilon_1^2) - 2\beta E\{\epsilon_1(\bar{y} - \bar{Y})\} \\ &= V(\bar{y}) + \beta^2 V(\bar{x}) - 2\beta \text{Cov}(\bar{x}, \bar{y}) \\ &= \frac{1-f}{n} [S_y^2 + \beta^2 S_x^2 - 2\beta S_{yx}] \\ &= \frac{1-f}{n} S_y^2 (1 - \rho^2). \end{aligned}$$

Therefore, if $b = \hat{\beta}$ is used instead of β , the variance becomes

$$V(\bar{y}_{lr}) = \frac{1-f}{n} [S_y^2 + b^2 S_x^2 - 2b S_{yx}].$$

This formula is valid if sample size n is large.

Corollary : In simple random sampling if sample size n is large, then the estimator of variance of linear regression estimator is

$$v(\bar{y}_{lr}) = \frac{1-f}{n(n-2)} \left[\sum_{i=1}^n (y_i - \bar{y})^2 - \frac{\left\{ \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right\}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

Proof: We know $V(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1 - \rho^2)$.

We define $u_i = y_i - \beta(x_i - \bar{X})$.

Then $S_u^2 = \frac{1}{N-1} \sum^N (u_i - \bar{U})^2$.

Estimate of S_u^2 is $s_u^2 = \frac{1}{n-1} \sum^n (u_i - \bar{u})^2$.

Again, $\bar{u} = \bar{y} + \beta(\bar{X} - \bar{x}) = \bar{y}_{lr}$,

$\therefore V(\bar{y}_{lr}) = V(\bar{u}) = \frac{1-f}{n} S_u^2$.

Hence, estimator of $V(\bar{y}_{lr})$ is $v(\bar{y}_{lr}) = \frac{1-f}{n} s_u^2$.

We have $u_i - \bar{u} = (y_i - \bar{y}) - \beta(x_i - \bar{x}) = [(y_i - \bar{y}) - b(x_i - \bar{x})] + (b - \beta)(x_i - \bar{x})$.

For large n the last term in the right-hand side becomes negligible. Then

$$u_i - \bar{u} = (y_i - \bar{y}) - b(x_i - \bar{x}).$$

$\therefore s_u^2 = \frac{1}{n-1} \sum^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$.

$\therefore v(\bar{y}_{lr}) = \frac{1-f}{n(n-2)} \left[\sum^n (y_i - \bar{y})^2 - \frac{\{\sum (x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum (x_i - \bar{x})^2} \right]$.

Here $(n-2)$ instead of $(n-1)$ is used since in regressions analysis s_u^2 is calculated using $(n-2)$ in the denominator.

15.15 Comparison of Regression Estimator, Ratio Estimator and Simple Estimator

Let y_1, y_2, \dots, y_n be a random sample of n observations selected from a finite population of size N . Then the variances of simple estimator, ratio estimator and regression estimator of population mean are, respectively,

$$V(\bar{y}) = \frac{1-f}{n} S_y^2$$

$$V(\hat{Y}_R) = \frac{1-f}{n} [S_y^2 + R^2 S_x^2 - 2RS_{yx}]$$

$$V(\bar{y}_{lr}) = \frac{1-f}{n} [S_y^2 + \beta^2 S_x^2 - 2\beta S_{yx}].$$

Let us assume that n is large. For large n the regression coefficient b of β tends to β . Then

$$V(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1 - \rho^2).$$

Since $|\rho| > 0$ $[-1 < \rho \leq 1]$,

$$V(\bar{y}_{lr}) < V(\bar{y}).$$

However, if $\rho = 0$, $V(\bar{y}_{lr}) = V(\bar{y})$.

The variance of regression estimator becomes less than the variance of ratio estimator, if

$$V(\bar{y}_{lr}) < V(\bar{Y}_R) \text{ or, } -\rho^2 S_y^2 < R^2 S_x^2 - 2RS_y S_x \rho$$

or, $(\rho S_y - RS_x)^2 > 0 \text{ or, } (\beta - R)^2 > 0.$

It is observed that, unless $\beta = R$ the regression estimator is more efficient than the ratio estimator. However, if y_i and x_i are linearly related and the regression line of y on x passes through the origin, then ratio estimator is the best.

Example 15.6 : There are 400 villages in a sub-division of a district. Thirty villages out of 400 villages are randomly selected to estimate the total cultivated land for jute in the district. The number of jute grower farmers (x) and amount of land cultivated (y in acres) for jute by these farmers are recorded by an inspection. The information are given below :

SL. No. of villages	y	x	SL. No. of villages	y	x	SL. No. of villages	y	x
1	5.4	3	11	6.2	3	21	20.2	5
2	10.6	5	12	4.4	2	22	18.5	6
3	15.2	10	13	8.5	5	23	12.2	4
4	12.7	8	14	20.8	10	24	15.0	7
5	8.5	4	15	24.0	10	25	8.2	2
6	10.0	4	16	20.0	8	26	10.5	5
7	16.2	8	17	18.8	6	27	12.6	8
8	15.5	6	18	14.2	7	28	17.2	6
9	12.2	7	19	11.3	12	29	5.6	2
10	10.5	3	20	14.4	8	30	8.5	4

- (i) Estimate total land area cultivated for jute.
- (ii) Estimate the variance of your estimator.
- (iii) Compare your estimator with ratio estimator and simple estimator.

Given that there are 2450 jute grower farmers in the study area.

Solution : (i) We have $N = 400, n = 30, X = 2450$. Then $\bar{X} = 6.125$. We have $\sum(y - \bar{y})^2 = 740.983, \bar{y} = 12.93, \bar{x} = 5.93, s_x^2 = 6.96. \sum(x - \bar{x})^2 = 201.87.$

$$s_y^2 = 25.55, \sum(x - \bar{x})(y - \bar{y}) = 259.56.$$

$$b = \hat{\beta} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{259.56}{201.87} = 1.286.$$

The estimate of mean land area cultivated for jute is

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) = 12.93 + 1.286(6.125 - 5.93) = 13.181 \text{ acres.}$$

The estimate of total land area cultivated for jute in the study area is

$$\hat{Y}_{lr} = N\bar{y}_{lr} = 400 \times 13.181 = 5272.40 \text{ acres.}$$

If $b_0 = 1$ is considered, then $\hat{Y}_{lr} = N[\bar{y} + b_0(\bar{X} - \bar{x})] = 5250.00 \text{ acres.}$

(ii) The estimator of variance of \hat{Y}_{lr} is

$$\begin{aligned} v(\hat{Y}_{lr}) &= N^2 v(\bar{y}_{lr}) = \frac{N^2(1-f)}{n(n-2)} \left[\sum (y - \bar{y})^2 - \frac{\{\sum (x - \bar{x})(y - \bar{y})\}}{\sum (x - \bar{x})^2} \right] \\ &= \frac{400^2(1-0.075)}{30(30-2)} \left[740.983 - \frac{(259.56)^2}{201.87} \right] \\ &= 71752.95. \end{aligned}$$

The estimate of variance of \hat{Y}_{lr} , when $b_0 = 1$, is an assumed value,

$$\begin{aligned} v(\hat{Y}_{lr}) &= N v(\bar{y}_{lr}) = N^2 \frac{1-f}{n} [s_y^2 + b_0^2 s_x^2 - 2b_0 s_{yx}] \\ &= \frac{(400)^2(1-0.075)}{30} [25.55 + 1^2 \times 6.96 - 2 \times 1 \times 8.95] \\ &= 72076.00. \end{aligned}$$

(iii) The simple estimate of total land area cultivated for jute is

$$\hat{Y} = N\bar{y} = 400 \times \frac{387.9}{30} = 5172.00 \text{ acres.}$$

$$v(\hat{Y}) = N^2 v(\bar{y}) = (400)^2 \frac{1-f}{n} s_y^2 = (400)^2 \frac{(1-0.075)}{30} \times 25.55 = 126046.67.$$

The ratio estimate of total land area cultivated for jute is

$$\hat{Y}_R = N\hat{Y}_R = N \frac{\bar{y}}{\bar{X}} = \frac{400 \times 12.93 \times 6.125}{5.93} = 5342.07 \text{ acres.}$$

The estimate of variance of \hat{Y}_R is

$$\begin{aligned} v(\hat{Y}_R) &= N^2 v(\hat{Y}_R) = \frac{N^2(1-f)}{n} [s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{yx}] \\ &= \frac{(400)^2(1-0.075)}{30} [25.55 + (2.18)^2 6.96 - 2 \times 2.18 \times 8.95] \\ &= 96716.54. \end{aligned}$$

It is observed that, even if β is estimated, the variance of \hat{Y}_{lr} is less than the variances of \hat{Y} and \hat{Y}_R . The relative efficiency of \hat{Y}_{lr} compared to \hat{Y} is

$$\frac{v(\hat{Y}) - v(\hat{Y}_{lr})}{v(\hat{Y}_{lr})} \times 100\% = \frac{126046.67 - 71752.95}{71752.95} \times 100\% = 75.67\%.$$

Therefore, the gain in precision of regression estimator compared to simple estimator is 75.67%. The gain in precision of regression estimator compared to ratio estimator is

$$\frac{v(\hat{Y}_R) - v(\hat{Y}_{lr})}{v(\hat{Y}_{lr})} \times 100\% = \frac{96716.54 - 71752.95}{71752.95} \times 100\% = 34.79\%.$$

15.16 Regression Estimator in Case of Stratified Random Sampling

We have already mentioned that the population units are divided into strata so that the observations within a stratum are internally homogeneous and then sampling units from each stratum are selected by simple random sampling. This process of selection of sampling units

from a population is known as stratified random sampling. Let \bar{Y}_h and \bar{X}_h be the population means of the study variable y and its related variable x of h -th stratum ($h = 1, 2, \dots, L$). Consider that N_h is the population size of h -th stratum, where the entire population units are $N = \sum_h^L N_h$. Let \bar{y}_h and \bar{x}_h be the simple estimate of \bar{Y}_h and \bar{X}_h , respectively observed from

sample size n_h selected from h -th stratum. The total sample size from all strata is $n = \sum_{h=1}^L n_h$.

The problem is to estimate population mean \bar{Y} using \bar{y}_h, \bar{x}_h and \bar{X}_h . This estimator can be found out by regression method of estimation.

Let $\bar{y}_{htr} = \bar{y}_h + b_h(\bar{X} - \bar{x}_h)$, $h = 1, 2, \dots, L$ be the regression estimator of \bar{Y}_h , where b_h is the sample regression coefficient of y on x . Using these separate regression estimators from all strata, we can find the regression estimator of \bar{Y} , where such an estimator is

$$\bar{y}_{lrs} = \sum_{j=1}^L W_h \bar{y}_{htr} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_{htr}; \quad W_h = \frac{N_h}{N}.$$

This estimator \bar{y}_{lrs} is a precise estimator of \bar{Y} , if the values of β_h vary too much. This estimator is known as *separate regression estimator*.

We can also define a combined regression estimator for population mean \bar{Y} . This is done if the values of β_h do not vary much. Here β_h is the regression coefficient of y on x for h -th stratum. The combined regression estimator is defined by

$$\bar{y}_{lrc} = \bar{y}_{st} + b(\bar{X} - \bar{x}_{st}),$$

where
$$\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h, \quad \bar{x}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{x}_h, \quad \bar{X} = \frac{1}{N} \sum_{h=1}^L \sum_{i=1}^{N_h} x_{hi}.$$

$$\therefore b = \frac{\sum_h^L \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{h=1}^L \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2} \quad \text{and} \quad b_h = \frac{\sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h)}{\sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}.$$

If the values of b_h and b are known, then it becomes easier to find the variance of \bar{y}_{lrs} and \bar{y}_{lrc} . Moreover, the estimator \bar{y}_{lrs} is an unbiased estimator of \bar{Y} , if b_h is known.

Theorem : If in stratified random sampling the sampling units from different strata are selected under SRS scheme and if the population regression coefficient of y on x for h -th stratum (β_h) is known, then the separate regression estimator \bar{y}_{lrs} is an unbiased estimator of \bar{Y} . The variance of this estimator is

$$V(\bar{y}_{lrs}) = \sum_{h=1}^L \frac{W_h^2(1 - f_h)}{n_h} [S_{hy}^2 + b_{ho}^2 S_{hx}^2 - 2b_{ho} S_{hyx}],$$

where b_{ho} is the known value of β_h .

Proof: Let β_h of h -th stratum be known and its value be b_{ho} . Then the regression estimator of \bar{Y}_h is

$$\bar{y}_{htr} = \bar{y}_h + b_{ho}(\bar{X}_h - \bar{x}_h).$$

Since b_{ho} is known, $E(\bar{y}_{htr}) = \bar{Y}_h$.

We have $\bar{y}_{irs} = \sum_{h=1}^L W_h \bar{y}_{htr}$.

$$E(\bar{y}_{irs}) = \sum_{h=1}^L W_h E(\bar{y}_{htr}) = \sum_{h=1}^L W_h \bar{Y}_h = \bar{Y}.$$

$$\text{Now, } V(\bar{y}_{irs}) = \sum_{h=1}^L W_h^2 V(\bar{y}_{htr}) = \sum_{h=1}^L W_h^2 \frac{1-f_h}{n_h} [S_{hy}^2 + b_{ho}^2 S_{hx}^2 - 2b_{ho} S_{hyx}],$$

where $f_h = \frac{n_h}{N_h}$, $h = 1, 2, \dots, L$.

This result is true since sample is selected from h -th stratum under SRS scheme.

Corollary : In stratified random sampling if b_{ho} is a known value of β_h , then the variance of \bar{y}_{irs} becomes minimum and this minimum variance is given by

$$V_{\min}(\bar{y}_{irs}) = \sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} \left(S_{hy}^2 - \frac{S_{yhx}^2}{S_{hx}^2} \right).$$

Corollary : In stratified random sampling if b_{ho} is a known value of β_h , then the estimator of $V(\bar{y}_{irs})$ is given by

$$v(\bar{y}_{irs}) = \sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} [s_{hy}^2 + b_{ho}^2 s_{hx}^2 - 2b_{ho} s_{hyx}].$$

In practice, β_h is not known. It is estimated from sample observations where the estimator is

$$b_h = \frac{\sum_{i=1}^{n_h} (x_{hi} - \bar{X}_h)(y_{hi} - \bar{y}_h)}{\sum_{i=1}^{n_h} (x_{hi} - \bar{X}_h)^2}.$$

Then the regression estimator of \bar{Y}_h is

$$\bar{y}_{htr} = \bar{y}_h + b_h(\bar{X}_h - \bar{X}_h).$$

But this estimator \bar{y}_{htr} is not unbiased and hence,

$$\bar{y}_{irs} = \sum_{h=1}^L W_h \bar{y}_{htr} = \sum_{h=1}^L W_h [\bar{y}_h + b_h(\bar{X}_h - \bar{X}_h)]$$

is also not an unbiased estimator of \bar{Y} . The variance of this estimator is

$$V(\bar{y}_{irs}) \approx \sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} [S_{hy}^2 + b_h^2 S_{hx}^2 - 2b_h S_{hyx}].$$

The estimator of this variance is

$$v(\bar{y}_{irs}) = \sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h(n_h-2)} \left[\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 - b_h^2 \sum_{i=1}^{n_h} n_h (x_{hi} - \bar{X}_h)^2 \right].$$

All these results are presented using the results of the previous section.

Let us now consider the combined regression estimator, where the combined estimator is

$$\bar{y}_{trc} = \bar{y}_{st} + b(\bar{X} - \bar{x}_{st}).$$

If b is known, $E(\bar{y}_{trc}) = \bar{Y}$, the combined regression estimator is unbiased. But, when b is not known, we need to replace it by its estimator. In that case, combined estimator is not an unbiased estimator of \bar{Y} , because

$$\begin{aligned} \bar{y}_{trc} - \bar{Y} &= \bar{y}_{st} - b(\bar{X} - \bar{x}_{st}) - \bar{Y} = (\bar{y}_{st} - \bar{Y}) - b(\bar{X} - \bar{x}_{st}) \\ &= [\bar{y}_{st} - \bar{Y} + \beta(\bar{X} - \bar{x}_{st})] + (b - \beta)(\bar{X} - \bar{x}_{st}). \end{aligned}$$

It is clear that, if the sampling error of b is not zero, the combined regression estimator will not be unbiased. However, for large sample size, $(b - \beta)$ can be considered negligible and then $V(\bar{y}_{trc})$ becomes

$$V(\bar{y}_{trc}) = \sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} [S_{hy}^2 + \beta^2 S_{hx}^2 - 2BS_{hyx}], \text{ if } b = \beta.$$

Here $\beta = b = \frac{\sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} \sum_{i=1}^{N_h} (x_{hi} - \bar{x}_h)^2}$.

In such a situation $V(\bar{y}_{trc})$ becomes minimum. The estimator of β from sample observations is

$$b = \frac{\sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}.$$

However, if sample from h -th stratum is selected under proportional allocation and if $(n_h - 1)$ is replaced by n_h , then

$$b = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{h=1}^L \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}.$$

Using this value of b the estimator of $V(\bar{y}_{trc})$ is

$$v(\bar{y}_{trc}) = \sum_{h=1}^L \frac{W_h^2(1-f_h)}{n_h} [s_{hy}^2 + b^2 s_{hx}^2 - 2bs_{hyx}]$$

Example 15.7 : A researcher has decided to estimate the land area cultivated for high yielding variety of rice. The study area is consisted of 74 villages. These villages are divided into 7 strata in respect of total cultivable land. From each stratum, villages are selected by proportional allocation. The information of total cultivable land (X_h) in all villages (N_h) of h -th stratum, the cultivable land in selected villages (x_{hi}), number of selected villages (n_h) and the amount of land cultivated for HYV rice in selected villages (y_{hi}) are shown below :

Estimate the total amount of cultivable land for HYV rice. Also estimate the variance of your estimator.

SL. No. stratum	N_h	n_h	X_h (in hectares)	x_{hi} (in hectares)	y_{hi}
1	15	4	5750	150, 370, 475, 320	20, 60, 75, 101
2	12	3	3180	225, 465, 570	75, 180, 270
3	11	3	3678	460, 392, 282	125, 220, 68
4	7	2	1877	312, 162	170, 92
5	10	3	2895	575, 603, 475	150, 200, 180
6	8	2	2662	380, 420	120, 210
7	11	3	3192	472, 662, 390	150, 212, 203
Total	74	20	23234		

Solution : We have $N = 74$, $n = 20$, $X = 23234$

$$y_1 = 256, \bar{y}_1 = 64.00, x_1 = 1315, \bar{x}_1 = 328.75, s_y^2 = 1147.33$$

$$y_2 = 525, \bar{y}_2 = 175.00, x_2 = 1260, \bar{x}_2 = 420.00, s_y^2 = 9525.00$$

$$y_3 = 413, \bar{y}_3 = 137.67, x_3 = 1134, \bar{x}_3 = 378.00, s_y^2 = 5896.33$$

$$y_4 = 262, \bar{y}_4 = 131.00, x_4 = 474, \bar{x}_4 = 237.00, s_y^2 = 3042.00$$

$$y_5 = 530, \bar{y}_5 = 176.67, x_5 = 1653, \bar{x}_5 = 551.00, s_y^2 = 633.33$$

$$y_6 = 330, \bar{y}_6 = 165.00, x_6 = 800, \bar{x}_6 = 400.00, s_y^2 = 4050.00$$

$$y_7 = 565, \bar{y}_7 = 188.33, x_7 = 1524, \bar{x}_7 = 508.00, s_y^2 = 1122.33$$

$$s_{1x}^2 = 18372.92, s_{2x}^2 = 31275.00, s_{3x}^2 = 8068.00$$

$$s_{4x}^2 = 11250.00, s_{5x}^2 = 4528.00, s_{6x}^2 = 800.00$$

$$s_{7x}^2 = 19468.00, b_1 = 0.16301, b_2 = 0.54316, b_3 = 0.42159$$

$$b_4 = 0.52, b_5 = 0.03533, b_6 = 2.25, b_7 = 0.0846.$$

$$\bar{X}_1 = 383.33, \bar{X}_2 = 265.00, \bar{X}_3 = 334.36, \bar{X}_4 = 268.14, \bar{X}_5 = 289.50,$$

$$\bar{X}_6 = 332.75, \bar{X}_7 = 290.18.$$

Now, $\bar{y}_{htr} = \bar{y}_h + b_h(\bar{X}_h - \bar{x}_h)$; $h = 1, 2, \dots, 7$.

So, we have

$$\bar{y}_{1tr} = 72.897, \bar{y}_{2tr} = 90.810, \bar{y}_{3tr} = 119.272, \bar{y}_{4tr} = 147.193,$$

$$\bar{y}_{5tr} = 167.431, \bar{y}_{6tr} = 13.687, \bar{y}_{7tr} = 169.902.$$

Therefore, from separate regression estimator, we have

$$\begin{aligned} \bar{y}_{lrs} &= \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_{htr} = \frac{1}{74} [15 \times 72.897 + 12 \times 90.81 + 11 \times 119.272 + 7 \times 147.193 \\ &\quad + 10 \times 167.431 + 8 \times 13.687 + 11 \times 169.902] \\ &= 110.517 \end{aligned}$$

The estimate of total cultivable land for HYV rice is

$$\hat{Y}_{lrs} = N\bar{y}_{lrs} = 74 \times 110.517 = 8178.258 \text{ (hectares).}$$

The estimator of variance of this estimator is

$$v(\hat{Y}_{lrs}) = \sum \frac{N_h^2(1-f_h)}{n_h(n_h-2)} \left[\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 - b_h^2 \sum (x_{hi} - \bar{x}_h)^2 \right].$$

SL. No.	$\frac{N_h^2(1-f_h)}{n_h(n_h-2)} = C_h$	$C_h \sum_i (y_{hi} - \bar{y}_h)^2$	$C_h b_h^2 \sum (x_{hi} - \bar{x}_h)^2$
1	20.615	70956.62	30193.35
2	36.000	685800.00	664332.31
3	29.322	345765.02	84094.98
4	—	—	—
5	23.333	29554.98	263.75
6	—	—	—
7	29.322	65817.92	8171.20
Total		1197894.54	787055.59

Hence, $v(\hat{Y}_{lrs}) = 1197894.54 - 757055.59 = 410838.95$.

In this selection, the sample from h -th stratum is selected by proportional allocation and hence, we can use the estimator of b as

$$b = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{h=1}^L \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}$$

$$= \frac{1294944 - \frac{2881 \times 8160}{20}}{3701338 - \frac{(8160)^2}{20}} = \frac{119496}{372058} = 0.3212.$$

$$\bar{y}_{st} = 143.91, \bar{x}_{st} = 406.57, \bar{X} = 313.973$$

$$\bar{y}_{lrc} = \bar{y}_{st} + b(\bar{X} - \bar{x}_{st})$$

$$= 143.91 + 0.3212(313.973 - 406.57) = 114.17.$$

Therefore, estimated total cultivable land for HYV rice using combined regression estimator is

$$\hat{Y}_{lrc} = N\bar{y}_{lrc} = 74 \times 114.17 = 8448.58 \text{ (hectares).}$$

The estimated variance of \hat{Y}_{lrc} is

$$v(\hat{Y}_{lrc}) = \sum \frac{N_h^2(1-f_h)}{n_h(n_h-1)} \sum_{i=1}^{n_h} [(y_{hi} - \bar{y}_h) - b(x_{hi} - \bar{x}_h)]^2$$

$$= \sum \frac{N_h^2(1-f_h)}{n_h} [s_{hy}^2 + b^2 s_{hx}^2 - 2b s_{hyx}].$$

SL. No. Strata	$\frac{N_h^2(1-f_h)}{n_h} = C_{1h}$	$C_{1h}s_{hy}^2$	$C_{1h}s_{hx}^2$	$2bC_{1h}s_{hyx}$
1	41.231	47305.56	78154.34	79327.95
2	36.000	342900.00	116158.47	392859.72
3	29.322	172892.19	24406.78	64062.77
4	17.493	53213.71	20303.36	65739.39
5	23.333	14777.49	10900.04	2398.26
6	24.000	97200.00	1980.85	27751.68
7	29.322	32908.96	58893.31	31073.64
Total		761197.91	310797.15	663163.41

$$\therefore v(\hat{Y}_{irc}) = 761197.91 + 310797.15 - 663163.41 = 408831.65.$$

Chapter 16

Cluster Sampling

16.1 Introduction

The basic need to select simple random sample, stratified random sample and systematic sample is the list of population units (frame). The preparation of such a list is not very difficult if the population size is smaller and if the population units are confined in a smaller area. But if the survey is to be conducted throughout the country, it is difficult, time taking and expensive to prepare a frame. For example, the objective of a survey is to estimate the number of couples adopting family planning methods in a specified area or in a state or in the country. The couples of child-bearing ages are the population units and they are spread over the entire survey area. It is very difficult to prepare a list of couples of child-bearing ages and hence, a representative group of couples cannot be selected using simple random sampling, stratified sampling or using systematic sampling. Even if the list is available, the sampling units will be spread over the entire survey area and it will not be convenient to conduct the survey within the stipulated time period and within the limit of resources for the survey.

To avoid the problem, the entire study area can be divided into smaller administrative units and some of the units are to be selected by simple random sampling. The ultimate sampling units of each selected administrative units are to be investigated to collect the data according to the pre-determined objective. For example, to estimate the proportion of adopter couples, the study area can be divided into villages (if it is rural area) and some of the villages are to be selected by simple random sampling. The couples of child-bearing ages of each selected village are to be investigated. Here a village is called a cluster and the selection of clusters is known as cluster sampling. This cluster sampling is called one-stage cluster sampling, where the couples of villages constitute the cluster units. To investigate the couples, some of the districts can be selected randomly; from the randomly selected districts, some of the smaller administrative units, some of the villages can be selected randomly and finally the couples of child-bearing ages of a village can be investigated. This type of sampling is known as multi-stage cluster sampling or simply known as multi-stage sampling.

In case of one-stage cluster sampling, data can be collected from a randomly selected group of couples instead of all couples of the cluster. Such sampling is known as two-stage sampling. In the subsequent chapters and sections, the different aspects of multi-stage sampling will be discussed.

From the above discussion it is clear that, if the population units are divided into groups, where the units within a group are adjacent irrespective of homogeneity or heterogeneity in characteristic under study are called *clusters*. If in any sampling scheme the clusters are considered as a sampling units and some of the clusters are selected randomly to investigate the ultimate units within a cluster is called *cluster sampling*.

The cluster sampling is applied profitably in selecting the population units where list of units are not available. However, the necessary pre-condition of cluster sampling is that each population unit must be accommodated once and only once in any one cluster. This ensures that

neither a unit has chance to be excluded from the survey nor a unit has the chance to be included more than once in the sample. The above mentioned condition is needed to avoid the bias in cluster sampling.

The important step in cluster sampling is to prepare the clusters. Therefore, the cluster size plays a vital role in cluster sampling. If the cluster sizes are smaller, the ultimate sampling units are expected to be more homogeneous. In such a case the precision of the estimate of population parameter is expected to be increased. In practice, in an agricultural survey the nearby farmlands can constitute a cluster; in socioeconomic survey the families living in a smaller area may constitute a cluster; in studying the family planning activities of couples, a group of couples living nearby may constitute a cluster.

Advantages of Cluster Sampling : The advantages of cluster sampling are :

- (i) Since clusters are formed with neighbouring population units, the data collection from units within a cluster is easier, less costly. The survey can be conducted within a short period of time. Cluster sampling is also advantageous from the point of view of administration of survey work.
- (ii) This sampling scheme is used easily if the frame of population units is not available.

16.2 Method of Estimation in Cluster Sampling

Let there be NM units in a population, where the units are divided into M clusters each of M units. The problem is to select n clusters from N clusters by a random process. Let y_{ij} ($i = 1, 2, \dots, M$) be the value of the variable under study of j th unit in i th cluster. Then the population mean per element of i th cluster is

$$\bar{Y}_i = \frac{1}{M} \sum_{j=1}^M y_{ij}, \quad i = 1, 2, \dots, N.$$

The cluster sample mean is

$$\bar{y}_c = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i = \frac{1}{nM} \sum_{i=1}^n \sum_{j=1}^M y_{ij}.$$

The population mean per cluster is

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

The population mean per element is

$$\bar{Y} = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M y_{ij} = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i.$$

The variance of observations within i th cluster is

$$S_i^2 = \frac{1}{M-1} \sum_{j=1}^M (y_{ij} - \bar{Y}_i)^2, \quad i = 1, 2, \dots, N.$$

The mean of variances of cluster is

$$S_w^2 = \frac{1}{N} \sum_{i=1}^N S_i^2.$$

The variance of the cluster means (mean square error of clusters) is

$$S_b^2 = \frac{1}{N-1} \sum_i^N (\bar{Y}_i - \bar{Y})^2.$$

The variance of population observations is

$$S^2 = \frac{1}{NM-1} \sum \sum (y_{ij} - \bar{Y})^2.$$

Intra-cluster correlation coefficient of observations is

$$\rho = \frac{E(y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{E(y_{ij} - \bar{Y})^2} = \frac{\sum_i \sum_j \sum_{k \neq j} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{(M-1)(NM-1)S^2}.$$

Since all clusters have equal number of observations, $\bar{Y}_N = \bar{Y}$. The estimate of this \bar{Y} is \bar{y}_c .

Theorem : If n clusters are selected from N clusters by simple random sampling without replacement (SRSWOR), then \bar{y}_c is an unbiased estimator of \bar{Y} with variance.

$$V(\bar{y}_c) = \frac{1-f}{n} S_b^2 \approx \frac{1-f}{nM} S^2 [1 + (M-1)\rho],$$

where ρ is the intra-cluster correlation coefficient of observations and M is the cluster size.

Proof : We have $\bar{y}_c = \frac{1}{nM} \sum_i^n \sum_j^M y_{ij} = \frac{1}{n} \sum_i^n \bar{Y}_i$.

Since clusters are selected by SRS scheme,

$$E(\bar{Y}_i) = E \left[\frac{1}{M} \sum_j^M y_{ij} \right] = \frac{1}{N} \sum \bar{Y}_i.$$

$$\therefore E(\bar{y}_c) = \frac{1}{n} \sum^n E(\bar{Y}_i) = \frac{1}{n} \sum^n \frac{1}{N} \sum \bar{Y}_i = \frac{1}{n} \sum^n \bar{Y} = \bar{Y}.$$

Again, as clusters are selected by SRS scheme, the variance of \bar{y}_c is

$$V(\bar{y}_c) = \frac{N-n}{nN} \frac{1}{N-1} \sum (\bar{Y}_i - \bar{Y})^2 = \frac{N-n}{nN} S_b^2 = \frac{1-f}{n} S_b^2, \quad f = \frac{n}{N}.$$

$$\begin{aligned} \text{But } \sum_i^N (\bar{Y}_i - \bar{Y})^2 &= \sum_i^N \left(\frac{1}{M} \sum_j^M y_{ij} - \bar{Y} \right)^2 = \frac{1}{M^2} \sum_i^N \left(\sum_j^M y_{ij} - M\bar{Y} \right)^2 \\ &= \frac{1}{M^2} \sum_i^N \sum_j^M (y_{ij} - \bar{Y})^2 + \frac{1}{M^2} \sum_i^N \sum_j^M \sum_{k \neq j} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) \\ &= \frac{(NM-1)S^2}{M^2} + \frac{\rho S^2 (M-1)(NM-1)}{M^2} = \frac{(NM-1)S^2}{M^2} [1 + (M-1)\rho] \end{aligned}$$

$$\begin{aligned}
\text{Then } V(\bar{y}_c) &= \frac{N-n}{nN} \frac{1}{N-1} \frac{MN-1}{M^2} S^2 [1 + (M-1)\rho] \\
&= \frac{1-f}{n} \frac{NM-1}{(N-1)M^2} S^2 [1 + (M-1)\rho] \approx \frac{1-f}{n} S^2 [1 + (M-1)\rho] \\
&= \frac{S^2}{n} [1 + (M-1)\rho], \text{ if } f = \frac{n}{N} \text{ is neglected.}
\end{aligned}$$

It is observed that, in case of cluster sampling the variance of cluster sample mean depends on cluster size, number of clusters and intra-cluster correlation coefficient. If the number of observations per cluster is $M = 1$, then the cluster sampling is equivalent to simple random sampling of size n from N units. Hence, cluster sampling and simple random sampling are equally efficient. But if $M > 1$ and ρ is positive, the variance of cluster sample mean is greater than the variance of the simple random sample mean. The variance of cluster sample mean will be less than the variance of simple random sample mean, if $\rho < 0$. Therefore, cluster sample is profitably applied, if $\rho < 0$.

Corollary : The estimate of population total in case of cluster sampling when n clusters are selected by SRS scheme from N clusters each of M units is

$$\hat{Y} = NM\bar{y}_c$$

and its variance is

$$V(\hat{Y}) = N^2 M^2 (1-f) \frac{S_b^2}{n} \approx \frac{M^2 N^2 (1-f)}{n} S^2 [1 + (M-1)\rho].$$

Corollary : Let there be NM elements in a population. Consider that nM elements are selected from the population by SRSWOR technique. Then the estimator of population mean is

$$\bar{y} = \frac{1}{nM} \sum_i^n \sum_j^M y_{ij}.$$

The variance of this estimator is

$$V(\bar{y}) = \frac{1-f}{nM} S^2, \quad \text{where } S^2 = \frac{1}{NM-1} \sum_i^N \sum_j^M (y_{ij} - \bar{Y})^2.$$

16.3 Method of Estimation in Cluster Sampling when Clusters are of Unequal Sizes

Let there be N clusters in a population and i th cluster be of size M_i ($i = 1, 2, \dots, N$) such that total number of population units are $M_0 = \sum_{i=1}^N M_i$.

Then the population mean is

$$\bar{Y} = \frac{1}{M_0} \sum_i^N \sum_j^{M_i} y_{ij} = \frac{1}{M_0} \sum_{i=1}^N M_i \bar{Y}_i,$$

where $\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$ = the mean of i th cluster.

The population mean per cluster is $\bar{Y}_N = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i$.

Let us consider that n clusters are selected by SRS scheme from N clusters. Let y_{ij} be the observation of j th unit in i th selected cluster ($i = 1, 2, \dots, n; j = 1, 2, \dots, M_i$). Now, we can define three estimators of population mean \bar{Y} . The estimators are :

$$\bar{y}_{c1} = \frac{N}{nM_0} \sum_{i=1}^n M_i \bar{Y}_i = \frac{1}{n\bar{M}} \sum_{i=1}^n M_i \bar{Y}_i, \quad \bar{M} = \frac{M_0}{N}$$

$$\bar{y}_{c2} = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i$$

and $\bar{y}_{c3} = \frac{1}{M_1} \sum_{i=1}^n M_i \bar{Y}_i$, where $\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$, $M_1 = \sum_{i=1}^n M_i$.

Theorem : If n clusters are selected by SRSWR scheme from N clusters, where i th cluster is of size M_i , then \bar{y}_{c1} is the unbiased estimator of \bar{Y} and the variance of this estimator is

$$\begin{aligned} V(\bar{y}_{c1}) &= \frac{1-f}{n} S_{bc}^2 \\ &= \frac{S_{bc}^2}{n}, \text{ if } f = \frac{n}{N} \text{ is neglected.} \end{aligned}$$

Here $S_{bc}^2 = \frac{1}{N-1} \sum_{i=1}^N \left(\frac{M_i}{M} \bar{Y}_i - \bar{Y} \right)^2$.

Proof : The estimator \bar{y}_{c1} is given by $\bar{y}_{c1} = \frac{1}{n\bar{M}} \sum_{i=1}^n M_i \bar{Y}_i$.

$$E(\bar{y}_{c1}) = \frac{1}{n\bar{M}} \sum_{i=1}^n E(M_i \bar{Y}_i) = \frac{1}{n\bar{M}} \sum_{i=1}^n \sum_{j=1}^{M_i} \frac{M_i}{N} \bar{Y}_i = \bar{Y}.$$

Hence, \bar{y}_{c1} is an unbiased estimator of \bar{Y} .

The variance of the estimator is

$$\begin{aligned} V(\bar{y}_{c1}) &= E(\bar{y}_{c1} - \bar{Y})^2 = E \left[\frac{\sum^n M_i \bar{Y}_i}{n\bar{M}} - \bar{Y} \right]^2 = \frac{1}{n^2} E \left[\frac{\sum^n M_i \bar{Y}_i}{\bar{M}} - n\bar{Y} \right]^2 \\ &= \frac{1}{n^2} E \left[\left(\frac{M_1}{\bar{M}} - \bar{Y} \right) + \left(\frac{M_2}{\bar{M}} - \bar{Y} \right) + \dots + \left(\frac{M_n}{\bar{M}} - \bar{Y} \right) \right]^2 \\ &= \frac{1}{n^2} \left[\sum_{i=1}^n E \left(\frac{M_i \bar{Y}_i}{\bar{M}} - \bar{Y} \right)^2 + E \sum_i \sum_{i \neq j} \left(\frac{M_i \bar{Y}_i}{\bar{M}} - \bar{Y} \right) \left(\frac{M_j \bar{Y}_j}{\bar{M}} - \bar{Y} \right) \right]. \end{aligned}$$

But $E \left(\frac{M_i \bar{Y}_i}{\bar{M}} - \bar{Y} \right)^2 = \frac{N-1}{N} S_{bc}^2$ and $E \left(\frac{M_i \bar{Y}_i}{\bar{M}} - \bar{Y} \right) \left(\frac{M_j \bar{Y}_j}{\bar{M}} - \bar{Y} \right) = -\frac{S_{bc}^2}{N}$

[According to the variance of simple random simple mean]

$$\begin{aligned}
\text{Therefore, } V(\bar{y}_{c1}) &= \frac{1}{n^2} \left[\sum_{i=1}^n \frac{N-1}{N} S_{bc}^2 - \sum_{i \neq j}^n \frac{S_{bc}^2}{N} \right] = \frac{N-n}{nN} S_{bc}^2 \\
&= \frac{1-f}{n} S_{bc}^2 \\
&= \frac{1}{n} S_{bc}^2, \text{ if } f = \frac{n}{N} \text{ is neglected.}
\end{aligned}$$

Theorem : If n clusters are selected under SRSWR scheme from N clusters of unequal sizes, then the sample mean \bar{y}_{c2} is not unbiased estimator of population mean. The bias and variance of this estimator are, respectively

$$\text{Bias } (\bar{y}_{c2}) = \frac{1}{M} \text{Cov}(\bar{Y}_i, M_i)$$

$$V(\bar{y}_{c2}) = \frac{1-f}{n} S_{bc2}^2(\bar{Y}_i, M_i),$$

$$\text{where } \bar{y}_{c2} = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i, \quad S_{bc2}^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y}_N)^2.$$

$$\text{Proof: We have } E(\bar{y}_{c2}) = \frac{1}{n} \sum_{i=1}^n E(\bar{Y}_i) = \frac{1}{n} \sum_{i=1}^n \frac{1}{N} \sum_{i=1}^N \bar{Y}_i = \bar{Y}_N \neq \bar{Y}.$$

Therefore, \bar{y}_{c2} is not unbiased estimator of \bar{Y} . The bias of this estimator is given by

$$\begin{aligned}
\text{Bias } (\bar{y}_{c2}) &= E(\bar{y}_{c2}) - \bar{Y} = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i - \frac{1}{NM} \sum_{i=1}^N M_i \bar{Y}_i = \frac{1}{NM} \sum_{i=1}^N (\bar{M} \bar{Y}_i - M_i \bar{Y}_i) \\
&= -\frac{1}{NM} \sum_{i=1}^N \bar{Y}_i (M_i - \bar{M}) = -\frac{\text{Cov}(\bar{Y}_i, M_i)}{M}.
\end{aligned}$$

It is clear that, if the $\text{Cov}(\bar{Y}_i, M_i)$ is small, \bar{y}_{c2} can be used as an unbiased estimator of population mean. This covariance will be smaller if the variation in M_i 's is smaller. If M_i and \bar{Y}_i are independent, covariance will be zero and \bar{y}_{c2} can be used as unbiased estimator of \bar{Y} .

Again, the variance of \bar{y}_{c2} is

$$V(\bar{y}_{c2}) = E(\bar{y}_{c2} - \bar{Y}_N)^2 = \frac{1-f}{n} S_{bc2}^2, \quad \text{where } S_{bc2}^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y}_N)^2.$$

16.4 Estimation of Sampling Variance in Case of Cluster Sampling

Let there be NM units in a population, where units are divided into N clusters each of size M . Consider that n clusters are selected under SRSWOR scheme. Let the sample mean be

$$\bar{y}_c = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i = \frac{1}{nM} \sum_i \sum_j^M y_{ij}.$$

Then the variance of sample observations is

$$s^2 = \frac{1}{nM-1} \sum_i \sum_j^M (\hat{y}_{ij} - \bar{y}_c)^2.$$

The total sum of squares $(nM - 1)s^2$ can be partitioned using the technique of analysis of variance. The analysis of variance table is shown below :

ANOVA Table of Sample Observations

Sources of variation	d.f.	M.S.
Intra-class	$n - 1$	$\frac{M}{n - 1} \sum_{i=1}^n (\bar{Y}_i - \bar{y}_c)^2 = MS_b^2$
Inter-class	$n(M - 1)$	$\frac{1}{n(M - 1)} \sum_{i=1}^n \sum_{j=1}^M (y_{ij} - \bar{Y}_i)^2 = s_w^2$
Total	$nM - 1$	$\frac{1}{nM - 1} \sum_{i=1}^n \sum_{j=1}^M (y_{ij} - \bar{y}_c)^2 = s^2$

Using the above results of ANOVA table it is easy to show that s_b^2 and s_w^2 are unbiased estimators of S_b^2 and S_w^2 , respectively. But s^2 is not an unbiased estimator of S^2 since cluster sample of size nM cannot be considered as a simple random sample from NM observations in the population. However, from the identity of analysis of variance, we have

$$S^2 = \frac{M(N - 1)S_b^2 + N(M - 1)S_w^2}{NM - 1}$$

But unbiased estimators of S_b^2 and S_w^2 are s_b^2 and s_w^2 , respectively. Hence, replacing S_b^2 by s_b^2 and S_w^2 by s_w^2 , we get an unbiased estimator of S^2 as

$$\hat{S}^2 = \frac{M(N - 1)s_b^2 + N(M - 1)s_w^2}{NM - 1}$$

Corollary : If n clusters are selected from N clusters by SRSWOR scheme, then the unbiased estimator of variance of \bar{y}_c is

$$v(\bar{y}_c) = \frac{1 - f}{n} s_b^2, \text{ where } s_b^2 = \frac{1}{n - 1} \sum_{i=1}^n (\bar{Y}_i - \bar{y}_c)^2$$

$$= \frac{s_b^2}{n}, \text{ if } f = \frac{n}{N} \text{ is neglected.}$$

This latter variance is obtained if the sample is selected with replacement.

Corollary : In cluster sampling if clusters are of equal sizes and they are selected without replacement, then the unbiased estimator of variance of the estimator of population total is given by

$$v(\hat{Y}_c) = \frac{N^2 M^2 (1 - f)}{n} s_b^2 = \frac{N^2 M^2}{n} s_b^2, \text{ if } f = \frac{n}{N} \text{ is neglected}$$

or if sample is selected with replacement.

Corollary : In cluster sampling if clusters are of unequal sizes, then the variance of \bar{y}_{c1} is estimated by

$$v(\bar{y}_{c1}) = \frac{1 - f}{n} s_{bc}^2, \text{ where } s_{bc}^2 = \frac{1}{n - 1} \sum_{i=1}^n \left(\frac{M_i \bar{Y}_i}{M} - \bar{y}_{c1} \right)^2$$

The unbiased estimator of variance of \bar{y}_{c2} and \bar{y}_{c3} are given respectively by

$$v(\bar{y}_{c2}) = \frac{1-f}{n} s_{bc2}^2, \quad \text{where } s_{bc2}^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{Y}_i - \bar{y}_{c2})^2$$

$$v(\bar{y}_{c3}) = \frac{1-f}{n} s_{bc3}^2, \quad \text{where } s_{bc3}^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{Y}_i - \bar{y}_{c3})^2 \frac{M_i^2}{M_1}$$

Here $\bar{M}_1 = \frac{1}{n} \sum_{i=1}^n M_i$.

Corollary : In cluster sampling if clusters are of equal size and sample is drawn under SRSWOR scheme, the unbiased estimator of intra-cluster correlation coefficient is given by

$$\hat{\rho} = \frac{(n-1)Ms_b^2 - ns_w^2}{(n-1)Ms_b^2 + n(M-1)s_w^2}$$

16.5 Relative Efficiency of Cluster Sampling

Let there be NM elements in a population. Consider that nM elements are selected by simple random sampling without replacement (SRSWOR) scheme. Then the variance of the sample mean \bar{y} is

$$V(\bar{y}) = \frac{1-f}{nM} S^2,$$

where $\bar{y} = \frac{1}{nM} \sum_{i=1}^n \sum_{j=1}^M y_{ij}$, $S^2 = \frac{1}{NM-1} \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{Y})^2$.

Also, consider that NM units are divided into N clusters each of size M and n clusters are selected from N clusters by SRSWOR scheme. Then the variance of the cluster sample mean is

$$V(\bar{y}_c) = \frac{1-f}{n} S_b^2,$$

where $\bar{y}_c = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i$, $\bar{Y}_i = \frac{1}{M} \sum_{j=1}^M y_{ij}$, $S_b^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2$.

$$\begin{aligned} \text{We have } (NM-1)S^2 &= \sum_i^N \sum_j^M (y_{ij} - \bar{Y})^2 = \sum_i^N \sum_j^M [(y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y})]^2 \\ &= \sum_i^N \sum_j^M (y_{ij} - \bar{Y}_i)^2 + M \sum_i^N (\bar{Y}_i - \bar{Y})^2 \\ &= (M-1) \sum_i^N S_i^2 + M(N-1)S_b^2 = N(M-1)S_w^2 + M(N-1)S_b^2, \end{aligned}$$

where $S_w^2 = \frac{1}{N(M-1)} \sum_i^N \sum_j^M (y_{ij} - \bar{Y}_i)^2$, $MS_b^2 = \frac{1}{N-1} \sum_i^N (\bar{Y}_i - \bar{Y})^2$.

Again, $MS_b^2 = \frac{1}{N-1} [(NM-1)S^2 - N(M-1)S_w^2]$.

Therefore, the relative efficiency of cluster sampling compared to simple random sampling is given by

$$\text{R.E. (Cluster)} = \frac{V(\bar{y})}{V(\bar{y}_c)} = \frac{S^2}{MS_b^2} = \frac{(N-1)S^2}{(NM-1)S^2 - N(M-1)S_w^2}$$

It is observed that, if the variance of the cluster means is minimum, the relative efficiency of cluster sampling is increased. Again, with the increase in the variance of observations within the clusters the efficiency is increased. Therefore, the cluster sample will provide efficient estimator of population parameter if clusters are formed in such a way that the variance of the observations within a cluster is more but the variance of the cluster means is less. But if nM units are selected randomly from NM units in the population and n clusters are formed, then the variance of the observations with clusters and the variance of the cluster means will be of same order and cluster sampling and SRS will be of same efficiency.

It is noted that the clusters are so formed that the variance of the observations within cluster becomes minimum. As a result the variance of sample mean of cluster sample is expected to be more than the variance of the sample mean when a sample of nM units are selected randomly from the population. The variance of cluster sample mean increases with the increase in number of clusters. The phenomenon can be observed if variance is expressed in terms of intra-cluster correlation coefficient. The intra-cluster correlation coefficient is

$$\rho = \frac{E(y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{E(y_{ij} - \bar{Y})^2}, \quad j \neq k = 1, 2, \dots, M.$$

$$\begin{aligned} \text{We have } V(\bar{y}_c) &= \frac{1-f}{n} \frac{(NM-1)}{M^2(N-1)} S^2 [1 + (M-1)\rho] \\ &\approx \frac{S^2}{nM} [1 + (M-1)\rho], \text{ if } f = \frac{n}{N} \text{ is neglected.} \end{aligned}$$

$$\text{Then R.E. (cluster)} = \frac{V(\bar{y})}{V(\bar{y}_c)} \approx \frac{S^2/nM}{\frac{S^2}{nM} [1 + (M-1)\rho]} \approx \frac{1}{1 + (M-1)\rho}$$

It is observed that, if $M = 1$, both sampling schemes are of equal efficiency. But, if $M > 1$, $(M-1)\rho$ increases and efficiency decreases. Hence, $(M-1)\rho$ is a measure of relative change in the sampling variance of cluster sampling. In practice, ρ is positive and its value decreases with the increase in size of M . However, the rate of decrease in the value of ρ is not similar to the rate of increase in the value of M . Hence, $V(\bar{y}_c)$ increases if M increases. If $\rho = 1$, $S_w^2 = 0$, and in such a situation, cluster sampling is not efficient.

Let us now discuss the relative efficiency of cluster sampling when clusters are of unequal sizes. We have noted that \bar{y}_{c1} is an unbiased estimator of population mean and its variance is

$$V(\bar{y}_{c1}) = \frac{1-f}{n} S_{bc}^2, \quad \text{where } \bar{y}_{c1} = \frac{1}{n\bar{M}} \sum_{i=1}^n M_i \bar{Y}_i.$$

Here the sample size $M_1 = \sum_{i=1}^n M_i$ is a random variable with expectation $n\bar{M}$. If a sample of size $n\bar{M}$ is selected from any population by SRS technique, then the variance of sample mean will be

$$V(\bar{y}) = \frac{\bar{M}N - n\bar{M}}{N\bar{M}n\bar{M}} S^2 = \frac{1-f}{n\bar{M}} S^2.$$

In such a case, the relative efficiency of cluster sampling compared to simple random sampling becomes

$$\text{R.E. (Cluster)} = \frac{V(\bar{y})}{V(\bar{y}_{c1})} = \frac{S^2}{MS_{bc}^2}.$$

It is observed that the efficiency of cluster sampling is increased if the variance of cluster means (S_{bc}^2) is decreased.

16.6 Cluster Sampling with Varying Probabilities

To select cluster sample we have mentioned that clusters are selected randomly by simple random sampling where all clusters have equal chance of selections. Clusters may be selected randomly with varying probabilities, i.e., i th cluster will be selected with probability p_i , where $p_i = M_i/M_0$ ($i = 1, 2, \dots, N$);

$$M_0 = \sum_{i=1}^N M_i.$$

This is specially done if clusters are of unequal sizes. The total probability of selection is

$$\sum p_i = 1 \quad (0 < p_i < 1).$$

The clusters can be selected using Lahiri's method of selection. The method is described below :

Let there be N clusters numbered $1, 2, \dots, N$. Consider that in any cluster the maximum number of units is M . Now we need to select a pair of random numbers to include a cluster in the sample. The first number of the pair is a random number from 1 to N : The second number of the pair is a random number from 1 to M . If this second number is less than or equal to the size (M) of cluster corresponding to the first selected number, then that cluster is included in the sample, otherwise this pair of random numbers is deleted and a second pair is selected. If the second pair corresponds to a cluster and second number in the pair of random numbers is less than or equal to size of the cluster, then the respective cluster is included in the sample. The process is repeated unless required n cluster are selected in the sample. For example, let us consider that there are 10 clusters in a population and the elements in the clusters are $M_1 = 150, M_2 = 90, M_3 = 240, M_4 = 25, M_5 = 98, M_6 = 140, M_7 = 80, M_8 = 75, M_9 = 120, M_{10} = 165$. Now, we need to select a random number from 1 to 10 and a number from 1 to 240 ($\because M_3 = 240$ is the largest size of the cluster). Using Random Number Table of Appendix, we get the pair (01,034). This pair corresponds to first cluster. Hence, first cluster is included in the sample. The second pair is (06,161). But sixth cluster has $M_6 = 140$ elements. Hence, this cluster is not included in the sample. The pair is deleted. Another pair is to be selected. The process is continued until the required sample is selected.

Let y_{ij} be the value of the variable under study of j th unit in i th cluster. Let us define a new variable Z such that

$$Z_{ij} = \frac{M_i y_{ij}}{M_0 p_i}; \quad i = 1, 2, \dots, N; \quad j = 1, 2, \dots, M_i.$$

Then $\bar{Z}_i = \frac{M_i \bar{Y}_i}{M_0 p_i}$, where $\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$.

Here \bar{Y}_i is the mean of i th cluster and \bar{Z}_i is the mean related to \bar{Y}_i . We have

$$E(\bar{Z}_i) = \sum_{i=1}^N p_i \bar{Z}_i = \sum_{i=1}^N p_i \frac{M_i \bar{Y}_i}{M_0 p_i} = \bar{Y}$$

Let us write $\bar{Y} = \bar{Z}_{..}$ and define an estimator \bar{Z} on the basis of n selected clusters such that

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n \bar{Z}_i = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i$$

This result is also true, if $p_i \neq M_i/M_0$.

Theorem : In cluster sampling if clusters are selected by probability proportional to size (PPS) of cluster with replacement, then sample mean \bar{Z} is an unbiased estimator of population mean \bar{Y} . The variance of \bar{Z} is given by

$$V(\bar{Z}) = \frac{1}{n} \sum_{i=1}^N p_i (\bar{Z}_i - \bar{Z}_{..})^2, \quad \text{where } \bar{Z} = \frac{1}{n} \sum_{i=1}^n \bar{Z}_i$$

Proof : $E(\bar{Z}) = \frac{1}{n} \sum_{i=1}^n E(\bar{Z}_i) = \frac{1}{n} \sum_{i=1}^n \bar{Y} = \bar{Y}$

We know $V(\bar{Z}) = E(\bar{Z}^2) - [E(\bar{Z})]^2$

$$= E \left[\frac{1}{n} \sum_{i=1}^n \bar{Z}_i \right]^2 - \bar{Z}_{..}^2 \quad [\because E(\bar{Z}) = \bar{Y} = \bar{Z}_{..}]$$

$$= \frac{1}{n^2} \left[E \sum_{i=1}^n \bar{Z}_i^2 + E \sum_{i \neq j} \bar{Z}_i \bar{Z}_j \right] - \bar{Z}_{..}^2$$

Again, $E(\bar{Z}_i^2) = \sum_{i=1}^N p_i \bar{Z}_i^2$ and $E(\bar{Z}_i \bar{Z}_j) = E(\bar{Z}_i) E(\bar{Z}_j) = \bar{Z}_{..}^2$

Therefore, $V(\bar{Z}) = \frac{1}{n} E(\bar{Z}_i^2) + \frac{n-1}{n} E(\bar{Z}_i) E(\bar{Z}_j) - \bar{Z}_{..}^2 = \frac{1}{n} \left[\sum_{i=1}^N p_i \bar{Z}_i^2 - \bar{Z}_{..}^2 \right]$

$$= \frac{1}{n} \sum_{i=1}^N p_i (\bar{Z}_i - \bar{Z}_{..})^2 = \frac{1}{n} \sum_{i=1}^N p_i (\bar{Y}_i - \bar{Y})^2$$

Corollary : In cluster sampling under PPS sampling scheme the unbiased estimator of variance of \bar{Z} is

$$v(\bar{Z}) = \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{Y}_i - \bar{y}), \quad \text{where } \bar{y} = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i$$

Corollary : In cluster sampling if i th cluster is selected with probability p_i , then the unbiased estimator of $V(\bar{Z})$ is

$$v(\bar{Z}) = \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{Z}_i - \bar{Z})^2$$

Example 16.1 : To estimate the family planning adopter couples in a police station area a survey is conducted. In that police station area there are 120 villages. These villages are divided into 40 clusters. The number of villages in clusters are different. Ten clusters are selected by probability proportional to number of villages in a cluster. The number of villages in a cluster, number of couples of child-bearing ages who adopted family planning are shown below :

Sl. No.	Number of villages in clusters M_i	Number of adopter couples $M_i \bar{Y}_i$	\bar{Y}_i	$p_i = \frac{M_i}{M_0}$	$\bar{Z}_i = \frac{M_i \bar{Y}_i}{M_0 p_i}$	Given p_i
1	4	150	37.5	0.0333	41.67	0.03
2	2	95	47.5	0.0167	39.58	0.02
3	3	175	58.33	0.025	58.33	0.025
4	2	120	60.00	0.0167	50.00	0.02
5	2	70	35.00	0.0167	29.17	0.02
6	4	250	62.50	0.0333	69.44	0.03
7	4	300	75.00	0.0333	83.33	0.03
8	3	275	91.67	0.025	91.67	0.025
9	3	120	40.00	0.025	40.00	0.025
10	2	135	67.5	0.0167	56.25	0.02

- (i) Estimate the total adopter couples in the police station area.
(ii) Estimate the standard error of your estimator.
(iii) In the above example, if $p_i \neq M_i/M_0$, then what would be your estimators.

Solution : (i) Here $p_i = M_i/M_0$. Hence, average adopter couple is

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i = \frac{575.0}{10} = 57.5.$$

The estimate of total adopter couples in the study area is

$$\hat{Y} = M_0 \bar{Z} = 120 \times 57.5 = 6900.$$

- (ii) The estimated variance of this estimator is

$$v(\hat{Y}) = M_0^2 v(\bar{Z}),$$

$$\begin{aligned} \text{where } v(\bar{Z}) &= \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{Y}_i - \bar{y})^2 = \frac{1}{n(n-1)} \left[\sum \bar{Y}_i^2 - \frac{(\sum \bar{Y}_i)^2}{n} \right] \\ &= \frac{1}{10(10-1)} \left[35980.7778 - \frac{(575)^2}{10} \right] = 32.4253. \end{aligned}$$

Therefore, $v(\hat{Y}) = (120)^2 32.4253 = 466924.32$.

$$\text{s.e.}(\hat{Y}) = \sqrt{v(\hat{Y})} = 683.32.$$

- (iii) If $p_i \neq M_i/M_0$, then we define

$$\bar{Z}_i = \frac{M_i \bar{Y}_i}{M_0 p_i} \quad \text{and} \quad \bar{Z} = \frac{1}{n} \sum \bar{Z}_i = \frac{559.44}{10} = 55.944.$$

Hence $\hat{Y} = M_0 \bar{Z} = 120 \times 55.944 = 67132.8 \approx 67133$.

$$v(\hat{Y}) = \frac{M_0^2}{n(n-1)} \left[\sum \bar{Z}_i^2 - \frac{(\sum \bar{Z}_i)^2}{n} \right] = \frac{(120)^2}{10(10-1)} \left[34989.497 - \frac{(559.44)^2}{10} \right]$$

$$= 590749.7024.$$

$$\text{s.e.}(\hat{Y}) = \sqrt{v(\hat{Y})} = 768.60.$$

Example 16.2 : To estimate the total production of marigold in a police station the area is divided into 55 clusters in such a way that in each cluster there are 10 farmers living nearby area. The production data of marigold of each farmer are shown below :

Sl.No. of clusters	Production of marigold (in kg) of different farmers										Total Y_i
	1	2	3	4	5	6	7	8	9	10	
1	4.5	5.0	6.2	7.5	4.0	4.0	5.6	8.2	7.6	9.5	62.1
2	5.0	5.6	7.2	8.5	9.0	5.0	5.0	6.7	6.0	6.0	64.0
3	8.0	8.0	8.0	7.6	7.8	6.2	4.5	4.2	4.0	4.0	62.3
4	5.5	3.2	3.4	6.8	8.0	9.0	9.2	9.0	6.7	8.2	69.8
5	4.0	5.5	6.0	6.0	6.5	6.5	6.5	8.0	8.7	8.5	66.2
6	6.2	3.7	3.8	4.5	4.0	9.7	4.6	4.8	5.6	6.7	53.6
7	4.0	8.0	9.3	9.8	3.6	2.8	5.5	5.0	5.0	6.5	59.5
8	8.6	8.9	6.3	9.7	4.2	4.0	3.8	3.0	3.5	7.6	59.6
9	6.0	4.6	4.8	3.0	9.7	8.7	4.2	4.5	6.0	8.7	60.2
10	8.0	9.0	8.5	8.0	4.2	5.0	8.7	5.6	3.2	3.0	63.2
11	4.0	9.0	6.5	6.2	8.7	3.5	3.5	4.2	4.8	5.0	55.4
12	7.5	7.2	6.0	8.7	4.8	9.6	4.8	5.0	6.6	6.0	66.2
13	3.5	3.0	4.0	5.5	6.7	6.2	6.4	7.0	8.0	4.2	54.5
14	5.0	6.0	6.3	4.8	7.0	7.2	5.0	5.5	5.6	8.8	61.2
15	4.4	6.2	6.0	5.8	6.3	9.0	6.7	8.0	8.2	4.8	65.4
16	6.0	3.0	4.4	5.0	5.8	6.2	8.7	6.4	7.8	6.5	59.8
17	8.4	9.6	4.5	6.0	6.7	6.6	3.8	3.0	4.0	5.7	58.3
18	5.5	5.0	5.2	6.2	4.8	7.6	7.0	6.2	6.4	6.0	59.9
19	6.2	6.0	4.5	6.0	8.5	7.2	6.0	4.5	5.0	5.2	59.1
20	8.0	7.2	7.0	6.2	4.0	4.5	7.0	4.8	6.7	9.2	64.6
21	7.2	4.0	5.2	6.3	4.8	6.6	8.2	5.2	3.4	3.0	53.9
22	6.0	8.2	9.6	9.0	4.0	5.2	5.3	6.0	6.4	3.5	63.2
23	7.0	7.5	6.0	6.6	5.0	5.0	4.3	3.8	4.1	4.0	53.3
24	7.2	6.0	4.1	4.2	4.0	4.0	3.8	9.7	6.8	7.4	57.2
25	6.0	6.2	5.0	5.0	5.5	5.5	6.0	7.2	3.8	4.5	54.7
26	4.5	4.0	6.2	7.0	7.2	8.0	5.6	5.0	5.0	4.8	57.3
27	4.0	6.0	5.0	4.8	3.6	4.4	5.2	4.6	6.7	7.2	51.5
28	6.4	7.2	3.0	3.4	2.7	6.6	4.8	6.4	4.2	4.5	49.2
29	5.0	5.2	9.8	6.7	6.8	4.4	6.7	4.2	5.0	4.9	58.7

Sl.No. of clusters	Production of marigold (in kg) of different farmers										Total Y_i
	1	2	3	4	5	6	7	8	9	10	
30	6.2	7.0	6.0	5.8	8.0	7.5	7.0	6.8	4.0	3.8	62.1
31	4.8	5.0	6.7	4.8	7.2	9.0	9.2	6.0	6.0	7.5	66.2
32	8.2	8.0	3.8	4.0	3.8	3.0	3.6	4.8	7.6	5.4	52.2
33	7.6	7.0	6.2	8.7	6.2	6.0	8.0	8.2	6.7	7.0	71.6
34	4.4	5.5	6.2	4.8	7.0	7.2	6.2	5.6	5.0	5.0	56.9
35	4.8	4.4	5.4	6.2	6.0	7.0	6.8	8.7	9.0	8.0	66.3
36	5.6	6.2	6.8	7.0	6.7	6.4	5.8	5.4	6.7	6.8	63.4
37	7.4	9.0	8.7	8.0	5.0	5.5	4.8	3.6	4.2	4.0	60.2
38	6.7	6.2	7.0	7.0	5.2	5.4	3.9	4.9	4.1	5.1	55.5
39	6.0	6.0	6.0	6.0	5.2	4.8	5.0	4.6	5.0	4.8	53.4
40	7.0	7.4	7.0	5.5	5.3	5.1	6.1	6.2	3.8	4.7	58.1
41	8.0	8.2	7.6	7.6	4.8	3.0	3.6	4.2	4.8	5.6	57.4
42	5.5	6.5	4.5	4.1	4.2	4.4	6.0	7.2	6.6	5.0	53.6
43	5.0	5.2	5.1	7.0	7.3	4.3	8.0	7.1	7.2	5.5	61.7
44	6.1	6.4	6.3	7.2	7.0	4.8	7.2	4.9	5.0	5.2	60.1
45	8.5	7.6	7.2	7.4	6.2	6.7	5.8	5.9	4.2	4.8	64.3
46	7.2	7.0	4.3	3.9	3.0	3.2	4.4	4.6	4.7	9.0	53.3
47	6.4	6.2	6.4	8.0	8.1	7.5	3.4	3.0	3.8	9.5	62.3
48	4.7	4.8	6.0	6.2	6.0	6.2	6.0	6.4	5.8	5.0	57.1
49	4.0	4.5	6.5	6.7	9.0	6.7	6.8	5.0	5.8	4.4	59.4
50	6.0	8.0	9.2	9.3	9.4	6.4	6.7	4.4	5.5	5.0	69.9
51	3.8	3.8	4.0	6.6	7.0	7.2	4.0	4.0	4.0	4.2	48.6
52	4.4	4.6	6.7	6.2	6.7	7.2	3.0	3.8	6.8	7.0	58.4
53	4.8	4.0	8.7	8.0	8.2	6.6	6.0	7.7	7.1	6.2	67.3
54	9.2	9.0	4.6	5.8	6.2	7.6	7.8	6.0	3.8	4.0	64.0
55	7.4	8.2	4.2	4.3	3.8	6.6	6.0	4.0	4.8	5.0	54.3

- (i) Select a cluster sample of size $n = 15$ to estimate the total marigold production in the area.
- (ii) Find the relative efficiency of cluster sampling compared to simple random sampling. Also estimate the relative efficiency.

Solution : (i) Given $N = 55$, $M = 10$. We need to select $n = 15$ clusters. The selected clusters are shown below :

Sl. No.	Random Number	Production of marigold (in kg) of different farmers										Total Y_i
		1	2	3	4	5	6	7	8	9	10	
1	51	3.8	3.8	4.0	6.6	7.0	7.2	4.0	4.0	4.0	4.2	48.6
2	16	6.0	3.0	4.4	5.0	5.8	6.2	8.7	6.4	7.8	6.5	59.8
3	48	4.7	4.8	6.0	6.2	6.0	6.2	6.0	6.4	5.8	5.0	57.1
4	28	6.4	7.2	3.0	3.4	2.7	6.6	4.8	6.4	4.2	4.5	49.2
5	40	7.0	7.4	7.0	5.5	5.3	5.1	6.1	6.2	3.8	4.7	58.1

Sl. No.	Random Number	Production of marigold (in kg) of different farmers										Total Y_i
		1	2	3	4	5	6	7	8	9	10	
6	11	4.0	9.0	6.5	6.2	8.7	3.5	3.5	4.2	4.8	5.0	55.4
7	39	6.0	6.0	6.0	6.0	5.2	4.8	5.0	4.6	5.0	4.8	53.4
8	09	6.0	4.6	4.8	3.0	9.7	8.7	4.2	4.5	6.0	8.7	60.2
9	17	8.4	9.6	4.5	6.0	6.7	6.6	3.8	3.0	4.0	5.7	58.3
10	10	8.0	9.0	8.5	8.0	4.2	5.0	8.7	5.6	3.2	3.0	63.2
11	30	6.2	7.0	6.0	5.8	8.0	7.5	7.0	6.8	4.0	3.8	62.1
12	29	5.0	5.2	9.8	6.7	6.8	4.4	6.7	4.2	5.0	4.9	58.7
13	21	7.2	4.0	5.2	6.3	4.8	6.6	8.2	5.2	3.4	3.0	53.9
14	19	6.2	6.0	4.5	6.0	8.5	7.2	6.0	4.5	5.0	5.2	59.1
15	26	4.5	4.0	6.2	7.0	7.2	8.0	5.6	5.0	5.0	4.8	57.3

The means and variances of sample clusters are shown below :

Sl.No.	Cluster Mean \bar{Y}_i	$\sum_{j=1}^M y_{ij}^2$	$M\bar{Y}_i^2$	$s_i^2 = \frac{1}{M-1} \left[\sum_{j=1}^M y_{ij}^2 - M\bar{Y}_i^2 \right]$
1	4.86	254.92	236.196	2.0804
2	5.98	381.18	357.604	2.6196
3	5.71	329.61	326.041	0.3966
4	4.92	266.10	242.064	2.6707
5	5.81	349.29	337.561	1.3032
6	5.54	343.56	306.916	4.0716
7	5.34	288.28	285.156	0.3471
8	6.02	408.56	362.404	5.1284
9	5.83	379.35	339.889	4.3846
10	6.32	450.18	399.424	5.6396
11	6.21	403.01	385.641	1.9299
12	5.87	370.11	344.569	2.8379
13	5.39	316.01	290.521	2.8321
14	5.91	363.07	349.281	1.5321
15	5.73	343.93	328.329	1.7334
Total	85.44			

The total production of marigold is

$$\hat{Y}_c = NM\bar{y}_c = NM \frac{1}{n} \sum \bar{Y}_i = \frac{55 \times 10}{15} \times 85.44 = 3132.8 \text{ kg.}$$

$$(ii) \text{ We have } s_b^2 = \frac{1}{n-1} \left[\sum \bar{Y}_i^2 - \frac{(\sum \bar{Y}_i)^2}{n} \right] = \frac{1}{14} \left[489.1596 - \frac{(85.44)^2}{15} \right] = 0.1781.$$

The estimated variance of \hat{Y}_c is

$$\begin{aligned} v(\hat{Y}_c) &= N^2 M^2 v(\bar{y}_c) = \frac{N^2 M^2 (1-f)}{n} s_b^2 \\ &= \frac{(55)^2 (10)^2 (1-0.27)}{15} \times 0.1781 = 2621.9288. \end{aligned}$$

$$\begin{aligned} \text{Also, we have } s^2 &= \frac{1}{nM-1} \left[\sum \sum y_{ij}^2 - \frac{(\sum \sum y_{ij})^2}{nM} \right] \\ &= \frac{1}{10 \times 15 - 1} \left[5247.16 - \frac{(854.4)^2}{10 \times 15} \right] = 2.5536. \end{aligned}$$

$$\begin{aligned} s_w^2 &= \frac{1}{n(M-1)} \left[\sum \sum y_{ij}^2 - \frac{\sum^n Y_i^2}{M} \right] \\ &= \frac{1}{15(10-1)} \left[5247.16 - \frac{48915.96}{10} \right] = 2.6338. \end{aligned}$$

Hence, the estimate of S^2 is

$$\begin{aligned} \hat{S}^2 &= \frac{(N-1)M s_b^2 + N(M-1)s_w^2}{NM-1} \\ &= \frac{(55-1)10 \times 0.1781 + 55(10-1)2.6338}{55 \times 10 - 1} = 2.5499. \end{aligned}$$

Now, the estimate of relative efficiency of cluster sampling compared to SRS is

$$\text{r.e. (cluster)} = \frac{\hat{S}^2}{M s_b^2} = \frac{2.5499}{10 \times 0.1781} = 143.17\%.$$

The estimate of intra-cluster correlation coefficient is

$$\begin{aligned} \hat{\rho} &= \frac{(n-1)M s_b^2 - n s_w^2}{(n-1)M s_b^2 + n(M-1)s_w^2} \\ &= \frac{14 \times 10 \times 0.1781 - 15 \times 2.6338}{14 \times 10 \times 0.1781 + 15 \times 9 \times 2.6338} = -0.0383. \end{aligned}$$

$$\begin{aligned} \text{Again, } S^2 &= \frac{1}{NM-1} \left[\sum \sum y_{ij}^2 - \frac{(\sum \sum y_{ij})^2}{NM} \right] \\ &= \frac{1}{55 \times 10 - 1} \left[21006.77 - \frac{(3284.5)^2}{55 \times 10} \right] = 2.5361. \end{aligned}$$

$$\begin{aligned} S_b^2 &= \frac{1}{N-1} \left[\sum Y_i^2 - \frac{(\sum Y_i)^2}{N} \right] \\ &= \frac{1}{55-1} \left[1972.9027 - \frac{(328.45)^2}{55} \right] = 0.2122. \end{aligned}$$

Therefore, the relative efficiency of cluster sampling compared to SRS is

$$\text{R.E. (Cluster)} = \frac{S^2}{MS_b^2} = \frac{2.5361}{10 \times 0.2122} = 119.51\%.$$

$$\begin{aligned} \text{We have } S_w^2 &= \frac{1}{N(M-1)} \left[\sum \sum y_{ij}^2 - \frac{\sum Y_i^2}{M} \right] \\ &= \frac{1}{55(10-1)} \left[21006.77 - \frac{197290.27}{10} \right] = 2.5813. \end{aligned}$$

$$\therefore \rho = \frac{\frac{N-1}{N} S_b^2 - \frac{S_w^2}{M}}{\frac{NM-1}{NM} S^2} = \frac{\frac{55-1}{55} \times 0.2122 - \frac{2.5813}{10}}{\frac{55 \times 10 - 1}{55 \times 10} \times 2.5361} = -0.0197.$$

It is observed that ρ is negative and cluster sampling is more efficient than simple random sampling.

Example 16.3 : In a police station area there are 120 villages. These villages are divided into 20 clusters, where a cluster is formed with 6 neighbouring villages. In each cluster there are different number of farmers who have milk producing cows. The number of cows of farmers in different clusters are shown below :

Number of cows (y_{ij}) of farmers in different clusters

Sl.No. of clusters	Cluster size M_i	Number of cows in clusters .
1	19	4, 3, 2, 1, 4, 4, 2, 2, 2, 2, 1, 1, 2, 2, 3, 3, 2, 4, 2
2	22	5, 1, 2, 4, 4, 2, 2, 3, 3, 3, 4, 4, 1, 1, 1, 1, 2, 2, 3, 3, 4, 4, 5
3	22	4, 4, 4, 4, 2, 2, 1, 1, 1, 1, 1, 1, 3, 3, 3, 5, 5, 1, 1, 2, 2, 2
4	21	1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 4, 4, 2, 2
5	39	2, 2, 2, 2, 2, 2, 2, 2, 4, 1, 3, 1, 2, 3, 4, 1, 4, 1, 4, 5, 2, 3, 1, 2, 1, 3, 3, 3, 3, 3, 4, 2, 2, 1, 1, 1, 1, 5
6	42	1, 1, 4, 4, 2, 3, 1, 2, 1, 1, 2, 2, 2, 3, 2, 2, 2, 1, 4, 1, 1, 2, 2, 3, 2, 2, 2, 2, 4, 3, 2, 1, 1, 2, 4, 2, 2, 2, 1, 1, 1, 2
7	18	2, 1, 1, 4, 2, 2, 3, 4, 1, 3, 1, 1, 2, 2, 2, 2, 1, 4
8	34	1, 1, 4, 1, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 4, 3, 1, 2, 2, 3, 2, 2, 2, 1, 1, 2, 2, 1, 1, 2
9	14	4, 1, 2, 2, 3, 1, 2, 3, 1, 2, 2, 1, 4, 2
10	20	4, 1, 4, 2, 2, 2, 2, 1, 1, 1, 1, 2, 2, 3, 3, 2, 2, 2, 4, 1
11	16	2, 2, 2, 2, 4, 1, 1, 3, 3, 1, 4, 2, 2, 1, 4, 2
12	18	1, 1, 2, 2, 2, 4, 2, 3, 3, 2, 2, 1, 1, 1, 1, 2, 2, 1
13	22	1, 1, 2, 3, 3, 3, 4, 5, 1, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 3, 2, 2
14	34	4, 2, 1, 1, 3, 2, 1, 1, 2, 2, 2, 1, 1, 1, 2, 2, 4, 4, 3, 3, 2, 2, 2, 1, 1, 3, 2, 1, 1, 1, 2, 2, 1, 1
15	19	2, 2, 1, 2, 3, 3, 2, 2, 1, 4, 2, 2, 1, 1, 2, 4, 1, 1, 2
16	33	2, 1, 1, 1, 2, 2, 3, 3, 4, 4, 4, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 2, 1, 2, 2, 4, 2, 2, 4, 3, 2, 2, 2
17	40	2, 2, 2, 1, 1, 1, 2, 2, 1, 1, 1, 1, 2, 2, 2, 3, 3, 1, 2, 3, 4, 4, 2, 2, 3, 3, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 4, 5, 1

Sl.No. of clusters	Cluster size M_i	Number of cows in clusters
18	37	1, 1, 1, 4, 5, 6, 1, 2, 2, 1, 2, 3, 3, 3, 2, 1, 4, 4, 4, 2, 1, 2, 2, 3, 3, 2, 1, 2, 3, 3, 4, 5, 2, 4, 3, 2, 1
19	34	2, 3, 3, 3, 2, 4, 5, 5, 4, 3, 2, 1, 1, 2, 2, 2, 3, 4, 1, 1, 4, 2, 3, 2, 2, 3, 2, 1, 1, 3, 2, 2, 1, 2
20	30	3, 2, 4, 5, 4, 4, 6, 4, 2, 2, 2, 1, 2, 2, 1, 1, 2, 2, 3, 1, 1, 1, 2, 2, 3, 2, 2, 4, 3

- (i) Select 40% cluster to estimate the total number of cows in the study area.
- (ii) Estimate the variance of your estimate.
- (iii) Estimate the relative efficiency of cluster sampling compared to SRS.

Solution : We have $N = 20$, $M_0 = \sum_{i=1}^N M_i = 534$. Forty per cent of clusters are $20 \times 0.40 = 8$. We have to select 8 clusters randomly. The selected clusters are shown below :

Sl.No.	Random Number	Cluster Size, M_i	Number of cows in clusters y_{ij}	Total Y_i
1	11	16	2, 2, 2, 2, 4, 1, 1, 3, 3, 1, 4, 2, 2, 1, 4, 2	36
2	16	33	2, 1, 1, 1, 2, 2, 3, 3, 4, 4, 4, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 2, 1, 2, 2, 4, 2, 2, 4, 3, 2, 2, 2	83
3	08	34	1, 1, 4, 1, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 4, 3, 1, 2, 2, 3, 2, 2, 2, 1, 1, 2, 2, 1, 1, 2	61
4	03	22	4, 4, 4, 4, 2, 2, 1, 1, 1, 1, 1, 1, 3, 3, 3, 5, 5, 1, 1, 2, 2, 2	53
5	15	19	2, 2, 1, 2, 3, 3, 2, 2, 1, 4, 2, 2, 1, 1, 2, 4, 1, 1, 2	38
6	14	34	4, 2, 1, 1, 3, 2, 1, 1, 2, 2, 2, 1, 1, 1, 2, 2, 4, 4, 3, 3, 2, 2, 2, 1, 1, 3, 2, 1, 1, 1, 2, 2, 1, 1	64
7	04	21	1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 4, 4, 2, 2	38
8	12	18	1, 1, 2, 2, 2, 4, 2, 3, 3, 2, 2, 1, 1, 1, 1, 2, 2, 1	33
Total		$M_1 = 197$		406

Other calculations related to sample observations

Cluster Mean	1	2	3	4	5	6	7	8	Total
\bar{Y}_i	2.25	2.52	1.79	2.41	2.00	1.88	1.81	1.83	16.49
$\sum_{j=1}^{M_i} y_{ij}^2$	98	225	131	169	87	150	84	73	1017
$M_i \bar{Y}_i^2$	81.00	209.56	108.94	127.78	76.00	120.17	68.80	60.28	
s_i^2	1.1333	0.4825	0.6685	1.9628	0.6111	0.9039	0.76	0.7482	

$$\text{Here } s_i^2 = \frac{1}{M_i - 1} \left[\sum_{j=1}^{M_i} y_{ij}^2 - M_i \bar{Y}_i^2 \right]$$

The estimate of total number of cows in the study area is

$$\hat{Y}_{c1} = M_0 \bar{y}_{c1} = \frac{N}{n} \sum_{i=1}^n M_i \bar{Y}_i = \frac{N}{n} \sum_{i=1}^n \sum_{j=1}^{M_i} y_{ij} = \frac{20 \times 406}{8} = 1015.$$

We have $\bar{M} = \frac{M_0}{N} = \frac{534}{20} = 26.7.$

The average number of cows per farmer is

$$\bar{y}_{c1} = \frac{N}{nM_0} \sum_{i=1}^n M_i \bar{Y}_i = \frac{N}{nM_0} \sum_{i=1}^n \sum_{j=1}^{M_i} y_{ij} = \frac{20 \times 534}{8 \times 537} = 2.486.$$

(ii)

Sl.No.	$\frac{M_i \bar{Y}_i}{M}$	$\left(\frac{M_i \bar{Y}_i}{M} - \bar{y}_{c1}\right)^2$	$M_i(\bar{Y}_i - \bar{y}_{c1})^2$
1	1.3483	1.2943	0.8911
2	3.1086	0.3876	0.0381
3	2.2846	0.0405	16.4701
4	1.9850	0.2510	0.1271
5	1.4232	1.1295	4.4877
6	2.3970	0.0079	12.4860
7	1.4232	1.1295	9.5965
8	1.2359	1.5626	7.7460
Total		5.8029	51.8426

$$s_{bc}^2 = \frac{1}{n-1} \sum_i \left(\frac{M_i \bar{Y}_i}{M} - \bar{y}_{c1}\right)^2 = \frac{5.8029}{8-1} = 0.82898.$$

$$v(\bar{y}_{c1}) = \frac{1-f}{n} s_{bc}^2 = \frac{1-0.4}{8} \times 0.82898 = 0.0621735.$$

Hence, the estimate of variance of total cows in the study area is

$$v(\hat{Y}_{c1}) = M_0^2 v(\bar{y}_{c1}) = (534)^2 0.0621735 = 17729.1466.$$

(iii) We have $(M_1 - 1)s^2 = \sum \sum (y_{ij} - \bar{y}_{c1})^2$

$$\begin{aligned} &= \sum \sum y_{ij}^2 + M_1 \bar{y}_{c1}^2 - 2\bar{y}_{c1} \left(\sum \sum y_{ij}\right) \\ &= 1017 + 197(2.486)^2 - 2 \times 2.486 \times 406 = 4253.1306. \end{aligned}$$

$$\begin{aligned} s_w^2 &= \frac{1}{M_1 - n} \left[(M_1 - 1)s^2 - \sum M_i (\bar{Y}_i - \bar{y}_{c1})^2 \right] \\ &= \frac{1}{197 - 8} [4253.1306 - 51.8426] = 22.2290. \end{aligned}$$

$$\begin{aligned}\hat{S}^2 &= \frac{1}{M_0 - 1} \left[\frac{N - 1}{n - 1} \sum M_i (\bar{Y}_i - \bar{y}_{c1})^2 + (M_0 - N) s_w^2 \right] \\ &= \frac{1}{534 - 1} \left[\frac{20 - 1}{8 - 1} 51.8426 + (534 - 20) 22.229 \right] = 21.7006.\end{aligned}$$

Therefore, the estimate of relative efficiency of cluster sampling compared to SRS is

$$\text{r.e. (cluster)} = \frac{\hat{S}^2}{M s_{bc}^2} = \frac{21.7006}{26.7 \times 0.82898} = 98.04\%.$$

It is observed that cluster sampling is less efficient than simple random sampling.

16.7 Cluster Sampling to Estimate Proportions

The problem of estimation of proportion arises if the variable under study is qualitative in nature. Let y_{ij} be the observation of j th element in i th cluster ($i = 1, 2, \dots, N; j = 1, 2, \dots, M$), where $y_{ij} = 1$, if the population element possesses the character under study, otherwise $y_{ij} = 0$. Also, consider that

$$a_i = \sum_{j=1}^{M_i} y_{ij} = \text{number of elements in } i\text{th cluster possessing the character under study.}$$

Then the proportion of element in i th cluster possessing the character is

$$P_i = \frac{a_i}{M}.$$

The population proportion of units possessing the character is

$$P = \frac{1}{NM} \sum_{i=1}^N a_i = \frac{1}{N} \sum_{i=1}^N P_i$$

The unbiased estimator of P is $\hat{P}_c = \frac{1}{n} \sum_{i=1}^n P_i = \bar{P}$.

The variance of \hat{P}_c is

$$V(\hat{P}_c) = \frac{1-f}{n(N-1)} \sum_{i=1}^N (P_i - P)^2 \approx \frac{1-f}{nN} \sum_{i=1}^N (P_i - P)^2.$$

Again, if N is large,

$$S^2 \approx S_b^2 + S_w^2 = PQ, \text{ where } Q = 1 - P$$

and
$$S_w^2 = \frac{1}{N} \sum_{i=1}^N P_i Q_i, \quad Q_i = 1 - P_i.$$

Then the intra-class correlation coefficient is

$$\rho = 1 - \frac{M \sum_{i=1}^N P_i Q_i}{(M-1)PQ}.$$

Hence,
$$V(\hat{P}_c) = \frac{(1-f)NPQ}{(N-1)nM} [1 + (M-1)\rho].$$

If a simple random sample of nM observations is selected from the population, then the estimate of population proportion is

$$\hat{P} = \frac{1}{nM} \sum_i^N \sum_j^M y_{ij} = \frac{1}{nM} \sum_{i=1}^n a_i = \frac{1}{n} \sum_{i=1}^n P_i.$$

The variance of \hat{P} is $V(\hat{P}) = \frac{1-f}{nM(N-1)} NPQ$.

Then the relative efficiency of cluster sampling compared to simple random sampling is

$$\text{R.E. (cluster)} = \frac{N-1}{(NM-1)} \frac{NPQ}{NPQ - \sum_{i=1}^N P_i Q_i}.$$

The estimate of variance of \hat{P}_c is

$$v(\hat{P}_c) = \frac{1-f}{n(n-1)} \sum_{i=1}^n (P_i - \bar{P})^2, \quad \bar{P} = \frac{1}{n} \sum_{i=1}^n P_i.$$

Let A be the total number of units in the population possessing the characteristic under study.

The estimate of A is $\hat{A}_c = NM\hat{P}_c$.

The variance of this estimator is

$$V(\hat{A}_c) = N^2 M^2 V(\hat{P}_c) = \frac{N^2 M^2 (1-f)}{n(N-1)} \sum (P_i - \bar{P})^2.$$

The estimator of this variance is

$$v(\hat{A}_c) = \frac{N^2 M^2 (1-f)}{n(n-1)} \sum_{i=1}^n (P_i - \bar{P})^2.$$

The above estimator can also be obtained if the clusters are of unequal sizes. Let the elements in i th cluster be M_i . Then

$$P_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \frac{a_i}{M_i}.$$

The estimator of population proportion is

$$\hat{P}_{c1} = \frac{N}{nM_0} \sum_{i=1}^n a_i = \frac{N}{nM_0} \sum_{i=1}^n M_i P_i = \frac{1}{n\bar{M}} \sum M_i P_i.$$

The variance of this estimator is

$$V(\hat{P}_{c1}) = \frac{1-f}{n} S_{bc}^2,$$

where $S_{bc}^2 = \frac{1}{N-1} \sum_{i=1}^N \left(\frac{M_i \bar{Y}_i}{\bar{M}} - \bar{Y} \right)^2 = \frac{N}{M_0^2 (N-1)} \left[\sum_{i=1}^N M_i^2 P_i^2 - \frac{\left(\sum_{i=1}^N M_i P_i \right)^2}{N} \right]$

The estimator of this variance is

$$v(\hat{P}_{c1}) = \frac{(1-f)N^2}{n(n-1)M_0^2} \left[\sum^N M_i^2 P_i^2 - \frac{\left(\sum^N M_i P_i \right)^2}{n} \right]$$

Example 16.4 : There are 105 villages in a police station area. Each village is considered a cluster. From the population clusters 18 clusters are selected by SRS scheme. The number of babies under age 5 years are investigated to study the proportion of babies who are given polio vaccine. The total number of babies under age 5 years in the selected villages and the number of babies who are given polio vaccine are shown below : The total number of babies in that police station area are $M_0 = 4882$.

Sl. No. of Clusters	Babies in Clusters M_i	Babies under Polio vaccine a_i	Proportion of Babies who are given polio vaccine, $P_i = \frac{a_i}{M_i}$
1	68	48	0.706
2	46	36	0.783
3	25	12	0.480
4	38	28	0.737
5	52	32	0.615
6	44	30	0.682
7	80	32	0.400
8	46	20	0.435
9	55	40	0.727
10	70	55	0.786
y11	28	15	0.536
12	56	42	0.750
13	60	40	0.667
14	32	25	0.781
15	46	20	0.435
16	66	28	0.424
17	42	24	0.571
18	58	40	0.690

- (i) Estimate the proportion of babies in the police station area who are given polio vaccine.
(ii) Also, estimate the variance of your estimate.

Solution : (i) The estimate of population proportion is

$$\hat{P}_c = \frac{N}{nM_0} \sum^n a_i = \frac{105 \times 567}{18 \times 488} = 0.677$$

- (ii) The estimate of variance of \hat{P}_c

$$\begin{aligned} v(\hat{P}_c) &= \frac{(1-f)N^2}{n(n-1)M_0^2} \left[\sum a_i^2 - \frac{(\sum a_i)^2}{n} \right], \quad f = \frac{n}{N} = 0.17 \\ &= \frac{(1-0.17)(105)^2}{18(18-1)(4882)^2} \left[20075 - \frac{(567)^2}{18} \right] = 0.00278 \end{aligned}$$

Chapter 17

Two-Stage Sampling

17.1 Definition

It has already been discussed that cluster samples are drawn when population units are divided into several groups (clusters). The number of elements in each cluster may be large. But cluster sampling principle states that some clusters are selected randomly and all elements within a cluster are investigated. If clusters are of larger sizes, the cost and time required to finish the survey are increased. To obviate the problem, some clusters are randomly selected and from each selected cluster some elements are randomly selected for a survey. Since ultimate sampling units are selected at two stages, the sampling technique is known as two-stage sampling. For example, let us consider the survey to estimate the total production of jute in a district. The entire district is divided into several administrative blocks. In each block there are many farmers who grow jute in their cultivable lands. Here blocks can be considered as clusters, and farmers within a block are considered as ultimate sampling units. Some blocks can be selected at random and from each selected block some jute growers can be selected at random.

The process of selection of jute growers as mentioned above is known as *two-stage sampling*.

In two-stage sampling the clusters are known as first-stage units (fsu) and elements within a cluster are called ultimate sampling units or second stage sampling units (ssu). The first stage units, also called primary sampling units, may be of equal size or of unequal sizes. The number of secondary units may also be equal or unequal. Whatever be the number of secondary units, equal or unequal, the list of the units is not essential to select the sample. However, the list of the primary units must be known so that some of the primary units can be selected at random. If the list of the secondary units is not known, it can be prepared for the selected primary units during pilot survey.

17.2 Advantage and Use of Two-Stage Sampling

The two-stage sampling is a mixture of cluster sampling and random sampling, where clusters are usually selected by SRS scheme. Hence, this sampling is expected to be more efficient than cluster sampling but less efficient than random sampling. This sampling is equivalent to one-stage sampling if the second stage selected units are $m = M$, where M is the size of second stage units in a primary unit. The main advantage of this sampling is that the sample can be selected even if the frame of all population units is not available. The list of the second stage units can be prepared after selecting the primary units.

This sampling scheme is used in agricultural as well as in social survey work. For example, let us consider the survey to estimate the proportion of children who are given BCG vaccine. If the survey is confined in a municipality, the study area is divided into streets, where each street can be considered as a cluster. In each cluster some families are living. The families can be considered as second stage units. Hence, as a primary units some streets are selected at random and from the selected streets some families are selected at random. The selected families are investigated for the survey purpose. The sampling design mentioned above to select families is a two-stage sampling technique.

17.3 Estimation of Parameter in Two-Stage Sampling

Let us first consider the estimation procedure with equal second stage units in every first stage unit. Let there be NM elements in a population, where population units are divided into N clusters each of size M . Consider that n clusters out of N are selected under SRSWR scheme and from each selected cluster m ($m < M$), second stage units are selected by SRSWR scheme.

Let y_{ij} be the value of the variable under study of j -th second stage unit within i -th first stage unit ($i = 1, 2, \dots, N$; $j = 1, 2, \dots, M$). Then the mean per element of i -th first stage unit is

$$\bar{Y}_i = \frac{1}{M} \sum_{j=1}^M y_{ij}.$$

The population mean per element is

$$\bar{Y} = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M y_{ij} = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i.$$

The variance of the means of first stage units is

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2.$$

The variance of the observations within the i -th first stage unit is

$$S_{wi}^2 = \frac{1}{M-1} \sum_{j=1}^M (y_{ij} - \bar{Y}_i)^2.$$

The variance of the observations within the first stage units is

$$S_w^2 = \frac{1}{N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{Y}_i)^2 = \frac{1}{N} \sum_{i=1}^N S_{wi}^2.$$

The mean of the selected second stage units from i -th first stage unit is

$$\bar{y}_i = \frac{1}{m} \sum_{j=1}^m y_{ij}.$$

The mean per element of sample observations is

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

Theorem : In two-stage sampling, if n primary units and from each selected primary unit if m secondary units are selected by SRSWOR scheme, then sample mean \bar{y} is an unbiased estimator of \bar{Y} and the variance of this estimator is

$$V(\bar{y}) = \frac{1-f}{n} S_b^2 + \frac{M-m}{M} \frac{S_w^2}{nm}.$$

Proof: We know $\bar{y} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij} = \frac{1}{n} \sum \bar{y}_i$.

Since second stage units are selected by SRSWOR scheme, \bar{y}_i is an unbiased estimator of $\bar{Y}_i [E_2(\bar{y}_i) = \bar{Y}_i]$. We can write,

$$E(\bar{y}) = E_1 \left[E_2 \left(\frac{1}{n} \sum_{i=1}^n \bar{y}_i \right) \right] = E_1 \frac{1}{n} \left[\sum_{i=1}^n E_2(\bar{y}_i) \right].$$

Here E_1 and E_2 are used to find expected values for first stage sampling and second stage sampling, respectively. Therefore,

$$\begin{aligned} E(\bar{y}) &= E_1 \left[\frac{1}{n} \sum_{i=1}^n \bar{Y}_i \right] = \frac{1}{n} \sum_{i=1}^n \frac{1}{n} \sum_{i=1}^N \bar{Y}_i \\ & \qquad \qquad \qquad \text{[according to the property of SRSWR scheme]} \\ &= \bar{Y}. \end{aligned}$$

Hence, sample mean \bar{y} is an unbiased estimator of \bar{Y} . The variance of \bar{y} is

$$V(\bar{y}) = V_1[E_2(\bar{y})] + E_1[V_2(\bar{y})].$$

Here also V_1 and V_2 are used for variances in case of first stage and second stage sampling, respectively. We have

$$E_2(\bar{y}) = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i.$$

Since \bar{Y}_i is sample mean under SRSWR scheme,

$$V_1[E_2(\bar{y})] = V_1 \left[\frac{1}{n} \sum \bar{Y}_i \right] = \frac{N-n}{Nn} S_b^2.$$

Again, $V_2(\bar{y}) = V_2 \left(\frac{1}{n} \sum \bar{y}_i \right) = \frac{1}{n^2} \sum_{i=1}^n V_2(\bar{y}_i)$.

But second stage units are also selected under SRSWR scheme. Hence,

$$V_2(\bar{y}_i) = \frac{M-m}{mM} S_{wi}^2 \quad \text{and} \quad V_2(\bar{y}) = \frac{1}{n^2} \sum_{i=1}^n \frac{M-m}{mM} S_{wi}^2.$$

Now, $E_1 V_2(\bar{y}) = \frac{1}{N} \sum_{i=1}^N V_2(\bar{y})$

$$= \frac{1}{N} \frac{M-m}{n^2 m M} \sum_{i=1}^n \sum_{i=1}^N S_{wi}^2 = \frac{M-m}{mM} \frac{S_w^2}{n}.$$

$\therefore V(\bar{y}) = \frac{N-n}{Nn} S_b^2 + \frac{M-m}{mM} \frac{S_w^2}{n}$

$$= \frac{1-f}{n} S_b^2 + \frac{1-f_1}{mn} S_w^2, \quad \text{where } f = \frac{n}{N}, \quad f_1 = \frac{m}{M}$$

$$= \frac{S_b^2}{n} + \frac{S_w^2}{nm}, \quad \text{if } f \text{ and } f_1 \text{ are negligible.}$$

Corollary : In two-stage sampling if first stage units are selected under SRSWR scheme and second stage units are selected under SRSWOR scheme, the sampling variance of sample mean is

$$V(\bar{y}) = \frac{S_b^2}{n} + (1 - f_1) \frac{S_w^2}{nm}$$

Corollary : In two-stage sampling if first stage units are selected under SRSWOR scheme and if all units in a primary units are selected in the sample, then

$$V(\bar{y}) = \frac{1 - f}{n} S_b^2 \quad (\because m = M).$$

Theorem : In two-stage sampling, if n first stage units are selected and from each selected first stage unit m second stage units are selected under SRSWOR scheme, then the estimator of variance of sample mean \bar{y} is given by

$$v(\bar{y}) = \frac{1 - f}{n} s_b^2 + \frac{f(1 - f_1)}{nm} s_w^2,$$

where $f = \frac{n}{N}$, $f_1 = \frac{m}{M}$, $s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2$ and $s_w^2 = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2$.

Proof : We have $(n-1)s_b^2 = \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 = \sum \bar{y}_i^2 - n\bar{y}^2 = n\bar{y}^2$.

Now, $(n-1)E_2(s_b^2) = E_2 \sum \bar{y}_i^2 - nE_2(\bar{y}^2)$.

Again, $E_2 \left[\sum_{i=1}^n \bar{y}_i^2 \right] = \left[\sum_{i=1}^n \bar{y}_i^2 + \sum_{i=1}^n \frac{M-m}{mM} S_{wi}^2 \right]$

$$E_1 \left[E_2 \left(\sum_{i=1}^n \bar{y}_i^2 \right) \right] = \frac{n}{N} \left[\sum_{i=1}^N \bar{Y}_i^2 + \frac{N(M-m)}{mM} S_w^2 \right]$$

$$V(\bar{y}) = E_2(\bar{y}^2) - \bar{Y}^2$$

$$nE_2(\bar{y}^2) = (1-f)S_b^2 + \frac{M-m}{mM} S_w^2 + n\bar{Y}^2.$$

Hence, putting the values of $E_2(\sum \bar{y}_i^2)$ and $E(\bar{y}^2)$ and on simplification, we get

$$E_2(s_b^2) = S_b^2 + \frac{M-m}{mM} S_w^2.$$

We know that for any value of i $E(s_i^2) = S_{wi}^2$, since s_i^2 is the variance of simple random sample observations. For different values of i , we have

$$E \left(\frac{1}{n} \sum s_i^2 \right) = \frac{1}{N} \sum_{i=1}^N S_{wi}^2 = S_w^2.$$

Hence, the unbiased estimator of S_w^2 is s_w^2 , where

$$s_w^2 = \frac{\sum_{i=1}^n (m-1)s_i^2}{n(m-1)} \quad \text{and} \quad s_i^2 = \frac{1}{m-1} \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2.$$

The unbiased estimator of S_b^2 is

$$\hat{S}_b^2 = s_b^2 - \left(\frac{M-m}{mM} \right) s_w^2.$$

Therefore, $v(\bar{y}) = \frac{1-f}{n} s_b^2 + \frac{f(1-f_1)}{nm} s_w^2$.

If $m = M$, the variance formula becomes

$$v(\bar{y}) = \frac{1-f}{n} s_b^2 \quad (\because f_1 = 1).$$

Again, if $f = \frac{n}{N}$ is negligible,

$$v(\bar{y}) = \frac{s_b^2}{n} + \frac{f(1-f_1)}{nm} s_w^2.$$

If f is negligible and $f_1 = 1$, then $v(\bar{y}) = \frac{s_b^2}{n}$.

Corollary : If n clusters are selected under SRSWR and if m elements are selected from each selected cluster under SRSWOR, then

$$v(\bar{y}) = \frac{s_b^2}{n} + \frac{1-f_1}{Nm} s_w^2.$$

Corollary : In two-stage sampling, if n clusters are selected under SRSWOR and m units from each selected clusters are selected under SRSWOR, then the estimator of population total is given by

$$\hat{Y} = NM\bar{y}.$$

The estimate of variance of this estimator is

$$v(\hat{Y}) = N^2 M^2 (1-f) \frac{s_b^2}{n} + N^2 M^2 \frac{f(1-f_1)}{nm} s_w^2.$$

Example 17.1 : The rural administrative unit under a police station is divided into 60 clusters in such a way that in each cluster there are 10 farmers living nearby and they have milking cows. Twenty per cent of the clusters are selected by simple random sampling without replacement (SRSWOR) and 50% farmers of each selected cluster are also selected by SRSWOR technique to estimate the total milk production in the study area. The population data of milk production (in kg) are given below. Estimate the total milk production and also estimate its standard error. Compare two-stage sampling with simple random sampling.

Milk production of different farmers in clusters

Sl. No. of clusters	Milk production (y_{ij} kg) of farmers in a day									
	1	2	3	4	5	6	7	8	9	10
1	5.0	6.2	7.4	10.2	2.8	3.0	1.5	1.5	2.0	3.0
2	6.0	3.0	4.5	2.5	3.0	2.0	7.5	6.5	5.0	4.0
3	4.0	6.0	5.0	4.0	4.0	6.2	9.5	10.0	11.2	10.2
4	5.0	3.0	3.0	5.0	5.0	5.0	4.5	9.0	8.0	7.5
5	8.0	4.5	5.0	3.0	3.5	4.0	2.0	1.5	1.0	6.0
6	7.5	3.0	8.0	3.0	4.0	5.0	5.0	5.0	2.5	2.4

Milk production of different farmers in clusters

Sl. No. of clusters	Milk production (y_{ij} kg) of farmers in a day									
	1	2	3	4	5	6	7	8	9	10
7	1.5	1.5	1.5	2.0	6.0	7.0	8.0	7.5	5.0	5.0
8	4.0	3.5	3.0	6.0	7.5	10.0	8.6	7.2	6.0	4.0
9	5.0	5.0	6.0	8.0	9.2	4.2	1.5	2.0	2.0	3.0
10	6.0	1.8	1.5	1.5	4.6	4.5	5.0	6.0	7.5	7.5
11	4.0	4.5	5.0	3.5	3.0	3.2	3.0	3.0	8.0	10.0
12	8.0	9.0	10.0	10.0	8.0	4.5	2.0	2.0	4.0	2.5
13	3.0	2.0	2.5	2.0	1.5	10.5	12.0	6.0	6.0	2.0
14	8.0	7.5	4.2	4.0	4.0	6.5	1.5	1.5	8.0	9.0
15	2.0	2.5	12.6	1.8	2.0	8.6	9.5	4.5	4.0	4.0
16	8.0	9.0	10.2	11.6	7.0	1.2	2.4	5.6	6.0	6.0
17	5.0	6.2	7.5	8.0	1.5	2.0	3.0	3.5	6.0	3.2
18	6.0	2.5	4.5	6.0	5.0	5.0	4.9	6.0	2.0	1.0
19	8.1	7.2	6.2	7.0	2.0	2.0	2.0	3.0	4.5	5.0
20	7.0	6.0	2.4	1.5	2.0	4.6	5.0	5.0	4.0	3.2
21	1.8	1.9	2.0	4.6	7.2	7.0	6.5	3.0	3.5	3.0
22	2.0	1.4	2.6	6.8	7.0	6.0	4.0	3.0	2.0	8.0
23	4.6	1.5	7.6	10.0	9.2	3.5	2.0	2.0	2.5	3.5
24	1.5	7.0	6.8	4.3	2.5	2.5	4.6	5.0	5.0	6.3
25	4.0	5.2	2.6	2.7	1.8	9.6	8.7	2.0	1.8	1.5
26	3.0	4.8	5.6	2.5	1.5	1.0	4.8	4.6	6.0	2.4
27	4.6	5.8	2.5	3.0	9.6	9.2	4.2	5.0	3.0	4.6
28	5.0	1.5	2.0	4.6	3.0	6.2	4.0	3.5	4.0	2.0
29	6.2	9.2	4.0	4.2	1.8	7.2	4.2	4.0	6.0	6.5
30	4.2	7.2	1.8	3.8	1.5	6.5	1.8	6.2	4.0	1.5
31	5.0	6.2	2.0	1.5	7.6	9.0	4.0	4.8	1.8	1.0
32	6.2	4.0	2.8	7.6	8.0	10.5	3.0	3.0	3.0	3.5
33	4.6	5.2	1.0	10.2	3.0	3.5	4.0	4.0	6.6	7.2
34	4.0	1.2	1.5	6.4	3.0	4.0	4.5	4.8	2.6	2.8
35	4.8	6.0	7.2	8.6	4.0	1.0	1.5	10.0	3.0	5.0
36	4.2	6.2	4.0	3.2	5.0	6.2	6.0	9.0	8.2	10.0
37	1.5	3.0	8.0	4.2	8.0	4.6	7.2	5.2	6.0	5.2
38	5.2	10.2	4.6	5.0	1.2	1.0	5.2	1.5	5.0	3.0
39	10.0	4.5	6.2	1.8	2.0	9.6	4.6	6.7	5.0	6.8
40	7.2	4.0	8.7	4.8	9.0	7.2	10.0	8.0	4.6	9.0
41	3.0	7.6	4.8	6.4	7.5	8.0	3.0	5.8	10.6	7.4
42	5.5	8.7	2.0	5.6	6.2	6.4	9.7	7.2	4.2	4.0
43	5.0	2.3	4.6	9.8	7.0	6.2	3.5	4.7	8.0	2.6
44	2.0	2.8	7.6	4.3	3.0	3.5	8.7	6.2	6.4	10.5
45	6.0	4.8	3.5	3.0	6.8	10.0	4.5	5.0	2.0	4.6
46	3.4	2.0	1.5	1.8	7.8	4.8	10.7	3.0	4.8	6.2

Milk production of different farmers in clusters

Sl. No. of clusters	Milk production (y_{ij} kg) of farmers in a day									
	1	2	3	4	5	6	7	8	9	10
47	5.0	4.8	7.6	9.2	3.5	1.5	1.5	1.5	6.4	7.0
48	4.0	2.2	10.8	4.6	1.5	2.8	4.6	7.0	1.8	2.6
49	2.0	1.8	7.4	6.3	1.3	1.4	10.6	9.8	6.4	5.0
50	8.7	4.2	3.5	4.6	6.7	1.8	2.5	2.5	4.6	5.8
51	2.2	8.0	9.7	5.2	6.0	4.8	3.0	3.0	4.2	5.0
52	3.5	4.6	2.8	5.6	8.0	2.0	1.8	4.0	6.7	3.8
53	3.0	8.7	10.6	5.2	4.6	7.0	6.8	2.8	9.7	4.2
54	6.2	4.6	8.7	4.0	3.0	4.8	5.0	6.7	4.8	3.0
55	1.8	9.0	7.2	3.5	5.0	6.0	6.0	4.0	3.2	9.2
56	8.2	3.0	4.6	2.0	1.8	1.5	4.6	10.5	3.6	6.0
57	7.2	4.0	2.8	4.6	3.0	3.2	4.8	5.0	2.8	4.6
58	9.6	3.2	4.8	5.0	1.6	4.7	7.6	3.0	3.0	3.8
59	4.0	5.0	6.2	7.4	3.0	3.5	4.0	2.8	4.6	6.0
60	7.2	4.8	1.8	4.0	5.5	3.0	6.2	2.8	3.5	4.0

Solution : We have $N = 60$, $M = 10$. We need to select $n = 60 \times 0.2 = 12$ primary units and from each selected primary unit we need to select $m = 10 \times 0.5 = 5$ second stage units. The selected primary units are shown below :

Sl. No.	Random	Milk production (y_{ij}) of farmers									
	No.	1	2	3	4	5	6	7	8	9	10
1	51	2.2	8.0	9.7	5.2	6.0	4.8	3.0	3.0	4.2	5.0
2	16	8.0	9.0	10.2	11.6	7.0	1.2	2.4	5.6	6.0	6.0
3	48	4.0	2.2	10.8	4.6	1.5	2.8	4.6	7.0	1.8	2.6
4	23	4.6	1.5	7.6	10.0	9.2	3.5	2.0	2.0	2.5	3.5
5	35	4.8	6.0	7.2	8.6	4.0	1.0	1.5	10.0	3.0	5.0
6	11	4.0	4.5	5.0	3.5	3.0	3.2	3.0	3.0	8.0	10.0
7	34	4.0	1.2	1.5	6.4	3.0	4.0	4.5	4.8	2.6	2.8
8	04	5.0	3.0	3.0	5.0	5.0	5.0	4.5	9.0	8.0	7.5
9	12	8.0	9.0	10.0	10.0	8.0	4.5	2.0	2.0	4.0	2.5
10	10	6.0	1.8	1.5	1.5	4.6	4.5	5.0	6.0	7.5	7.5
11	25	4.0	5.2	2.6	2.7	1.8	9.6	8.7	2.0	1.8	1.5
12	24	1.5	7.0	6.8	4.3	2.5	2.5	4.6	5.0	5.0	6.3

From the selected primary units the ultimate selected sample observations are shown below :

Sl. No.	Milk production (y_{ij}) of farmers in the sample					Mean \bar{y}_i	$\sum y_{ij}^2$	$\sum y_{ij}^2 - m\bar{y}_i^2$
	1	2	3	4	5			
1	2.2	4.8	3.0	9.7	6.0	5.14	166.97	33.3255
2	8.0	11.6	9.0	1.2	6.0	7.16	317.00	60.672
3	1.5	4.6	2.8	1.8	2.2	2.58	39.33	6.048
4	4.6	2.0	2.5	9.2	2.0	4.06	120.05	37.632
5	1.0	1.5	7.2	4.8	5.0	3.90	103.13	27.080
6	8.0	3.5	4.0	3.0	3.0	4.30	110.25	17.800
7	1.5	4.5	6.4	4.0	1.2	3.52	80.90	18.948
8	3.0	5.0	5.0	5.0	8.0	5.20	148.00	12.800
9	10.0	8.0	10.0	2.0	9.0	7.80	349.00	44.800
10	5.0	1.5	1.8	4.5	4.6	3.48	71.90	11.348
11	8.7	2.6	1.8	1.8	1.5	3.28	91.18	37.388
12	5.0	5.0	4.6	4.3	2.5	3.28	95.90	4.308
Total						54.7		

The mean milk production per farmer is

$$\bar{y} = \frac{1}{n} \sum \bar{y}_i = \frac{54.7}{12} = 4.56 \text{ kg.}$$

The estimate of total milk production in the study area is

$$\hat{Y} = NM\bar{y} = \frac{60 \times 10}{n} \sum_{i=1}^n \bar{y}_i = 2735.00 \text{ kg.}$$

$$\begin{aligned} \text{We have } s_b^2 &= \frac{1}{n-1} \sum (\bar{y}_i - \bar{y})^2 = \frac{1}{n-1} \left[\sum \bar{y}_i^2 - n\bar{y}^2 \right] \\ &= \frac{1}{12-1} [275.9828 - 12(4.56)^2] = 2.4054. \end{aligned}$$

$$\begin{aligned} s_w^2 &= \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2 = \frac{1}{n(m-1)} \sum \left[\sum y_{ij}^2 - m\bar{y}_i^2 \right] \\ &= \frac{1}{12(5-1)} \times 312.1495 = 6.5031. \end{aligned}$$

The estimated variance of \hat{Y} is

$$\begin{aligned} v(\hat{Y}) &= N^2 M^2 v(\bar{y}) = N^2 M^2 (1-f) \frac{s_b^2}{n} + N^2 M^2 + f(1-f) \frac{s_w^2}{nm} \\ &= (60)^2 (10)^2 \left[(1-0.2) \frac{2.4054}{12} + 0.2(1-0.5) \frac{6.5031}{12 \times 5} \right] \\ &= 360000 [0.16036 + 0.0108385] = 61631.46. \end{aligned}$$

Hence, the standard error of the estimate of total milk production is

$$\text{s.e.}(\hat{Y}) = \sqrt{v(\hat{Y})} = \sqrt{61631.46} = 248.26.$$

The comparison of two stage sampling with simple random sampling is done considering 60 observations in the sample drawn under SRSWOR scheme. In such a case,

$$v(\bar{y}) = \left(\frac{1}{nm} - \frac{1}{NM} \right) s^2,$$

$$\begin{aligned} \text{where } s^2 &= \frac{1}{NM-1} \left[M(N-1)s_b^2 + \left\{ N(M-1) - (M-m) \frac{N-1}{m} s_w^2 \right\} \right] \\ &= \frac{1}{60 \times 10 - 1} \left[10(60-1)2.4054 + \left\{ 60(10-1) - (10-5) \frac{(60-1)6.5031}{5} \right\} \right] \\ &= 2.3386. \end{aligned}$$

$$\text{Hence, } v(\bar{y}) = \left(\frac{1}{12 \times 5} - \frac{1}{60 \times 10} \right) 2.3386 = 0.035079.$$

Again, $v(\bar{y})$ two-stage = 0.1712 > $v(\bar{y})$ random.

Hence, simple random sampling is more efficient than two-stage sampling.

17.4 Estimation of Parameters in Two-Stage Sampling with Unequal First Stage Units

Consider that i -th ($i = 1, 2, \dots, N$) first stage unit has M_i elements (second stage unit) such that total elements in the population are $M_0 = \sum_{i=1}^N M_i$. Also consider that from N primary units n are selected under SRSWOR scheme and from i -th selected primary units m_i second stage units are selected at random by simple random sampling without replacement.

The population mean of M_0 elements is

$$\bar{Y} = \frac{1}{M_0} \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij} = \frac{1}{NM} \sum_{i=1}^N M_i \bar{Y}_i = \frac{1}{N} \sum_{i=1}^N u_i \bar{Y}_i.$$

$$\text{where } u_i = \frac{M_i}{M}, \quad M = \frac{1}{N} M_0, \quad \bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}.$$

$$\text{Again, } \bar{Y}_N = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i.$$

The mean of second stage units selected from i -th first stage unit is

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}.$$

Total sample observations are

$$m_0 = \sum_{i=1}^n m_i.$$

Now, the estimators of population mean \bar{Y} are

$$\bar{y}_1 = \frac{1}{n} \sum_{i=1}^n u_i \bar{y}_i = \frac{1}{nM} \sum_{i=1}^n M_i \bar{y}_i, \quad \bar{y}_2 = \frac{1}{n} \sum_{i=1}^n \bar{y}_i.$$

Moreover, a ratio estimator of population mean is also available. For this, let us define a variable X with values x_{ij} = values of correlated auxiliary variable corresponding to j -th second stage unit within i -th first stage unit. Then

$$\bar{X} = \frac{1}{M_0} \sum_{i=1}^N \sum_{j=1}^{M_i} x_{ij} = \frac{1}{M_0} \sum_{i=1}^N M_i \bar{X}_i = \frac{1}{NM} \sum_{i=1}^N M_i \bar{X}_i = \frac{1}{N} \sum_{i=1}^N u_i \bar{X}_i,$$

where $\bar{X}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} x_{ij}$.

Let $\bar{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}$ and $\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n u_i \bar{x}_i = \frac{1}{nM} \sum_{i=1}^n M_i \bar{x}_i$.

Then the ratio estimator of population mean \bar{Y} is

$$\bar{y}_{1R} = \frac{\bar{y}_1}{\bar{x}_1} \bar{X}.$$

Theorem : In two-stage sampling in case of unequal first stage units if simple random sampling without replacement is used at each stage, then \bar{y}_1 is an unbiased estimator of population mean \bar{Y} . The variance of \bar{y}_1 is

$$V(\bar{y}_1) = \frac{N-n}{Nn} S_{1b}^2 + \frac{1}{nN} \sum_{i=1}^N u_i^2 \frac{M_i - m_i}{m_i M_i} S_i^2.$$

Here $S_{1b}^2 = \frac{1}{N-1} \sum_{i=1}^N (u_i \bar{Y}_i - \bar{Y})^2$

and $S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2$.

Proof : Given $\bar{y}_1 = \frac{1}{n} \sum_{i=1}^n u_i \bar{y}_i = \frac{1}{NM} \sum_{i=1}^n M_i \bar{y}_i$.

$$\begin{aligned} E(\bar{y}_1) &= E_1 \left[\frac{1}{nM} \sum_{i=1}^n E_2(M_i \bar{y}_i) \right] = E_1 \left[\frac{1}{nM} \sum_{i=1}^n M_i \bar{Y}_i \right] \\ &= E_1 \left[\frac{1}{n} \sum_{i=1}^N u_i \bar{Y}_i \right] = \frac{1}{n} \sum_{i=1}^n \frac{1}{N} \sum_{i=1}^N u_i \bar{y}_i = \bar{Y}. \end{aligned}$$

Hence, \bar{y}_1 is an unbiased estimator of \bar{Y} .

$$\begin{aligned} V(\bar{y}_1) &= V_1 E_2 \left(\frac{\bar{y}_1}{n} \right) + E_1 V_2 \left(\frac{\bar{y}_1}{n} \right) \\ &= V_1 \left[E_2 \frac{1}{n} \sum_{i=1}^n u_i \bar{y}_i \right] + E_1 \left[\frac{1}{n^2} \sum_{i=1}^n u_i^2 V_2(\bar{y}_i) \right] \\ &= \frac{1-f}{n} S_{1b}^2 + E_1 \left[\frac{1}{n^2} \sum_{i=1}^n u_i^2 \frac{M_i - m_i}{m_i M_i} S_i^2 \right] \end{aligned}$$

$$\begin{aligned}
 &= \frac{1-f}{n} S_{1b}^2 + \frac{1}{nN} \sum_{i=1}^N u_i^2 \frac{M_i - m_i}{m_i M_i} S_i^2 \\
 &= \frac{1-f}{n} S_{1b}^2 + \frac{1}{nN} \sum_{i=1}^N \frac{M_i^2(1-f_i)}{M^2} \frac{S_i^2}{m_i}, \text{ where } f_i = \frac{m_i}{M_i}.
 \end{aligned}$$

Theorem : In two-stage sampling in case of unequal first-stage units if sample is selected by SRSWOR scheme at each stage, then \bar{y}_2 is biased estimator of \bar{Y} . The bias and variance of \bar{y}_2 are respectively

$$\text{Bias } (\bar{y}_2) = -\frac{1}{N\bar{M}} \sum_{i=1}^N (M_i - \bar{M})(\bar{Y}_i - \bar{Y}_N)$$

$$V(\bar{y}_2) = \frac{N-n}{nN} S_{2b}^2 + \frac{1}{nN} \sum_{i=1}^N \frac{M_i - m_i}{m_i M_i} S_i^2,$$

where $S_{2b}^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y}_N)^2.$

Proof : Given $\bar{y}_2 = \frac{1}{n} \sum_{i=1}^n y_i.$

$$E(\bar{y}_2) = E_1 \left[\frac{1}{n} \sum_{i=1}^n E_2(\bar{y}_i) \right] = E_1 \left[\frac{1}{n} \sum_{i=1}^n \bar{Y}_i \right] = \frac{1}{N} \sum_{i=1}^N Y_i = \bar{Y}_N \neq \bar{Y}.$$

Hence, \bar{y}_2 is not an unbiased estimator of \bar{Y} .

The bias of \bar{y}_2 is

$$\begin{aligned}
 \text{Bias } (\bar{y}_2) &= \bar{Y}_N - \bar{Y} = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i - \bar{Y} \\
 &= \frac{1}{N} \sum_{i=1}^N \bar{Y}_i - \frac{1}{N\bar{M}} \sum_{i=1}^N M_i \bar{Y}_i = -\frac{1}{N\bar{M}} \left[\sum_{i=1}^N M_i \bar{Y}_i - \bar{M} \sum_{i=1}^N \bar{Y}_i \right] \\
 &= -\frac{1}{N\bar{M}} \sum_{i=1}^N (M_i - \bar{M})(\bar{Y}_i - \bar{Y}_N).
 \end{aligned}$$

Again, $V(\bar{y}_2) = V_1 E_2(\bar{y}_2/n) + E_1[V_2(\bar{y}_2/n)]$

$$\begin{aligned}
 &= V_1 \left(E_2 \frac{1}{n} \sum_{i=1}^n \bar{y}_i \right) + E_1 \left[\frac{1}{n^2} \sum_{i=1}^n V(\bar{y}_i) \right] \\
 &= \frac{N-n}{nN} S_{2b}^2 + E_1 \left[\frac{1}{n^2} \sum_{i=1}^N \frac{M_i - m_i}{m_i M_i} S_i^2 \right] \\
 &= \frac{N-n}{nN} S_{2b}^2 + \frac{1}{nN} \sum_{i=1}^N \frac{M_i - m_i}{m_i M_i} S_i^2.
 \end{aligned}$$

Therefore, the mean square error (M.S.E.) of \bar{y}_2 is

$$\text{MSE}(\bar{y}_2) = \frac{N-n}{nN} S_{2b}^2 + \frac{1}{nN} \sum_{i=1}^N \frac{M_i - m_i}{m_i M_i} S_i^2 + (\bar{Y}_N - \bar{Y})^2.$$

Corollary : In two-stage sampling in case of unequal first stage units the unbiased estimator of population mean is

$$\hat{Y} = \bar{y}_2 + \frac{N-1}{NM} \frac{1}{n-1} \sum_{i=1}^n (M_i - \bar{M}_1)(\bar{y}_i - \bar{y}_2), \text{ where } \bar{M}_1 = \frac{1}{n} \sum_{i=1}^n M_i.$$

Corollary : In two-stage sampling in case of unequal first stage units the unbiased estimator of population total is

$$\hat{Y} = M_0 \bar{y}_1 = \frac{N}{n} \sum_{i=1}^n M_i \bar{y}_i.$$

The variance of this estimator is

$$V(\hat{Y}) = M_0^2 V(\bar{y}_1) = \frac{N^2 M_0^2 (1-f) S_{1b}^2}{n} + \frac{N}{n} \sum_{i=1}^N \frac{M_i^2 (1-f_i) S_i^2}{m_i}.$$

Theorem : In two-stage sampling in case of unequal first stage units the ratio estimator \bar{y}_{1R} is not an unbiased estimator of \bar{Y} . The mean square error of \bar{y}_{1R} is

$$\begin{aligned} \text{MSE}(\bar{y}_{1R}) &= \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N u_i^2 (\bar{Y}_i - R\bar{X}_i)^2 + \frac{1}{nN} \sum_{i=1}^N u_i^2 \frac{M_i - m_i}{m_i M_i} D_i^2 \\ &= \frac{N-n}{nN} (S_{1by}^2 - 2RS_{1bxy} + R^2 S_{1bx}^2) + \frac{1}{nN} \sum_{i=1}^N u_i^2 \frac{M_i - m_i}{m_i M_i} D_i^2, \end{aligned}$$

where $D_i^2 = S_{iy}^2 - 2RS_{ixy} + R^2 S_{ix}^2$.

$$S_{1by}^2 = \frac{1}{N-1} \sum_{i=1}^N (u_i \bar{Y}_i - \bar{Y})^2, \quad S_{1bx}^2 = \frac{1}{N-1} \sum_{i=1}^N (u_i \bar{X}_i - \bar{X})^2.$$

$$S_{1bx} = \frac{1}{N-1} \sum_{i=1}^N (u_i \bar{Y}_i - \bar{Y})(u_i \bar{X}_i - \bar{X}), \quad S_{iy}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2.$$

$$S_{ix}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (x_{ij} - \bar{X}_i)^2, \quad S_{ixy} = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (x_{ij} - \bar{X}_i)(y_{ij} - \bar{Y}_i).$$

Proof : According to the property of ratio estimator, we can write,

$$E(\bar{y}_{1R}) \approx \bar{Y} \left[1 + \frac{V(\bar{x}_1)}{\bar{X}^2} - \frac{\text{Cov}(\bar{x}_1, \bar{y}_1)}{\bar{X}\bar{Y}} \right] R.$$

Hence, \bar{y}_{1R} is not an unbiased estimator of \bar{Y} . The relative bias of \bar{y}_{1R} is

$$B \approx \left[\frac{V(\bar{x}_1)}{\bar{X}^2} - \frac{\text{Cov}(\bar{x}_1, \bar{y}_1)}{\bar{X}\bar{Y}} \right].$$

Again,
$$\text{Cov}(\bar{x}_1, \bar{y}_1) = \frac{N-n}{nN} S_{1bxy}^2 - \frac{1}{nN} \sum_{i=1}^N u_i^2 \frac{M_i - m_i}{m_i M_i} S_{ixy}^2.$$

Therefore,
$$B \approx \frac{N-n}{Nn} \left(\frac{S_{ibx}^2}{\bar{X}^2} - \frac{S_{ibxy}^2}{\bar{X}\bar{Y}} \right) + \frac{1}{nN} \sum_{i=1}^N u_i^2 \frac{M_i - m_i}{m_i M_i} \left(\frac{S_{ix}^2}{\bar{X}^2} - \frac{S_{ixy}}{\bar{X}\bar{Y}} \right).$$

Hence, up to first degree approximation,

$$\begin{aligned} \text{MSE}(\bar{y}_{1R}) &\approx \frac{N-n}{nN} \frac{1}{N-1} \sum u_i^2 (\bar{Y}_i - R\bar{X}_i)^2 + \frac{1}{nN} \sum u_i^2 \frac{M_i - m_i}{m_i M_i} D_i^2 \\ &= \frac{N-n}{nN} (S_{1by}^2 - 2RS_{1bxy} + R^2 S_{1bx}^2) + \frac{1}{nN} \sum_{i=1}^N u_i^2 \frac{M_i - m_i}{m_i M_i} D_i^2. \end{aligned}$$

Sukhatme and Sukhatme (1970) have mentioned that if the values of M_i do not vary too much. \bar{y}_{1R} is more precise compared to other estimators. We have mentioned that simple random sampling scheme without replacement is followed in two-stage sampling. As a result, the value of $V(\bar{y}_1)$ depends on the totals of first stage units. Hence, if M_i varies too much, the value of S_{1b}^2 will be more. Again, M_i and S_i^2 are positively correlated and hence, the second term in $V(\bar{y}_1)$ will also be larger. Thus, $V(\bar{y}_1)$ is expected to be larger. As a result, the use of estimator \bar{y}_1 is not suitable though it is unbiased.

Theorem : In two-stage sampling, in case of unequal first stage units the unbiased estimator of variance of \bar{y}_1 is

$$v(\bar{y}_1) = \frac{N-n}{nN} s_{1b}^2 + \frac{1}{nN} \sum_{i=1}^n u_i^2 \frac{M_i - m_i}{m_i M_i} s_i^2,$$

where
$$s_{1b}^2 = \frac{1}{n-1} \sum_{i=1}^n (u_i \bar{y}_i - \bar{y}_1)^2,$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2.$$

Proof : It is easy to show that

$$E(s_{1b}^2) = S_{1b}^2 + \frac{1}{N-1} \sum_{i=1}^N \frac{M_i - m_i}{m_i M_i} u_i^2 S_i^2.$$

[In a similar way as it is done in case of two-stage sampling with equal first stage unit].

$$E(s_i^2) = E \left[\frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2 \right] = S_i^2.$$

[∵ second stage units are selected under SRSWOR scheme].

Then
$$E \left\{ \frac{1}{n} \sum_{i=1}^n u_i^2 \left[\frac{M_i - m_i}{m_i M_i} \right] s_i^2 \right\} = \frac{1}{N} \sum_{i=1}^N \frac{M_i - m_i}{m_i M_i} S_i^2 u_i^2.$$

Hence, the unbiased estimator of S_{1b}^2 is

$$\hat{S}_{1b}^2 = s_{1b}^2 - \frac{1}{N} \sum u_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) s_i^2.$$

Now, we have

$$v(\bar{y}_1) = \frac{N-n}{nN} s_{1b}^2 + \frac{1}{nN} \sum_{i=1}^n u_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) s_i^2.$$

If N is large enough $v(\bar{y}_1) = \frac{s_{1b}^2}{n}$.

Corollary : In two-stage sampling in case of unequal first stage units the unbiased estimator of variance of \bar{y}_2 is

$$v(\bar{y}_2) = \frac{N-n}{nN} s_{2b}^2 + \frac{1}{nN} \sum_{j=1}^1 \frac{M_j - m_j}{m_j M_j} s_j^2,$$

where $s_{2b}^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y}_2)^2$.

Corollary : In two-stage sampling in case of unequal first stage units the estimator of variance of ratio estimator \bar{y}_{1R} is

$$v(\bar{y}_{1R}) = \frac{N-n}{nN} \left[s_{1by}^2 - 2\hat{R}s_{1bxy} + \hat{R}^2 s_{1bx}^2 \right] \\ + \frac{1}{nN} \sum_{i=1}^n u_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) [s_{i_y}^2 - 2\hat{R}s_{i_{xy}} + \hat{R}^2 s_{i_x}^2],$$

where $s_{1by}^2 = \frac{1}{n-1} \sum_{i=1}^n (u_i \bar{y}_i - \bar{y}_1)^2$,

$$s_{1bx}^2 = \frac{1}{n-1} \sum_{i=1}^n (u_i \bar{x}_i - \bar{x}_1)^2,$$

$$s_{1bxy} = \frac{1}{n-1} \sum_{i=1}^n (u_i \bar{y}_i - \bar{y}_1)(u_i \bar{x}_i - \bar{x}_1),$$

$$s_{i_y}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2,$$

$$s_{i_x}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i)^2.$$

$$s_{i_{xy}} = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i), \quad \hat{R} = \bar{y}_1 / \bar{x}_1.$$

Example 17.2 : To estimate the total unemployed graduates in a police station area a survey is conducted. The number of administrative blocks in the police station area is 15. In each block there are different numbers of villages. Total villages in the study area are 120. Six administrative blocks are selected at random and from the administrative blocks some villages are selected at random. The number of unemployed graduates in the villages are shown below.

Estimate the total unemployed graduates in the police station area. Also estimate standard error of your estimate.

Number of unplayed graduates (y_{ij}) in the selected villages

SL. No.	No. of villages in the administrative block M_i	No. of unemployed graduates in the selected villages y_{ij}	No. of selected villages m_i	\bar{y}_i $= \frac{1}{m_i} \sum y_{ij}$	$\sum y_{ij}^2$
1	15	12, 8, 6, 7, 14	5	9.40	489
2	13	10, 9, 6, 16	4	10.25	473
3	8	18, 11	2	14.50	445
4	10	12, 16, 8	3	12.00	464
5	12	14, 20, 8, 12	4	13.50	804
6	9	15, 13, 8	3	12.00	458
Total	67		21	71.65	3133

Solution : We have $N = 15$, $M_0 = 120$, $n = 6$, $\sum_{i=1}^n m_i = 21$, $\bar{M} = \frac{1}{N} \sum_{i=1}^N M_i = \frac{120}{15} = 8$.

The average unemployed graduates per village is

$$\bar{y}_1 = \frac{1}{n\bar{M}} \sum_{i=1}^n M_i \bar{y}_i = \frac{780.25}{6 \times 8} = 16.25.$$

Estimate of total unemployed graduates in the police station area is

$$\hat{Y} = M_0 \bar{y}_1 = 120 \times 16.25 = 1950.$$

$u_i = \frac{M_i}{\bar{M}}$	$\frac{1}{m_i - 1} [\sum y_{ij}^2 - m_i \bar{y}_i^2] = s_i^2$	$(u_i \bar{y}_i - \bar{y}_1)^2$	$\frac{u_i^2 (M_i - m_i)}{m_i M_i} s_i^2$
0.625	11.80	107.6406	0.61458
0.500	17.583	123.7656	0.76080
0.250	24.500	159.3906	0.57422
0.375	16.000	138.0625	0.525
0.500	25.000	90.2500	1.04167
0.375	13.000	138.0625	0.40625
Total		757.1718	3.92252

$$s_{1b}^2 = \frac{1}{n-1} \sum (u_i \bar{y}_i - \bar{y}_1)^2 = \frac{757.1718}{5} = 151.43436.$$

$$\begin{aligned} v(\bar{y}_1) &= \frac{N-n}{nN} s_{1b}^2 + \frac{1}{nN} \sum u_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) s_i^2 \\ &= \frac{15-6}{6 \times 15} \times 151.4346 + \frac{1}{6 \times 15} \times 3.92252 = 15.187019. \end{aligned}$$

$$\therefore v(\hat{Y}) = M_0^2 v(\bar{y}_1) = (120)^2 15.187019 = 218693.0816.$$

$$\text{s.e.}(\hat{Y}) = \sqrt{v(\hat{Y})} = 467.65.$$

17.5 Estimation of Proportion in Two-Stage Sampling with Equal First Stage Units

Let us consider that the variable under study in two-stage sampling is qualitative in nature. Then

$y_{ij} = 0$, if j -th second stage unit within i -th first stage unit does not possess the character.

$= 1$, if j -th second stage unit within i -th first stage unit possesses the character.

$i = 1, 2, \dots, N$; $j = 1, 2, \dots, M$. Then, the proportion of units in the selected i -th first stage unit possessing the character is

$$p_i = \frac{1}{m} \sum_{j=1}^M y_{ij} = \frac{a_i}{m}, \quad i = 1, 2, \dots, n.$$

Let P be the proportion of units in the population possessing the character, where

$$P = \frac{1}{NM} \sum_i^N \sum_{j=1}^M y_{ij} = \frac{\sum_{i=1}^N A_i}{NM} = \frac{A}{NM} = \frac{1}{N} \sum P_i,$$

where $P_i = \frac{1}{M} \sum_{j=1}^M y_{ij} = \frac{A_i}{M}$.

The estimator of P is

$$\hat{P} = \frac{1}{mn} \sum_{i=1}^n a_i = \frac{1}{n} \sum_{i=1}^n p_i.$$

The estimator of variance of this \hat{P} is

$$\begin{aligned} v(\hat{P}) &= \frac{1-f}{n(n-1)} \sum_{i=1}^n (p_i - \hat{P})^2 + \frac{f(1-f_1)}{n^2(m-1)} \sum_{i=1}^n p_i q_i \\ &= \frac{1-f}{n} s_b^2 + \frac{f(1-f_1)}{nm} s_w^2, \end{aligned}$$

where $s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (p_i - \hat{P})^2$, $s_w^2 = \frac{m}{n(m-1)} \sum_{i=1}^n p_i q_i$, $q_i = 1 - p_i$.

Example 17.3 : To estimate the proportion of patients admitted in a hospital in a month who suffer from Hepatitis-B virus, a survey is conducted. In the month total patients in the hospital are 500. These patients are treated in 10 wards. In each ward there are 50 patients. Each ward is considered as a cluster. Five clusters are randomly selected. From each randomly selected cluster 50% patients are randomly selected and investigated. The number of patients suffering from Hepatitis-B in each selected ward are shown below :

Sl. No.	Patients suffering from Hepatitis-B a_i	$p_i = \frac{a_i}{m}$	$q_i = 1 - p_i$	$p_i q_i$
1	12	0.48	0.52	0.2496
2	8	0.32	0.68	0.2176
3	10	0.40	0.60	0.2400
4	16	0.64	0.36	0.2304
5	10	0.40	0.60	0.2400

Estimate the total patients suffering from Hepatitis-B and also estimate the standard error of your estimate.

Solution : $N = 10$, $n = 5$, $M = 50$, $m = 25$

The proportion of patients suffering from hepatitis-B virus is

$$\hat{P} = \frac{1}{n} \sum_{i=1}^n p_i = \frac{2.24}{5} = 0.448.$$

The estimate of total patients is

$$\hat{A} = NM\hat{P} = 10 \times 50 \times 0.448 = 224.$$

The estimator of variance of \hat{P} is

$$v(\hat{P}) = \frac{1-f}{n} s_b^2 + \frac{f(1-f_1)}{nm} s_w^2, \text{ where } f = \frac{n}{N} = 0.5, f_1 = \frac{m}{M} = 0.5,$$

$$\begin{aligned} s_b^2 &= \frac{1}{n-1} \sum_{i=1}^n (p_i - \hat{P})^2 = \frac{1}{n-1} \left[\sum p_i^2 - n\hat{P}^2 \right] \\ &= \frac{1}{5-1} [1.0624 - 5 \times (0.448)^2] = 0.01472. \end{aligned}$$

$$s_w^2 = \frac{m}{n(m-1)} \sum_{i=1}^n p_i q_i = \frac{25}{5(25-1)} \times 1.1776 = 0.24533.$$

$$v(\hat{P}) = \frac{1-0.5}{5} \times 0.01472 + \frac{0.5 \times 0.5}{5 \times 25} \times 0.24533 = 0.0019625.$$

Therefore, $v(\hat{A}) = N^2 M^2 v(\hat{P}) = (10)^2 (50)^2 \times 0.0019625 = 490.65$.

The estimate of standard error of \hat{A} is

$$\text{s.e.}(\hat{A}) = \sqrt{v(\hat{A})} = 22.15.$$

17.6 Allocation of Sample Sizes in Two-Stages

Let us discuss the problem in two-stage sampling in case of equal first-stage units. The variance of sample mean of two-stage sampling when first-stage units are equal is

$$V(\bar{y}) = \frac{N-n}{nN} S_b^2 + \frac{M-m}{mM} \frac{S_w^2}{n}.$$

It is observed that the precision of the estimator \bar{y} depends on the value of n , m , S_b^2 and S_w^2 . The expenditure of the survey depends on n and m . Hence, the values of n and m are to be determined in such a way that for a fixed cost $V(\bar{y})$ is minimum.

Let the fixed cost function $C = anm$, where $a =$ proportionality constant. Let the fixed cost for the survey be C_0 such that

$$C_0 = anm \Rightarrow m = \frac{C_0}{na}.$$

Putting the value of m in the formula for $V(\bar{y})$, we get

$$V(\bar{y}) = \frac{N-n}{Nn} S_b^2 + \frac{M - C_0/na}{M \frac{C_0}{an}} S_w^2 = \left(S_b^2 - \frac{S_w^2}{M} \right) \frac{1}{n} + \frac{aS_w^2}{C_0} - \frac{S_b^2}{N}.$$

This variance is a monotone decreasing function of n . Therefore, $V(\bar{y})$ will be minimum if n takes highest value when $S_b^2 - S_w^2/M > 0$. Let us consider that the estimate of m is $\hat{m} = 1$. Then the estimated value of n is $\hat{n} = C_0/a$. Again, if $S_b^2 - S_w^2/M < 0$, $V(\bar{y})$ is a monotone increasing function of n and the variance will be minimum if n is minimum. The value of \hat{n} will be minimum when $\hat{m} = M$, because

$$\hat{n} = \frac{C_0}{aM}.$$

The values of \hat{m} and \hat{n} are decided for a fixed cost C_0 . These estimated values may also be found out for a definite amount of precision of estimator. Let the fixed value of variance of \bar{y} is

$$V_0 = \frac{N-n}{nN} S_b^2 + \frac{M-m}{mM} \frac{S_w^2}{n}.$$

From this variance formula, we have

$$n = \frac{S_b^2 + \frac{M-m}{mM} S_w^2}{V_0 + \frac{S_b^2}{N}}.$$

Putting the value of n in the cost function $C = anm$, we get

$$C = am \left[\frac{S_b^2 - \frac{S_w^2}{M}}{V_0 + \frac{S_b^2}{N}} \right] + \frac{aS_w^2}{V_0 + \frac{S_b^2}{N}}.$$

From the above value of C it may be concluded that if $S_b^2 - \frac{S_w^2}{M} > 0$, C will be minimum when m is minimum. Thus, C is minimum if $\hat{m} = 1$. Again, if $S_b^2 - \frac{S_w^2}{M} < 0$, C will be minimum when m becomes maximum. The maximum value of m is $\hat{m} = M$.

The value of m and n can be derived for a general cost function also. Let this general function be $C = C_1n + C_2nm$, where C_1 is the cost of including a first-stage unit in the sample and C_2 is the cost of inclusion of m second stage sampling units in the sample.

The variance of sample mean is

$$V(\bar{y}) = \frac{1}{n} \left(S_b^2 - \frac{S_w^2}{M} \right) + \frac{S_w^2}{nm} - \frac{S_b^2}{N}.$$

The last term in the right-hand side of $V(\bar{y})$ does not depend on n or m , the middle term is a function of variance of first stage units and the first term is the function of variance of first stage units and the variance of means of first stage units. The first term becomes minimum if n is maximum and the middle term becomes minimum if both n and m become maximum. Therefore, it is noted for maximum values of n and m $V(\bar{y})$ will be minimum. But, when n and m become maximum, the cost of survey will be increased. Therefore, the values of n and m are to be estimated in such a way that the estimator becomes precise for a fixed amount of cost.

Putting the value of n from cost function in the formula of $V(\bar{y})$, we get

$$\left\{ V(\bar{y}) + \frac{S_b^2}{N} \right\} C = C_1 \left(S_b^2 - \frac{S_w^2}{N} \right) + C_2 S_w^2 + C_2 m \left(S_b^2 - \frac{S_w^2}{M} \right) + \frac{C_1 S_w^2}{m}.$$

The optimum values of n and m are derived by minimizing the above variance function.

Let $S_b^2 - \frac{S_w^2}{M} > 0$. Then the above variance function can be written as

$$C \left\{ V(\bar{y}) + \frac{S_b^2}{N} \right\} = \left[\sqrt{C_1 \left(S_b^2 - \frac{S_w^2}{M} \right) + \sqrt{C_2 S_w^2}} \right]^2 + \left[\sqrt{C_2 m \left(S_b^2 - \frac{S_w^2}{M} \right) - \sqrt{\frac{C_1 S_w^2}{m}}} \right]^2$$

The above variance function becomes minimum if the last term in the right-hand side becomes zero. That is, if

$$\sqrt{C_2 m \left(S_b^2 - \frac{S_w^2}{M} \right)} = \sqrt{\frac{C_1 S_w^2}{m}}$$

Therefore, we have

$$\hat{m} = \sqrt{\frac{C_1}{C_2} \frac{S_w^2}{\left(S_b^2 - \frac{S_w^2}{M} \right)}} \approx \sqrt{\frac{C_1}{C_2} \left(\frac{1}{\rho} - 1 \right)},$$

where ρ is the intra-class correlation coefficient of observations of first-stage units.

Again, if $S_b^2 - S_w^2/M < 0$, the right-hand side term of the above variance function becomes minimum, when m becomes maximum. Now, if the total cost is fixed, i.e., $C = C_0$ and $C_0 > C_1 + C_2 M$, then $\hat{m} = M$ and the value of \hat{n} will be the maximum value of

$$\hat{n} \leq \frac{C_0}{C_1 + C_2 M}$$

Again, if $C_0 < C_1 + C_2 M$, then \hat{m} will be the maximum, if $\hat{n} \leq \frac{C_0 - C_1}{C_2}$ and $\hat{n} = 1$.

It is observed that the value of \hat{m} depends on C_1, C_2 and ρ . Generally, $S_b^2 - \frac{S_w^2}{M} > 0$. Hence, m will be minimum, if

- (i) the cost of travelling of first-stage units is less,
- (ii) the cost of selection of second-stage units from the selected first-stage units is more,
- (iii) ρ is maximum.

Corollary : In two-stage sampling, if $C_0 > C_1 + C_2 M$, the formula to decide optimum m from sample observations is

$$\hat{m} = \left[\frac{C_1}{C_2} \frac{s_w^2}{s_b^2 - \frac{s_w^2}{M}} \right]^{\frac{1}{2}}$$

Corollary : In two-stage sampling the minimum variance of \bar{y} for optimum m and n is

$$V_{\min}(\bar{y}) = \frac{S_b^2}{N} + \frac{1}{C_0} \left[C_1 \sqrt{\left(S_b^2 - \frac{S_w^2}{M} \right) + \sqrt{C_1 S_w^2}} \right]^2$$

The estimator of this variance is

$$v_{\min}(\bar{y}) = \frac{s_b^2}{N} + \frac{1}{C_0} \left[C_1 \sqrt{\left(s_b^2 - \frac{s_w^2}{M} \right) + \sqrt{C_2 s_w^2}} \right]^2$$

Allocation of sample size in two stages when first stage units are unequal

Let there be M_0 elements in a population. The population units are divided into N clusters, the size of i -th cluster is $M_i (i = 1, 2, \dots, N)$. According to the definition of two-stage sampling, we need to select n clusters (primary units) from N clusters by a random process and then from i -th primary units m_i second stage units are to be selected by a random process. The problem is to decide the values of n and $m_i (i = 1, 2, \dots, n)$. The value of m_i may be proportional to M_i , it may be independent of M_i or by any arbitrary process all m_i 's may be same [$m_1 = m_2 = \dots = m_n$]. However, the value of m_i is to be chosen in such way that the sample observations must provide more precise estimator of population parameter. The precision is to increase in such a way that the cost of the survey becomes minimum or the estimator is to be found out for a pre-fixed precision so that the cost is reduced.

Let the total cost of the survey be

$$C = C_1 n + C_2 \frac{n}{N} \sum_{i=1}^N m_i.$$

Here $C_1 n$ is the cost of survey of first-stage units and $C_2 \frac{n}{N} \sum m_i$ is the average cost of survey of second-stage units. Then

$$\left\{ V(\bar{y}_1) + \frac{S_{1b}^2}{N} \right\} C = C_1 \Delta + \frac{C_2}{N^2} \sum_{i=1}^N u_i^2 S_i^2 + \frac{1}{N} \sum_{i=1}^N \left(C_2 \Delta m_i + C_1 u_i^2 \frac{S_i^2}{m_i} \right) + \frac{C_2}{N^2} \sum_{i>i'=1}^N \left(\frac{m_i}{m_{i'}} u_{i'} S_{i'}^2 + \frac{m_{i'}}{m_i} u_i S_i^2 \right),$$

where $\Delta = S_{1b}^2 - \frac{1}{NM} \sum_{i=1}^N u_i S_i^2$.

If Δ is positive, then right-hand side of the variance function is written as

$$\frac{1}{N} \sum_{i=1}^N \left(\sqrt{C_2 \Delta m_i} - \sqrt{\frac{C_1}{m_i}} u_i S_i \right)^2 + \frac{C_2}{N^2} \sum_{i>i'=1}^N \left(\sqrt{\frac{m_i}{m_{i'}}} u_{i'} S_{i'} - \sqrt{\frac{m_{i'}}{m_i}} u_i S_i \right)^2 + \text{terms independent of } m_i$$

This quantity becomes minimum if both the squares are zero. In such a case, the estimated value of m_i is

$$\hat{m}_i = \sqrt{\frac{C_1}{C_2 \Delta}} u_i S_i \quad (i = 1, 2, \dots, N).$$

It is observed that, the value of \hat{m}_i depends on the cost of survey, the selected first-stage units and on the variance of the observations of first-stage units. In practice, the value of S_i^2 is not known. Hence, \hat{m}_i should be independent of S_i^2 . To get \hat{m}_i independent of S_i^2 , it is not essential for \hat{m}_i to be optimum.

If S_i^2 's are same for all $i (i = 1, 2, \dots, N)$, then m_i can be taken proportional to M_i . However, S_i^2 's are not equal, in practice, rather it increases with the increase in the value of M_i , though the rate of increase of S_i^2 is less than the rate of increase of M_i . To avoid the problem, strata can

be formed with clusters of equal sizes and m_i can be taken proportional to M_i . Let $m_i = KM_i$, where K is a constant and

$$K = \sqrt{\frac{C_1}{C_2} \frac{1}{NM^2} \sum_{i=1}^N u_i S_i^2}$$

The value of k leads to get minimum variance.

Example 17.4 In an agricultural research station an experiment is conducted to study the productivity of improved variety of mango. 400 mango plants are planted in different areas. In each area 20 plants are nursed. To estimate the total production of mango, 10 areas are randomly selected and from each selected area 5 plants are randomly selected. The total mango production per selected plant are shown below :

Sl. No.	Production of mango (y_{ij} in number) in selected plants					Mean \bar{y}_i	$\sum_{j=1}^{m_i} y_{ij}^2$
	1	2	3	4	5		
1	15	28	22	40	35	28.00	4318
2	42	12	42	48	55	39.80	9001
3	22	32	35	52	60	40.20	9037
4	52	38	30	24	28	34.40	6408
5	50	15	45	48	53	42.20	9863
6	18	24	37	46	62	37.40	8229
7	19	38	46	42	36	36.20	6981
8	20	24	54	50	52	40.00	9096
9	56	17	15	38	44	34.00	7030
10	29	44	40	48	31	38.40	7642
Total						370.60	77605

- (i) Estimate the total number of mango produced in the farm.
- (ii) Estimate the standard error of your estimate.
- (iii) If the cost function for the survey is $1000 = 20n + 10nm$, then find the optimum values of n and m .
- (iv) Find the minimum estimated variance of the estimator of total production.

Solution : (i) Given $N = 20$, $M = 20$, $n = 10$, $m = 5$.

The estimated mean production of mango is

$$\bar{y} = \frac{1}{nm} \sum \sum y_{ij} = \frac{1}{n} \sum \bar{y}_i = \frac{370.6}{10} = 37.06$$

The estimated total mango production is

$$\hat{Y} = MN\bar{y} = 20 \times 20 \times 37.06 = 14824$$

(ii) $s_b^2 = \frac{1}{n-1} \left[\sum \bar{y}_i^2 - n\bar{y}^2 \right] = \frac{1}{5-1} [13888.04 - 10 \times (37.06)^2] = 38.401$

$$s_w^2 = \frac{1}{n(m-1)} \sum_i^n \left[\sum_{j=1}^m y_{ij}^2 - m\bar{y}_i^2 \right] = \frac{1}{n(m-1)} \left[\sum \sum y_{ij}^2 - m \sum_{i=1}^n \bar{y}_i^2 \right]$$

$$= \frac{1}{10(5-1)} [77605 - 5 \times (13888.04)] = 204.12.$$

Therefore, the estimate of variance of \hat{Y} is

$$v(\hat{Y}) = N^2 M^2 \left[\frac{1-f}{n} s_b^2 + \frac{f(1-f_1)}{nm} s_w^2 \right]$$

$$= (20)^2 (20)^2 \left[\frac{1-0.5}{10} 38.401 + \frac{0.5(1-0.25)}{10 \times 5} \times 204.12 \right]$$

$$= 552152.00.$$

$$\therefore \text{s.e.}(\hat{Y}) = \sqrt{v(\hat{Y})} = 743.07.$$

(iii) The cost function $C = C_1 n + C_2 nm$, where $C = 1000 = C_0$, $C_1 = 20$, $C_2 = 10$. It is observed that $C_0 > C_1 + C_2 N$.

$$\hat{m} = \left[\frac{C_1}{C_2} \frac{s_w^2}{s_b^2 - \frac{s_w^2}{m}} \right]^{\frac{1}{2}}$$

$$\text{But } s_b^2 - \frac{s_w^2}{m} = 38.401 - \frac{204.12}{5} = -2.423 < 0.$$

$$\text{Therefore, } \hat{m} = M = 20 \text{ and } \hat{n} = \frac{C_0}{C_1 + C_2 M} = \frac{1000}{20 + 10 \times 20} = 4.5 \approx 5.$$

Now, the $V_{\min}(\hat{Y})$ for \hat{m} and \hat{n} is

$$u_{\min}(\hat{Y}) = M^2 N^2 \left\{ \frac{s_b^2}{N} + \frac{1}{C_0} \left[\sqrt{C_1 \left(s_b^2 - \frac{s_w^2}{M} \right)} + \sqrt{C_2 s_w^2} \right]^2 \right\}$$

$$= (20)^2 (20)^2 \left\{ \frac{38.401}{20} + \frac{1}{1000} \left[\sqrt{20 \left(38.401 - \frac{204.12}{20} \right)} + \sqrt{10 \times 204.12} \right]^2 \right\}$$

$$= 160000 \left\{ 1.92005 + \frac{1}{1000} [75.0933 + 45.1796]^2 \right\}$$

$$= 2621699.276.$$

17.7 Two-Stage Sampling with Varying Probabilities

In two-stage sampling, sampling units are selected at each stage by simple random sampling. This sampling scheme is good if the sizes of first stage units are smaller and if the variance of observation within first-stage unit is smaller. The efficiency of the sampling scheme is decreased if the sizes of first-stage units are large and if the observations within a first-stage unit vary too much. The efficiency of the sampling plan is not increased even if smaller number of first-stage units are selected and even if equal number of second-stage units are selected from the selected first-stage units. Therefore, an alternative sampling technique is needed to avoid the problem mentioned above.

The problem is obviated if sampling units are selected with varying probabilities. This is possible if first-stage units are grouped into strata according to the homogeneity of sizes of first-stage units and from each stratum if first-stage units are selected with probability proportional to sizes of first-stage units. The second-stage units are selected from each selected first-stage unit with equal probability.

Hausen and Hurwitz (1943, 1949) have discussed the technique in details. Singh (1954) has compared two estimators when first-stage units are selected by probability proportional to sizes of units without replacement and second-stage units are selected with equal probability with replacement and without replacement. Rao (1966) have proposed alternative estimators in case of above mentioned sampling procedure.

17.8 Method of Estimation in Two-Stage Sampling with Varying Probabilities

Let there be $M_0 = \sum_{i=1}^N M_i$ elements in a population which are divided into N -clusters. The number of second-stage units in i -th first-stage unit is M_i ($i = 1, 2, \dots, N$). Consider that a sample of n primary units is needed and these will be selected with replacement. Let the selected first primary unit is i -th unit. Then from this selected i -th unit m_i second-stage units are to be selected without replacement. Since first-stage units are selected with replacement, the i -th unit has chance to be included in the sample repeatedly. In that case, to select the second-stage units the selected m_i units are replaced again and again. Consider that i -th unit is included t_i times in the sample. Then m_i second-stage units are included in the sample t_i times and every time the sub-sampling will be independent.

Let p_i ($i = 1, 2, \dots, N$) be the probability of selection of i -th unit in the sample such that $\sum_{i=1}^N p_i = 1$. Let us define a variable Z_{ij} which is observed from j -th second-stage unit of selected i -th unit, where

$$Z_{ij} = \frac{M_i y_{ij}}{M_0 p_i} \quad \text{and} \quad \bar{Z}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Z_{ij}.$$

Then $\bar{Z}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Z_{ij}$. Using this \bar{Z}_i , we can formulate an estimator of population mean \bar{Y} , where the estimator is

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n \bar{Z}_i.$$

Theorem : In two-stage sampling, if first-stage units are selected by probability proportional to size (PPS) of first stage unit, then the unbiased estimator of population mean \bar{Y} is \bar{Z} and the variance of this estimator is

$$V(\bar{Z}) = \frac{1}{n} \sum_{i=1}^N \frac{M_i}{M_0} (\bar{Y}_i - \bar{Y})^2 + \frac{1}{n} \sum_{i=1}^N \frac{M_i}{M_0} \left(\frac{M_i - m_i}{m_i M_i} \right) S_{iz}^2,$$

where $S_{iz}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (Z_{ij} - \bar{Z})^2$.

Proof: The estimator \bar{Z} is defined by $\bar{Z} = \frac{1}{n} \sum_{i=1}^n \bar{Z}_i$.

$$\begin{aligned} E(\bar{Z}) &= E\left[\frac{1}{n} \sum_{i=1}^n \bar{Z}_i\right] = E\left[\frac{1}{n} \sum_{i=1}^n E(\bar{Z}_i)/i\right] \\ &= E\frac{1}{n} \sum_{i=1}^n \bar{Z}_i = \frac{1}{n} \sum_{i=1}^n E(\bar{Z}_i) = \bar{Z}_{..}, \end{aligned}$$

$$\text{where } \bar{Z}_{..} = \sum_{i=1}^N p_i \bar{Z}_i = \sum_{i=1}^N p_i \frac{1}{M_i} \sum_{j=1}^{M_i} Z_{ij} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{p_i M_i y_{ij}}{M_i M_0 p_i} = \frac{1}{M_0} \sum_i \sum_{j=1}^{M_i} y_{ij} = \bar{Y}.$$

Hence, \bar{Z} is an unbiased estimator of \bar{Y} .

The variance of \bar{Z} is written as

$$\begin{aligned} V(\bar{Z}) &= V\left[\frac{1}{n} \sum_{i=1}^n \bar{Z}_i\right] \\ &= V\left[E\left(\frac{1}{n} \sum_{i=1}^n \bar{Z}_i/n\right)\right] + E\left[V\left(\frac{1}{n} \sum_{i=1}^n \bar{Z}_i/n\right)\right] \\ &= V\left[\frac{1}{n} \sum_{i=1}^n \bar{Z}_i\right] + E\left[\frac{1}{n^2} \sum_{i=1}^n \frac{M_i - m_i}{m_i M_i} S_{iz}^2\right]. \end{aligned}$$

Again, we have

$$V\left[\frac{1}{n} \sum_{i=1}^n \bar{Z}_i\right] = \frac{1}{n} \sum_{i=1}^N p_i (\bar{Z}_i - \bar{Z}_{..})^2, \quad \bar{Z}_{..} = \frac{1}{M_0} \sum_{i=1}^N \sum_{j=1}^{M_i} Z_{ij}.$$

$$\text{Then } V(\bar{Z}) = \frac{1}{n} \sum_{i=1}^N p_i (\bar{Z}_i - \bar{Z}_{..})^2 + \frac{1}{n} \sum_{i=1}^N p_i \frac{M_i - m_i}{m_i M_i} S_{iz}^2.$$

$$\text{But } p_i = \frac{M_i}{M_0} \quad [i = 1, 2, \dots, N] \text{ gives } Z_{ij} = y_{ij}$$

$$\text{and } V\left(\frac{1}{n} \sum_{i=1}^n \bar{Z}_i\right) = \sum_{i=1}^N \frac{M_i}{M_0} (\bar{Y}_i - \bar{Y})^2.$$

$$\text{Also } S_{iz}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2 = S_i^2.$$

$$\text{Hence, } V(\bar{Z}) = \frac{1}{n} \sum_{i=1}^N \frac{M_i}{M_0} (\bar{Y}_i - \bar{Y})^2 + \frac{1}{n} \sum_{i=1}^N \frac{M_i}{M_0} \left(\frac{M_i - m_i}{m_i M_i}\right) S_i^2.$$

Theorem : In two-stage sampling if first-stage units are selected by PPS sampling scheme, then the unbiased estimator of variance of \bar{Z} is

$$v(\bar{Z}) = \frac{s_{\bar{z}}^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{Z}_i - \bar{Z})^2.$$

$$\text{Proof: Given } s_{bz}^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{Z}_i - \bar{Z})^2 = \frac{1}{n-1} \left[\sum \bar{Z}_i^2 - n\bar{Z}^2 \right]$$

$$\begin{aligned} E[s_{bz}^2] &= \frac{1}{n-1} \left[E \left(\sum \bar{Z}_i^2 \right) - nE(\bar{Z}^2) \right] \\ &= \frac{1}{n-1} \left[E \sum_{i=1}^n E(\bar{Z}_i^2 / i) - nE(\bar{Z}^2) \right] \\ &= \frac{1}{n-1} \left[E \sum_{i=1}^n \left\{ \bar{Z}_i^2 + \left(\frac{M_i - m_i}{m_i M_i} \right) S_{iz}^2 \right\} - n\{\bar{Z}^2 + V(\bar{Z})\} \right] \\ &= \frac{1}{n-1} \left[n \sum_{i=1}^N p_i \left\{ \bar{Z}_i^2 + \frac{M_i - m_i}{m_i M_i} S_{iz}^2 \right\} - n\{\bar{Z}^2 + V(\bar{Z})\} \right]. \end{aligned}$$

Now, putting the value of $V(\bar{Z})$ and on simplification, we get

$$E(s_{bz}^2) = \sum_{i=1}^N p_i (\bar{Z}_i - \bar{Z}_{..})^2 + \sum_{i=1}^N p_i \left(\frac{M_i - m_i}{m_i M_i} \right) S_{iz}^2.$$

$$\therefore V(\bar{Z}) = \frac{s_{bz}^2}{n}.$$

Example 17.5 : In a poultry farm there are 120 sheds to rear the chicks. The sheds are divided into 10 blocks. In each block there are different numbers of chicks. Below is given the number of chicks in each shed and in different blocks.

SL. No. of blocks	No. of sheds in a block	Number of chicks (y_{ij}) per shed
1	8	20, 22, 18, 17, 15, 25, 18, 16
2	12	10, 15, 12, 14, 15, 11, 13, 12, 10, 15, 12, 10
3	15	9, 6, 15, 18, 12, 12, 12, 10, 14, 12, 12, 11, 13, 10, 10
4	9	15, 12, 8, 12, 16, 18, 12, 10, 8,
5	14	10, 10, 10, 10, 12, 11, 9, 9, 10, 12, 15, 16, 12, 10
6	14	12, 11, 11, 11, 8, 10, 10, 12, 12, 11, 13, 14, 10, 11
7	8	15, 18, 14, 20, 20, 18, 20, 17
8	13	15, 16, 17, 13, 16, 16, 17, 18, 20, 18, 19, 10, 12
9	12	16, 20, 19, 8, 14, 16, 12, 14, 18, 17, 15, 10
10	15	10, 12, 16, 9, 15, 18, 19, 20, 12, 16, 17, 18, 13, 14, 15

- (i) Select a sample of 5 blocks proportional to size of sheds in blocks with replacement and select 25% sheds from selected blocks.
- (ii) Estimate the total number of chicks in the farm.
- (iii) Estimate the standard error of your estimator.

Solution : (i) The sample can be selected using Lahiri's method of sample selection. According to the method, first block is to be selected selecting a pair of numbers. The first

value of the pair is a selected number from 1 to N and the second value of the pair is a random number from 1 to M , where M is the largest size of the clusters. If this pair corresponds to a cluster along with its size, the cluster is to be included in the sample. For example, the first selected number from 1 to 10 ($\because N = 10$), using random number in Appendix, is 01 and the second value of the pair is from 1 to 15 is 01. Hence, the pair is (01, 01). This pair corresponds to first block. Hence, first block is included in the sample. Other clusters are selected in a similar way. Accordingly, the other pairs are (06, 06), (08, 08), (03, 03) and (05, 05). The sample blocks and sheds along with number of chicks in sheds are shown below :

SL. No. of blocks selected	No. of sheds in blocks M_i	No. of sheds selected in the sample m_i	Probability of selection $p_i = \frac{M_i}{M_0}$	No. of chicks in sample	\bar{y}_i
1	8	2	0.06667	20, 25	22.5
6	14	4	0.11667	12, 11, 8, 12,	10.75
8	13	3	0.10833	13, 16, 16,	15.00
3	15	4	0.125	12, 12, 18, 12,	13.50
5	14	4	0.11667	10, 10, 15, 10,	11.25

We have $N = 10$, $M_0 = 120$, $n = 5$.

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n \bar{Z}_i = \frac{1}{n} \sum_{i=1}^n \bar{y}_i = \frac{73.00}{5} = 14.60.$$

Estimate of total chicks is

$$\hat{Y} = M_0 \bar{Z} = 120 \times 14.60 = 1752.$$

$$\begin{aligned} \text{(ii) } V(\bar{Z}) &= \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{Z}_i - \bar{Z})^2 = \frac{1}{n(n-1)} \left[\sum \bar{y}_i^2 - n\bar{Z}^2 \right] \\ &= \frac{1}{5(5-1)} [1155.625 - 5(14.60)^2] = 4.49125. \end{aligned}$$

Hence, the estimator of variance of \hat{Y} is

$$v(\hat{Y}) = M_0^2 v(\bar{z}) = (120)^2 4.49125 = 64674.$$

$$\text{s.e. } (\hat{Y}) = \sqrt{v(\hat{Y})} = 254.31.$$

Example 17.6 : Estimate the total number of chicks in the farm using data of example 17.5. Select 5 blocks by probability proportional to number of chicks per block. Also estimate the standard error of your estimator.

Solution : $N = 10$, $\hat{n} = 5$,

SL. No. of blocks	1	2	3	4	5	6	7	8	9	10	Total
No. of chicks Y_i	151	149	176	111	156	156	142	207	179	224	1651
No. of sheds M_i	8	12	15	9	14	14	8	13	12	15	120

Using Lahiri's method we can select the sample. The selected pairs are (01, 066), (03, 033), (03, 165), (05, 057), (01, 111). The sample observations and related results are shown below :

SL. No. of selected cluster	No. of sheds M_i	No. of sheds in sample m_i	Probability of selection $p_i = \frac{Y_i}{Y}$	Sample observations y_{ij}	$Z_{ij} = \frac{M_i y_{ij}}{M_0 p_i}$	\bar{z}_i
1	8	2	0.09146	18, 20	13.12, 14.58	13.85
8	13	3	0.12538	20, 10, 13	17.28, 8.64 11.23	12.38
3	15	4	0.10660	12, 12, 18, 11	14.07, 14.07 21.11, 12.90	15.54
5	14	4	0.09449	12, 10, 16, 15	14.82, 12.35 19.76, 18.52	16.36
1	8	2	0.09146	20, 18	14.58, 13.12	13.85
Total						71.98

The average chicks per shed is

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n \bar{z}_i = \frac{71.98}{5} = 14.396.$$

The estimate of total chicks is

$$\hat{Y} = M_0 \bar{Z} = 120 \times 14.396 = 1728.$$

The estimate of variance of \bar{Z} is

$$\begin{aligned} v(\bar{Z}) &= \frac{1}{n(n-1)} \sum (\bar{z}_i - \bar{Z})^2 = \frac{1}{n(n-1)} \left[\sum \bar{z}_i^2 - n\bar{Z}^2 \right] \\ &= \frac{1}{5(5-1)} [1046.0506 - 5(14.396)^2] = 0.491326 \end{aligned}$$

Then $V(\hat{Y}) = M_0^2 v(\bar{Z}) = (120)^2 0.491326 = 7075.0944$

$$\text{s.e.}(\hat{Y}) = \sqrt{v(\hat{Y})} = 84.11.$$

17.9 Two-Stage Sampling With Varying Probabilities at Each Stage

We have discussed the method of selection of first-stage units by probability proportional to the size of first-stage unit. The second-stage units can also be selected by probability proportional to size or any other measure of size of second-stage units. Let us now discuss the selection of units of both stages by PPS sampling scheme.

Let x_{ij} = the measure of size of j -th second-stage unit in i -th first-stage unit; $i = 1, 2, \dots, N$; $j = 1, 2, \dots, M_i$.

y_{ij} = the value of the variable under study of j -th second-stage unit in i -th first-stage unit.

$$X_i = \sum_{j=1}^{M_i} x_{ij} = \text{the total size of } i\text{-th first-stage unit.}$$

$Y_i = \sum_{j=1}^{M_i} y_{ij}$ = the total value of the variable of i -th first-stage unit.

$X = \sum_{i=1}^N X_i = \sum_{i=1}^N \sum_{j=1}^{M_i} x_{ij}$ = total of population units.

$Y = \sum_{i=1}^N Y_i = \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}$ = total value of the variable in the population.

$r_{ij} = \frac{y_{ij}}{x_{ij}}$ = ratio of variable to size of j -th second-stage unit in i -th first-stage unit.

$R_i = \frac{Y_i}{X_i}$ = ratio of total value of i -th first-stage unit to size of i -th first-stage unit.

$R = \frac{Y}{X}$ = ratio of total value of total size in the population.

Theorem : In two-stage sampling if sample is selected by PPS scheme at each stage, then the unbiased estimator of population ratio R is \bar{r} and the variance of this estimator is

$$V(\bar{r}) = \frac{1}{n} \sum_{i=1}^N \frac{X_i}{X} (R_i - R)^2 + \frac{1}{nm} \sum_{j=1}^{M_i} \frac{x_{ij}}{X_i} (r_{ij} - R_i)^2,$$

where $\bar{r} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m r_{ij} = \frac{1}{n} \sum_{i=1}^n \bar{r}_i$.

Proof : The probability of inclusion of j -th second-stage unit from i -th first-stage unit in the sample is

$$p_{ij} = \frac{x_{ij}}{X_i}, \quad i = 1, 2, \dots, N; \quad j = 1, 2, \dots, M_i.$$

Again, the probability of inclusion of i -th first-stage unit in the sample is

$$p_i = \frac{X_i}{X}, \quad i = 1, 2, \dots, N.$$

Given $r_{ij} = \frac{y_{ij}}{x_{ij}}$ and $\bar{r} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m r_{ij}$.

$$\begin{aligned} E(\bar{r}) &= \frac{1}{nm} E \sum_i^n \sum_j^m E(r_{ij}/i) \\ &= \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \sum_{j=1}^{M_i} r_{ij} \frac{x_{ij}}{X_i} = E \sum_{j=1}^{M_i} \frac{y_{ij}}{x_{ij}} \frac{x_{ij}}{X_i} = E(R_i). \end{aligned}$$

But $E(R_i) = \sum_{i=1}^N \frac{Y_i}{X_i} \frac{X_i}{X} = \frac{Y}{X} = R$.

Hence, \bar{r} is an unbiased estimator of R .

$$V(\bar{r}) = V \left[\frac{1}{n} \sum_{i=1}^n \bar{r}_i \right] = V \left[E \left\{ \frac{1}{n} \sum_i \bar{r}_i/n \right\} \right] + E \left[V \left(\frac{1}{n} \sum \bar{r}_i/n \right) \right]$$

$$= V \left[\frac{1}{n} \sum_{i=1}^n R_i \right] + E \left[\frac{1}{n^2} \sum_{i=1}^n V(\bar{r}_i/n) \right]$$

$$= \frac{1}{n^2} \sum_{i=1}^n V(R_i) + E \left[\frac{1}{n^2} \sum V(\bar{r}_i/i) \right].$$

But $V(R_i) = \sum_{i=1}^N p_i (R_i - \bar{R})^2$ and $V(r_{ij}/i) = \sum_{j=1}^{M_i} p_{ij} (r_{ij} - R_i)^2$.

$$V(\bar{r}_i/i) = \frac{1}{m} \sum_{j=1}^{M_i} \frac{x_{ij}}{X_i} (r_{ij} - R_i)^2.$$

Hence, $V(\bar{r}) = \frac{1}{n^2} \sum \sum \frac{X_i}{X} (R_i - R)^2 + E \left[\frac{1}{n^2 m} \sum \sum \frac{x_{ij}}{X_i} (r_{ij} - \bar{R}_i)^2 \right]$

$$= \frac{1}{n^2} \sum_{i=1}^N \frac{X_i}{X} (R_i - R)^2 + \frac{1}{nm} \sum_{i=1}^N \frac{X_i}{X} \sum_{j=1}^{M_i} \frac{x_{ij}}{X_i} (r_{ij} - R_i)^2.$$

Corollary : In two-stage sampling if sampling units are selected by PPS sampling scheme at each stage, then the unbiased estimator of $V(\bar{r})$ is

$$v(\bar{r}) = \frac{1}{n(n-1)} \sum_{i=1}^n (r_i - \bar{r})^2.$$

Example 17.7 : In an area there are 106 villages. The villages are divided into 10 administrative blocks.

The number of villages in the blocks are different. Each village has cultivable land to produce potato. The amount of potato and the amount of cultivable land for potato are shown below.

SL. No.	No. of villages M_i	Amount of lands in villages (x_{ij} acre)	Amount of potato produced in villages (y_{ij} Q)
1	12	15.6, 12.2, 11.8, 16.6, 15.0, 10.2, 9.6, 12.5, 16.0, 17.2, 10.0, 13.0	240.0, 225.0, 180.0, 280.7, 200.0, 195.5, 150.0, 205.0, 255.0, 240.0, 160.5, 172.0
2	8	10.5, 17.5, 16.2, 8.6, 12.5, 13.0, 14.0, 10.0	200.0, 295.0, 265.0, 190.4, 260.0, 261.2, 242.0, 210.5
3	15	5.2, 6.7, 12.0, 15.0, 7.2, 10.5, 16.0, 11.5, 10.0, 12.6, 18.2, 15.5, 9.5, 6.2, 10.0	140.0, 165.0, 180.0, 270.0, 148.0, 175.0, 140.0, 200.0, 180.0, 192.0, 288.0, 212.0, 150.0, 135.0, 200.0
4	11	16.0, 17.2, 15.0, 8.8, 12.0, 15.0, 6.7, 10.5, 13.5, 14.0, 9.2	295.0, 312.0, 275.0, 165.0, 195.0, 298.0, 140.4, 175.2, 240.5, 260.2, 198.0
5	13	10.0, 11.2, 12.6, 9.9, 14.2, 13.4, 10.2, 11.5, 15.6, 10.2, 7.4, 8.8, 6.7	180.5, 215.0, 220.0, 175.0, 240.0, 250.5, 190.0, 210.0, 280.0, 210.0, 150.0, 165.0, 140.0

SL. No.	No. of villages M_i	Amount of lands in villages (x_{ij} acre)	Amount of potato produced in villages (y_{ij} Q)
6	7	15.5, 20.2, 20.0, 18.6 10.2, 14.2, 9.8	280.0, 260.0, 300.0, 290.0 180.0, 225.0, 160.0
7	10	6.0, 8.7, 12.6, 13.0 15.8, 14.6, 9.5, 10.0 16.2, 20.0	150.0, 165.0, 195.0, 180.0 225.0, 235.0, 200.0, 165.0 235.0, 320.0
8	8	15.0, 18.0, 10.2, 9.5 12.6, 17.2, 11.4, 12.0	300.0, 324.0, 180.0, 195.0 225.0, 275.0, 220.0, 165.0
9	14	12.0, 12.0, 14.6, 18.0 15.8, 16.2, 6.4, 5.5 6.7, 9.5, 10.0, 11.0 10.0, 10.0	195.0, 180.0, 240.0, 310.0 250.8, 195.0, 120.0, 100.0 125.0, 175.0, 175.0, 190.0 145.0, 160.0
10	8	15.0, 16.2, 18.6, 20.0 5.0, 10.2, 12.0, 13.0	245.0, 265.0, 300.0, 320.0 112.0, 220.0, 230.0, 245.0

- (i) Select 5 blocks and 3 villages from each selected block by PPS sampling scheme without replacement.
- (ii) Estimate the per acre production of potato in the area.
- (iii) Estimate the standard error of your estimate.

Solution : (i) Given $N = 10$, $n = 5$, $m = 3$. The blocks are to be selected with probability $p_i = X_i/X$ and blocks are selected by Lahiri's method. The p_i values are shown below :

SL. No.	No. of villages	$X_i = \sum_j x_{ij}$	$p_i = \frac{X_i}{X}$
1	12	159.7	0.1213
2	8	102.3	0.0777
3	15	166.1	0.1262
4	11	137.9	0.1048
5	13	141.7	0.1076
6	7	108.5	0.0824
7	10	126.4	0.0960
8	8	105.9	0.0805
9	14	157.7	0.1198
10	8	110.0	0.0836
Total	106	1316.2	

To select blocks we need to select five pairs of random numbers on the basis of serial number of blocks and the values of X_i . Using random number table of Appendix, the pairs of numbers are (10, 102), (01, 011), (04, 043), (05, 055), (09, 092). Therefore, the selected blocks are 1, 4, 5, 9, 10.

The second-stage units are also selected using Lahiri's method. Pairs of random number are selected using serial number of villages and the values of x_{ij} . The selected pairs of values, selected villages and the probability of selection of villages are shown below :

SL. No. of blocks	Pairs of random numbers	SL. No. of selected villages	$p_{ij} = x_{ij}/X_i$
1	(10,10), (09,09), (01, 01)	10, 9, 1	0.1077, 0.1002, 0.0977
4	(05, 05), (10, 10), (09, 09)	5, 10, 9	0.0807, 0.1015, 0.0979
5	(02, 02), (05, 05), (01, 01)	2, 5, 1,	0.0790, 0.1002, 0.0706
9	(01, 01), (07, 07), (08, 08)	1, 7, 8	0.0761, 0.0406, 0.0349
10	(8, 8), (5, 5), (7,7)	8, 5, 7	0.1182, 0.0454, 0.1091

Selected observations in the sample.

SL. No.	x_{ij}	y_{ij}	$r_{ij} = \frac{y_{ij}}{x_{ij}}$	$\bar{r}_i = \frac{\sum r_{ij}}{m}$
1	17.2, 16.0, 15.6	240.0, 255.0, 240.0	13.95, 15.94, 15.38	15.09
4	12.0, 14.0, 13.5	195.0, 260.2, 240.5	16.25, 18.59, 17.81	17.55
5	11.2, 14.2, 10.0	215.0, 240.0, 180.5	19.20, 16.90, 18.05	18.05
9	12.0, 6.4, 5.5	195.0, 120.0, 100.0	16.25, 18.75, 18.18	17.73
10	13.0, 5.0, 12.0	245.0, 112.0, 230.0	18.85, 22.40, 19.17	20.14

(ii) Per acre production of potato in the study area is

$$\bar{r} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m r_{ij} = \frac{1}{n} \sum_{i=1}^n \bar{r}_i = \frac{88.56}{5} = 17.712 Q.$$

(iii) The estimated variance of the estimator of per acre production of potato is

$$\begin{aligned} v(\bar{r}) &= \frac{1}{n(n-1)} \sum (\bar{r}_i - \bar{r})^2 = \frac{1}{n(n-1)} \left[\sum \bar{r}_i^2 - n\bar{r}^2 \right] \\ &= \frac{1}{5(5-1)} [1581.4856 - 5(17.712)^2] = 0.645544. \end{aligned}$$

$$\therefore \text{s.e.}(\bar{r}) = \sqrt{v(\bar{r})} = 0.8034.$$

Chapter 18

Three-Stage Sampling

18.1 Three-Stage Sampling with Equal First-Stage Units

Let there be NMT elements in a population. The elements are divided into NM sub-groups of T elements each. The sub-groups are divided into N clusters of M elements each. The elements in sub-groups are called *third-stage units*, the N clusters are called *primary-stage units* and the M elements in each cluster are called *second-stage units*. The problem is to select a random sample of nmt elements.

Consider that n primary units are selected by SRSWOR technique. In each selected cluster, there are MT elements divided into M second-stage units. A simple random sample without replacements of m second-stage units are to be selected for the sample. In each selected second-stage unit, there are T third-stage units. A simple random sample of t units are to be selected without replacement. The resultant sample is a three-stage sample.

Let y_{ijl} be the value of the variable under study which is recorded from l th third-stage unit in j th second-stage unit within i th first-stage unit [$i = 1, 2, \dots, N; j = 1, 2, \dots, M; l = 1, 2, \dots, T$]. Then

$$\bar{Y}_{ij} = \frac{1}{T} \sum_{l=1}^T y_{ijl} = \text{the mean of third-stage units.}$$

$$\bar{Y}_i = \frac{1}{MT} \sum_j \sum_l y_{ijl} = \frac{1}{M} \sum_{j=1}^M \bar{Y}_{ij} = \text{the mean of observations of } i\text{th first-stage unit.}$$

The population mean of the observations is given by

$$\bar{Y} = \frac{1}{NMT} \sum_{i=1}^N \sum_{j=1}^M \sum_{l=1}^T y_{ijl} = \frac{1}{NM} \sum_i \sum_j \bar{Y}_{ij} = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i.$$

The corresponding sample means of the above population means are

$$\bar{y}_{ij} = \frac{1}{t} \sum_{l=1}^t y_{ijl}, \quad \bar{y}_i = \frac{1}{mt} \sum \sum y_{ijl} = \frac{1}{m} \sum_j \bar{y}_{ij}$$

$$\text{and } \bar{y} = \frac{1}{nmt} \sum \sum \sum y_{ijl} = \frac{1}{nm} \sum_i \sum_j \bar{y}_{ij} = \frac{1}{n} \sum \bar{y}_i.$$

Let us now define the variances of observations of different stages. The variances are

$$S_{ij}^2 = \frac{1}{T-1} \sum_{l=1}^T (y_{ijl} - \bar{Y}_{ij})^2 = \text{the variance of observations of third-stage units of } j\text{th second-stage unit in } i\text{th first-stage unit.}$$

$$S_i^2 = \frac{1}{M-1} \sum_{j=1}^M (\bar{Y}_{ij} - \bar{Y}_i)^2 = \text{the variance of the means of } j\text{th second-stage unit in } i\text{th first-stage unit.}$$

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2 = \text{the variance of the means of first-stage units.}$$

Also, we can define

$$S_w^2 = \frac{1}{N} \sum_{i=1}^N S_i^2 = \frac{1}{N(M-1)} \sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_i)^2$$

$$S_T^2 = \frac{1}{NM} \sum_i \sum_j S_{ij}^2 = \frac{1}{NM(T-1)} \sum_i \sum_j \sum_{t=1}^T (y_{ijt} - \bar{Y}_{ij})^2.$$

Theorem : In three-stage sampling, if n first-stage units are selected and from each selected first-stage units m second-stage units are selected and finally if from each second-stage unit t third stage units are selected by SRSWOR scheme at each stage, then the sample mean \bar{y} is an unbiased estimator of population mean \bar{Y} . The variance of \bar{y} is

$$V(\bar{y}) = \frac{N-n}{nN} S_b^2 + \frac{M-m}{mM} \frac{S_w^2}{n} + \frac{T-t}{tT} \frac{S_T^2}{nm}.$$

Proof : We have $\bar{y} = \frac{1}{nmt} \sum \sum \sum y_{ijt} = \frac{1}{nm} \sum_i \sum_j \bar{y}_{ij} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$.

$$\begin{aligned} \text{Now } E(\bar{y}) &= E_1 \left[E_2 \left\{ E_3 \frac{1}{nm} \sum_i \sum_j \bar{y}_{ij} \right\} \right] = E_1 \left[E_2 \left\{ \frac{1}{nm} \sum_i \sum_j E_3(\bar{y}_{ij}) \right\} \right] \\ &= E_1[E_2(\bar{Y}_{ij})], \quad \because \bar{y}_{ij} \text{ is a simple random sample mean of } t \text{ observations} \\ &= E_1 \left[\frac{1}{M} \sum_{j=1}^M \bar{Y}_{ij} \right], \quad \because j\text{th second-stage units are selected by SRSWOR scheme.} \\ &= E[\bar{Y}_i] = \frac{1}{N} \sum_{i=1}^M \bar{Y}_i = \bar{Y}, \quad \because i\text{th first-stage units are selected by SRSWOR scheme.} \end{aligned}$$

Here E_1 , E_2 and E_3 are used to indicate the expectation of first-stage, second-stage and third-stage samplings, respectively.

$$\begin{aligned} V(\bar{y}) &= V \left[\frac{1}{n} \sum \bar{y}_i \right] = V \left[E \left(\frac{1}{n} \sum \bar{y}_{i/n} \right) \right] + E \left[V \left(\frac{1}{n} \sum \bar{y}_{i/n} \right) \right] \\ &= V \left[\frac{1}{n} \sum_{i=1}^n \bar{Y}_i \right] + E \left[\frac{1}{n^2} \sum_{i=1}^n V(\bar{y}_{i/n}) \right]. \end{aligned}$$

But \bar{y}_i is the mean observed from two-stage sampling. Hence, its variance is

$$V(\bar{y}_i) = \frac{M-m}{mM} S_i^2 + \frac{T-t}{tT} \frac{1}{nm} \sum_{j=1}^M S_{ij}^2.$$

Again, $V\left[\frac{1}{n}\sum\bar{Y}_i\right] = \frac{N-n}{nN}S_b^2$, \therefore this mean is a mean of simple random sample.

$$\begin{aligned} \text{Hence, } V(\bar{y}) &= V\left[\frac{1}{n}\sum_i^n\bar{Y}_i\right] + E\left[\frac{1}{n^2}\sum_{i=1}^n\left\{\frac{M-m}{mM}S_i^2 + \frac{T-t}{tT}\frac{1}{nm}\sum_j^M S_{ij}^2\right\}\right] \\ &= \frac{N-n}{nN}S_b^2 + \frac{1}{nN}\frac{M-m}{mM}\sum_{i=1}^N S_i^2 + \frac{1}{nN}\frac{T-t}{tT}\frac{1}{mM}\sum_i^N\sum_{j=1}^M S_{ij}^2 \\ &= \frac{N-n}{nN}S_b^2 + \frac{1}{n}\frac{M-m}{mM}S_w^2 + \frac{1}{nm}\frac{T-t}{tT}S_T^2 \\ &= (1-f)\frac{S_b^2}{n} + \frac{1}{n}(1-f_1)\frac{S_w^2}{m} + (1-f_2)\frac{S_T^2}{nmt}, \end{aligned}$$

where $f = \frac{n}{N}$, $f_1 = \frac{m}{M}$, $f_2 = \frac{t}{T}$.

$$V(\bar{y}) = \frac{S_b^2}{n} + \frac{S_w^2}{nm} + \frac{S_T^2}{nmt}, \text{ if } f, f_1 \text{ and } f_2 \text{ are negligible.}$$

It can also be mentioned that $V(\bar{y})$ is

$$V(\bar{y}) = \frac{S_b^2}{n} + \frac{S_w^2}{nm} + \frac{S_T^2}{nmt},$$

if sample is selected with replacement at each stage.

If all third-stage units are included in the sample ($t = T$), then the variance stands

$$V(\bar{y}) = \frac{S_b^2}{n} + \frac{S_w^2}{nm}.$$

Again, if all elements of first-stage units are included in the sample ($m = M$), then

$$V(\bar{y}) = \frac{S_b^2}{n}.$$

If second-stage units are selected from all primary-stage units ($n = N$), then

$$V(\bar{y}) = (1-f_1)\frac{S_w^2}{mn} + (1-f_2)\frac{S_T^2}{nmt}.$$

Corollary : In three-stage sampling if sample observations are selected by SRSWOR scheme at each stage, then the estimator of variance of sample mean is

$$v(\bar{y}) = (1-f)\frac{s_b^2}{n} + (1-f_1)\frac{s_w^2}{Nm} + (1-f_2)\frac{s_T^2}{NMt}.$$

where $s_b^2 = \frac{1}{n-1}\sum(\bar{y}_i - \bar{y})^2$, $s_w^2 = \frac{1}{n}\sum_{i=1}^n s_i^2$, $s_T^2 = \frac{1}{nm}\sum_i^n\sum_j^m s_{ij}^2$,

$$s_i^2 = \frac{1}{m-1}\sum_j^m(\bar{y}_{ij} - \bar{y}_i)^2, s_{ij}^2 = \frac{1}{t-1}\sum_{l=1}^t(\bar{y}_{ijl} - \bar{y}_{ij})^2.$$

Example 18.1 : For administrative convenience the entire garments industry of an owner is divided into 10 units. In each unit there are 10 machines. The products of each machine are recorded after every two-hours. The machines run from morning 7.00 AM to 7.00 PM. in the evening. To estimate the total production (in pieces) of the industry in a day, 5 units are randomly selected, from each selected unit 4 machines are randomly selected and the products of selected machines are recorded four times at random. The products of selected machines are shown below :

Primary units	Second-stage units	Products of selected machines (y_{ijl})	\bar{y}_{ij}	$s_{ij}^2 = \frac{1}{t-1} \left[\sum y_{ijl}^2 - \frac{(\sum y_{ijl})^2}{t} \right]$
1	1	15, 18, 16, 15	16.0	2.00
	2	18, 20, 18, 20	19.0	1.3333
	3	14, 12, 12, 12	12.5	1.00
	4	16, 16, 16, 16	16.0	0.00
2	1	11, 10, 12, 12	11.25	0.9167
	2	13, 15, 13, 13	13.50	1.00
	3	17, 17, 17, 16	16.75	0.25
	4	15, 16, 16, 16	15.75	0.25
3	1	18, 18, 18, 18	18.0	0.00
	2	17, 17, 18, 18	17.5	0.3333
	3	15, 15, 15, 16	15.25	0.25
	4	14, 14, 14, 15	14.25	0.25
4	1	12, 11, 12, 12	11.75	0.25
	2	13, 15, 15, 14	14.25	0.9167
	3	15, 15, 15, 15	15.00	0.00
	4	16, 15, 16, 14	15.25	0.9167
5	1	10, 12, 12, 10	11.00	0.3333
	2	15, 15, 15, 15	15.00	0.00
	3	18, 18, 18, 19	18.25	0.25
	4	14, 14, 14, 14	14.00	0.00
Total			300.25	10.25

Estimate the total production of a day. Also estimate the standard error of your estimator.

Solution : Given $N = 10$, $M = 10$, $T = 6$, $n = 5$, $m = 4$, $t = 4$.

The average production per machine is

$$\bar{y} = \frac{1}{nm} \sum_i^n \sum_j^m \bar{y}_{ij} = \frac{300.25}{5 \times 4} = 15.0125.$$

Estimate of total production in all machine is

$$\hat{Y} = NMT\bar{y} = 10 \times 10 \times 6 \times 15.0125 = 9007.5 \approx 9008.$$

Sl.No.	$\bar{y}_i = \frac{1}{m} \sum_j \bar{y}_{ij}$	$s_i^2 = \frac{1}{m-1} \left[\sum \bar{y}_{ij}^2 - \frac{(\sum \bar{y}_{ij})^2}{m} \right]$
1	15.875	7.0625
2	14.3125	6.0156
3	16.25	3.2083
4	14.0625	2.5573
5	14.5625	8.9323
Total	75.0625	27.776

$$s_w^2 = \frac{1}{n} \sum_{i=1}^n s_i^2 = \frac{27.776}{5} = 5.5552.$$

$$s_b^2 = \frac{1}{n-1} \left[\sum \bar{y}_i^2 - \frac{(\sum \bar{y}_i)^2}{n} \right] = \frac{1}{5-1} \left[1130.746094 - \frac{(75.0625)^2}{5} \right] = 0.967578.$$

$$s_T^2 = \frac{1}{nm} \sum \sum s_{ij}^2 = \frac{10.25}{5 \times 4} = 0.5125.$$

$$v(\bar{y}) = (1-f) \frac{s_b^2}{n} + (1-f_1) \frac{s_w^2}{Nm} + (1-f_2) \frac{s_T^2}{NMT},$$

where $f_1 = \frac{n}{N} = 0.5$, $f_2 = \frac{m}{M} = 0.4$, $f_3 = \frac{t}{T} = 0.67$.

$$v(\bar{y}) = (1-0.5) \frac{0.967578}{5} + (1-0.4) \frac{5.5552}{10 \times 4} + (1-0.67) \frac{0.5125}{10 \times 10 \times 4} = 0.1805086.$$

Hence, $v(\hat{Y}) = (NMT)^2 v(\bar{y}) = (10 \times 10 \times 6)^2 \times 0.1805086 = 64983.1005$.

\therefore s.e. $(\hat{Y}) = \sqrt{v(\hat{Y})} = 254.92$.

18.2 Three-Stage Sampling with Unequal First-Stage and Second-Stage Units

Let there be T_0 elements in a population. The population elements are divided into N clusters. Consider that the size of i th cluster is M_i ($i = 1, 2, \dots, N$). Thus, the primary units are unequal. Also consider that each cluster is sub-divided into sub-clusters and the size of j th sub-cluster (second-stage units) in i th cluster is T_{ij} ($j = 1, 2, \dots, M_i$). The M_i elements in each sub-cluster are called third-stage or ultimate sampling units. Here

$$T_0 = \sum_{i=1}^N \sum_{j=1}^{M_i} T_{ij} = \sum_{i=1}^N T_i = \sum_{i=1}^N M_i \bar{T}_i, \quad \bar{T}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} T_{ij}.$$

Let y_{ijl} be the value of a variable under study recorded from l th third-stage unit of j th second-stage unit within i th cluster (first-stage unit). For example, let us consider the estimation of dropout students from primary schools in a district. The district has some police stations. In each police station area there are some administrative blocks. In each block there are some primary schools. Here primary schools within a block can be considered as third-stage units; the blocks are second-stage units and the police stations are primary units. For the survey

to estimate the number of drop out students in a year some police stations can be selected at random. From the selected police stations some administrative blocks and from the selected blocks some primary schools can be selected at random. Here y_{ijl} is the number of dropout students in a year from l th primary school in j th block within i th police stations. Then,

$$\bar{Y}_{ij} = \frac{1}{T_{ij}} \sum_{l=1}^{T_{ij}} y_{ijl} = \text{mean per element of third-stage unit in } j\text{th second-stage unit within } i\text{th cluster}$$

$$\begin{aligned} \bar{Y}_i &= \frac{1}{T_i} \sum_{j=1}^{M_i} \sum_{l=1}^{T_{ij}} y_{ijl} = \frac{1}{T_i} \sum_{j=1}^{M_i} T_{ij} \bar{Y}_{ij} = \frac{1}{M_i T_i} \sum_{j=1}^{M_i} T_{ij} \bar{Y}_{ij} \\ &= \frac{1}{M_i} \sum_{j=1}^{M_i} k_{ij} \bar{Y}_{ij} = \text{mean of } i\text{th first-stage unit.} \end{aligned}$$

Here $T_i = \sum_{j=1}^{M_i} T_{ij}$ and $k_{ij} = \frac{T_{ij}}{T_i}$.

$$\begin{aligned} \bar{Y} &= \frac{1}{T_0} \sum_{i=1}^N \sum_{j=1}^{M_i} \sum_{l=1}^{T_{ij}} y_{ijl} = \frac{1}{T_0} \sum_i \sum_j T_{ij} \bar{Y}_{ij} = \frac{1}{\sum M_i \bar{T}_i} \sum_i \sum_j T_{ij} \bar{Y}_{ij} \\ &= \frac{1}{\sum M_i \bar{T}_i} \sum_{i=1}^N \bar{T}_i M_i \bar{Y}_i = \frac{1}{\sum_{i=1}^N Q_i} \sum Q_i \bar{Y}_i = \frac{1}{N} \sum_{i=1}^N W_i \bar{Y}_i \end{aligned}$$

= population mean.

Here $Q_i = M_i \bar{T}_i$, $\bar{Q} = \frac{1}{N} \sum_{i=1}^N Q_i$, $W_i = \frac{Q_i}{\bar{Q}}$.

Let us consider that, for a survey n primary units are selected. From the selected i th primary units m_i second-stage units and from the selected second-stage units t_{ij} third-stage units are selected. At each stage the selection is done by SRSWOR ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, m_i$; $l = 1, 2, \dots, t_{ij}$). Then

$$\bar{y}_{ij} = \frac{1}{t_{ij}} \sum_{l=1}^{t_{ij}} y_{ijl} = \text{mean per element of selected third-stage units.}$$

Let us define a simple mean $\bar{y}_{3s} = \frac{1}{n} \sum_{i=1}^n W_i \frac{1}{m_i} \sum_{j=1}^{m_i} k_{ij} \bar{y}_{ij}$.

Theorem : In three-stage sampling, if first-stage and second-stage units are unequal and if at each stage sampling units are selected by SRSWOR scheme, then the sample mean \bar{y}_{3s} is an unbiased estimator of population mean \bar{Y} . The variance of \bar{y}_{3s} is

$$V(\bar{y}_{3s}) = \frac{N-n}{nN} S_{1b}^2 + \frac{1}{nN} \sum_{i=1}^N W_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) S_{1i}^2 + \frac{1}{nN} \sum_{i=1}^N \frac{W_i^2}{m_i M_i} \sum_{j=1}^{m_i} k_{ij}^2 \left(\frac{T_{ij} - t_{ij}}{t_{ij} T_{ij}} \right) S_{ij}^2.$$

$$\text{Here } S_{1b}^2 = \frac{1}{N-1} \sum_{i=1}^N (W_i \bar{Y}_i - \bar{Y})^2, S_{1i}^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} (k_{ij} \bar{Y}_{ij} - \bar{Y}_i)^2, S_{ij}^2 = \frac{1}{T_{ij}-1} \sum_{l=1}^{T_{ij}} (y_{ijl} - \bar{Y}_{ij})^2.$$

$$\text{Proof: We have } \bar{y}_{3s} = \frac{1}{n} \sum_{i=1}^n W_i \frac{1}{m_i} \sum_{j=1}^{m_i} k_{ij} \bar{y}_{ij}.$$

$$\begin{aligned} E(\bar{y}_{3s}) &= E_1 \left[\frac{1}{n} \sum_{i=1}^n W_i \left\{ E_2 \frac{1}{m_i} \sum_{j=1}^{m_i} E_3(k_{ij} \bar{y}_{ij}) \right\} \right] = E_1 \left[\frac{1}{n} \sum_{i=1}^n W_i \left\{ E_2 \frac{1}{m_i} \sum_{j=1}^{m_i} k_{ij} \bar{Y}_{ij} \right\} \right] \\ &= E_1 \left[\frac{1}{n} \sum_{i=1}^n W_i \frac{1}{M_i} \sum_{j=1}^{M_i} k_{ij} \bar{Y}_{ij} \right] = E_1 \left[\frac{1}{n} \sum_{i=1}^n W_i \frac{1}{M_i} \sum_{j=1}^{M_i} \frac{T_{ij} M_i \bar{Y}_{ij}}{T_i} \right], \quad T_i = \sum_{j=1}^{M_i} T_{ij} \\ &= E_1 \frac{1}{n} \sum_{i=1}^n W_i \bar{Y}_i = \frac{1}{n} \sum_{i=1}^n \frac{1}{N} \sum_{i=1}^N W_i \bar{Y}_i = \bar{Y}. \end{aligned}$$

Hence, \bar{y}_{3s} is an unbiased estimator of \bar{Y} .

$$V(\bar{y}_{3s}) = V \left[\frac{1}{n} \sum_{i=1}^n W_i \frac{1}{m_i} \sum_{j=1}^{m_i} k_{ij} \bar{y}_{ij} \right] = V_1 E_2 E_3(\bar{y}_{3s}) + E_1 V_2 E_3(\bar{y}_{3s}) + E_1 E_2 V(\bar{y}_{3s}).$$

Now, using the concept of variance of sample mean in case of two-stage sampling with unequal first-stage units, we can write :

$$V(\bar{y}_{3s}) = \frac{N-n}{nN} S_{1b}^2 + \frac{1}{nN} \sum_{i=1}^n W_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) S_{1i}^2 + \frac{1}{nN} \sum_{i=1}^n \frac{W_i}{m_i M_i} \sum_{j=1}^{M_i} k_{ij}^2 \left(\frac{T_{ij} - t_{ij}}{t_{ij} T_{ij}} \right) S_{ij}^2.$$

Corollary : In three-stage sampling in case of unequal first-stage and second-stage units if sampling units are selected by SRSWOR scheme at every stage, the estimator of variance of \bar{y}_{3s} is

$$v(\bar{y}_{3s}) = \frac{N-n}{nN} s_{1b}^2 + \frac{1}{nN} \sum_{i=1}^n W_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) s_{1i}^2 + \frac{1}{nN} \sum_{i=1}^n \frac{W_i^2}{m_i M_i} \sum_{j=1}^{m_i} k_{ij}^2 \left(\frac{T_{ij} - t_{ij}}{t_{ij} T_{ij}} \right) s_{ij}^2,$$

$$\text{where } s_{1b}^2 = \frac{1}{n-1} \sum_{i=1}^n (W_i \bar{y}_i - \bar{y}_{3s})^2, \quad s_{1i}^2 = \frac{1}{m_i-1} \sum_{j=1}^{m_i} (k_{ij} \bar{y}_{ij} - \bar{y}_i)^2,$$

$$s_{ij}^2 = \frac{1}{t_{ij}-1} \sum_{l=1}^{t_{ij}} (y_{ijl} - \bar{y}_{ij})^2, \quad \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} k_{ij} \bar{y}_{ij}.$$

Example 18.2 : In a tea garden the tea plants are planted into 50 small areas, where each area is divided into several blocks. In each block, different number of plants are nurtured. To estimate the total production of tea in a day 10 areas are randomly selected. From the randomly selected areas 20% blocks are selected randomly and from each selected block 10% plants are selected at random. In each stage of selection sampling is done by SRSWOR. The tea production of a day is recorded from each plant. The data are shown below :

Sl. No. of area	Number of blocks in each area		Number of plants in a block		Production of tea in a day (in kg) y_{ij}
	In population M_i	In sample m_i	In population T_{ij}	In sample t_{ij}	
1	15	3	20	2	0.25, 0.22
			40	4	0.30, 0.18, 0.20, 0.15
			20	2	0.18, 0.20
2	20	4	25	3	0.13, 0.15, 0.14
			28	3	0.18, 0.22, 0.26
			20	2	0.17, 0.19
			30	3	0.20, 0.22, 0.24
3	16	3	16	2	0.20, 0.17
			18	2	0.18, 0.15
			20	2	0.12, 0.13
4	25	5	30	3	0.15, 0.16, 0.16
			25	3	0.14, 0.18, 0.17
			22	2	0.20, 0.20
			20	2	0.25, 0.15
			30	3	0.10, 0.15, 0.18
5	20	4	18	2	0.22, 0.18
			20	2	0.15, 0.16
			25	3	0.15, 0.15, 0.15
			20	2	0.15, 0.18
6	10	2	40	4	0.10, 0.10, 0.12, 0.12
			45	5	0.10, 0.12, 0.14, 0.12, 0.10
7	18	4	25	3	0.15, 0.16, 0.15
			20	2	0.18, 0.18
			20	2	0.20, 0.22
			30	3	0.20, 0.20, 0.18
8	20	4	30	3	0.15, 0.16, 0.16
			40	4	0.18, 0.18, 0.20, 0.20
			30	3	0.20, 0.15, 0.12
			25	3	0.16, 0.16, 0.18
9	12	2	36	4	0.20, 0.20, 0.20, 0.20
			38	4	0.21, 0.18, 0.16, 0.16
10	15	3	30	3	0.16, 0.18, 0.22
			22	2	0.25, 0.24
			28	3	0.25, 0.22, 0.22
Total	171	$m_{1i} = 34$	906	$t_{10} = 95$	

- (i) Estimate the total tea production in the garden in a day. Given the total number of plants in the garden is 2000.

(ii) Estimate standard error of your estimator.

Solution : Given $N = 50$, $M_1 = \sum M_i = 171$, $T_1 = 80$, $T_2 = 103$, $T_3 = 54$, $T_4 = 127$, $T_5 = 83$, $T_6 = 85$, $T_7 = 95$, $T_8 = 125$, $T_9 = 74$, $T_{10} = 80$.

Calculation Related to Estimator of Mean and Variance

Sl. No.	\bar{T}_i	$k_{ij} = \frac{T_{ij}}{T_i}$	$Q_i = M_i \bar{T}_i$	$W_i = \frac{Q_i}{Q}$	\bar{y}_{ij}	$s_{ij}^2 = \frac{1}{t_{ij}-1} \left[\sum y_{ijl}^2 - \frac{(\sum y_{ijl})^2}{t_{ij}} \right]$
1	26.67	0.75	400.05	0.875	0.235	0.00045
		1.50			0.2075	0.00422
		0.75			0.19	0.00020
2	25.75	0.97	515	1.126	0.14	0.0001
		1.09			0.22	0.0016
		0.78			0.18	0.0002
		1.16			0.22	0.0004
3	18	0.89	288	0.629	0.185	0.00045
		1.00			0.165	0.00045
		1.11			0.125	0.00005
4	25.4	1.18	635	1.388	0.157	0.00003
		0.98			0.163	0.00043
		0.87			0.20	0.00
		0.79			0.20	0.005
		1.18			0.143	0.00163
5	20.75	0.87	415	0.907	0.20	0.0008
		0.96			0.155	0.00005
		1.20			0.15	0.00
		0.96			0.165	0.00045
6	42.5	0.94	425	0.929	0.11	0.00013
		1.06			0.116	0.00028
7	23.75	1.05	427.5	0.935	0.153	0.00003
		0.84			0.18	0.00
		0.84			0.21	0.0002
		1.26			0.193	0.00013
8	31.25	0.96	625	1.366	0.157	0.00003
		1.28			0.19	0.00013
		0.96			0.157	0.00163
		0.80			0.167	0.00013
9	37	0.97	444	0.971	0.20	0.00
		1.03			0.177	0.00056
10	26.67	1.12	400	0.874	0.187	0.00093
		0.82			0.245	0.00005
		1.05			0.23	0.0003

$$\bar{Q} = \frac{1}{N} \sum Q_i = \frac{4574.55}{10} = 457.455$$

Calculation continued

Sl. No.	$k_{ij}\bar{y}_{ij}$	$\bar{y}_i = \frac{\sum k_{ij}\bar{y}_{ij}}{m_i}$	$s_{1i}^2 = \frac{\sum (k_{ij}\bar{y}_{ij} - \bar{y}_i)^2}{m_i - 1}$	$\sum_j k_{ij}^2 \left(\frac{T_{ij} - t_{ij}}{t_{ij}T_{ij}} \right) s_{ij}^2$	$W_i\bar{y}_i$
1	0.17625 0.31125 0.1425	0.21	0.0080992	0.0043814	0.18375
2	0.1358 0.2398 0.1404 0.2552	0.1928	0.0040325	0.0008096	0.21709
3	0.16465 0.165 0.13875	0.1561	0.0002267	0.0003837	0.09819
4	0.18526 0.15974 0.174 0.158 0.16874	0.1691	0.0001240	0.0022188	0.23471
5	0.174 0.1488 0.18 0.1584	0.1653	0.0002039	0.0004765	0.14993
6	0.1034 0.12296	0.1132	0.0001913	0.0000818	0.10516
7	0.16065 0.1512 0.1764 0.24318	0.1828	0.0017253	0.0001351	0.17092
8	0.15072 0.2432 0.15072 0.1336	0.1696	0.0024753	0.0005313	0.23167
9	0.194 0.18231	0.1881	0.0000683	0.0001329	0.18265
10	0.20944 0.2009 0.2415	0.2173	0.0004582	0.0004637	0.18992

Now, we have the estimator of mean production

$$\bar{y}_{3s} = \frac{1}{n} \sum_{i=1}^n W_i \bar{y}_i = \frac{1.76399}{10} = 0.176399$$

Calculation continued

Sl. No.	$(w_i \bar{y}_i - \bar{y}_{3s})^2$	$w_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) s_{1i}^2$	$\frac{w_i^2}{m_i M_i} \left[\sum_j k_{ij}^2 \left(\frac{T_{ij} - t_{ij}}{t_{ij} T_{ij}} \right) s_{ij}^2 \right]$
1	0.00005404	0.00165359	0.00007754
2	0.00165576	0.00102254	0.00001283
3	0.00611665	0.00002429	0.00000316
4	0.00340017	0.00003822	0.00003420
5	0.00070061	0.00003355	0.00000490
6	0.00507499	0.00006604	0.00000353
7	0.00003002	0.00029328	0.000001640
8	0.00305488	0.00092376	0.00001239
9	0.00003907	0.00002683	0.00000421
10	0.00018282	0.00009334	0.00000787
Total	0.02030901	0.00417544	0.00016227

The estimate of total tea production in a day is

$$\hat{Y} = T_0 \bar{y}_{3s} = 2000 \times 0.176399 = 352.80 \text{ kg.}$$

$$(ii) \text{ We have } s_{1b}^2 = \frac{1}{n-1} \sum (W_i \bar{y}_i - \bar{y}_{3s})^2 = \frac{0.02030901}{10-1} = 0.00225656.$$

$$\begin{aligned} v(\bar{y}_{3s}) &= \frac{N-n}{nN} s_{1b}^2 + \frac{1}{nN} \sum_{i=1}^n W_i^2 \left(\frac{M_i - m_i}{m_i M_i} \right) s_{i1}^2 \\ &\quad + \frac{1}{nN} \sum_{i=1}^n \frac{W_i^2}{m_i M_i} \sum_{j=1}^{m_i} k_{ij}^2 \left(\frac{T_{ij} - t_{ij}}{t_{ij} T_{ij}} \right) s_{ij}^2 \\ &= \frac{50-10}{10 \times 50} \times 0.00225656 + \frac{1}{10 \times 50} \times 0.00417544 + \frac{1}{10 \times 50} \times 0.00016227 \\ &= 0.00018920022. \end{aligned}$$

$$\text{Hence, } v(\hat{Y}) = T_0^2 v(\bar{y}_{3s}) = (2000)^2 \times 0.00018920022 = 756.8009.$$

$$\text{s.e. } (\hat{Y}) = \sqrt{v(\hat{Y})} = 27.51.$$

18.3 Allocation of Sample Sizes in Three-Stages

Let there be NMT elements in a population. The elements are divided into N clusters. There are MT elements in each cluster. Each cluster is sub-divided into M sub-clusters. In each sub-cluster there are T elements. The clusters are primary units, the sub-clusters are second-stage units and the elements in each sub-cluster are known as third-stage units. The problem is to select n primary units, m second-stage units and t third-stage units. We need a decision about the values of n , m and t .

Let us consider that the survey is to be conducted within the limit of a fixed cost C_0 . Let the cost function for the survey be

$$C = C_1 n + C_2 nm + C_3 nmt,$$

where C_1 , C_2 and C_3 are the cost to include each first-stage, second-stage and third-stage units, respectively in the sample. The values of n , m and t are to be found out so that the survey

is completed within the fixed cost but the variance of the estimator is minimum. We know variance of the sample mean is

$$\begin{aligned} V(\bar{y}) &= \frac{N-n}{nN} S_b^2 + \frac{1}{n} \frac{M-m}{mM} S_w^2 + \frac{1}{nm} \frac{T-t}{tT} S_T^2 \\ &= (1-f) \frac{S_b^2}{n} + (1-f_1) \frac{S_w^2}{nm} + (1-f_2) \frac{S_T^2}{nmt}. \end{aligned}$$

The values of n , m and t are also found out so that the $V(\bar{y})$ is fixed but the cost of the survey is minimum. We have

$$n = \frac{C}{(C_1 + C_2 m + C_3 m t)}.$$

Putting the value of n in the variance formula and on simplification, we get

$$\begin{aligned} C \left[V + \frac{S_b^2}{N} \right] &= \left\{ C_2 \left(S_b^2 - \frac{S_w^2}{M} \right) m + \left(S_w^2 - \frac{S_T^2}{T} \right) \frac{C_1}{m} \right\} + \left\{ C_3 \left(S_w^2 - \frac{S_T^2}{T} \right) t + \frac{C_2 S_T^2}{t} \right\} \\ &\quad + \left\{ C_3 \left(S_b^2 - \frac{S_w^2}{M} \right) m t + \frac{C_1 S_T^2}{m t} \right\} + \text{quantities independent of } m \text{ and } t. \end{aligned}$$

Here $V(\bar{y}) = V$. This variance function is to be minimised with respect to m and t .

Let us consider that $(S_b^2 - S_w^2/M)$ and $(S_w^2 - S_T^2/T)$ are positive. Then the variance given above can be written as

$$\begin{aligned} C \left[V + \frac{S_b^2}{N} \right] &= \left[\sqrt{C_2 \left(S_b^2 - \frac{S_w^2}{M} \right) m} - \sqrt{\frac{1}{m} C_1 \left(S_w^2 - \frac{S_T^2}{T} \right)} \right]^2 \\ &\quad + \left[\sqrt{C_3 \left(S_w^2 - \frac{S_T^2}{T} \right) t} - \sqrt{\frac{C_2 S_T^2}{t}} \right]^2 + \left[\sqrt{C_3 \left(S_b^2 - \frac{S_w^2}{M} \right) m t} - \sqrt{\frac{C_1 S_T^2}{m t}} \right]^2 \\ &\quad + \text{terms free of } m \text{ and } t. \end{aligned}$$

This variance function is minimum when the three squares in the right-hand side are zero. Then, we have

$$\hat{m} = \sqrt{\frac{C_1 \left(S_w^2 - \frac{S_T^2}{T} \right)}{C_2 \left(S_b^2 - \frac{S_w^2}{M} \right)}} \quad \text{and} \quad \hat{t} = \sqrt{\frac{C_2 S_T^2}{C_3 \left(S_w^2 - \frac{S_T^2}{T} \right)}}.$$

However, \hat{m} and \hat{t} will be the integer.

Let us now consider that $\left(S_b^2 - \frac{S_w^2}{M} \right) < 0$ and $\left(S_w^2 - \frac{S_T^2}{T} \right) > 0$.

In such a case, the value of \hat{m} will be maximum. The value of \hat{m} may be M . The value of \hat{t} is to be decided so that

$$t C_3 \left(S_w^2 - \frac{S_T^2}{T} \right) + \frac{1}{t} C_2 S_T^2 + \hat{m} t C_3 \left(S_b^2 - \frac{S_w^2}{M} \right) + \frac{1}{\hat{m} t} C_1 S_T^2$$

becomes minimum. This will be minimum, if

$$\left\{ C_3 \left(S_w^2 - \frac{S_T^2}{T} \right) + C_3 \hat{m} \left(S_b^2 - \frac{S_w^2}{M} \right) \right\} < 0.$$

But if this term is positive, then

$$t = \sqrt{\frac{\left(\frac{1}{m}C_1 + C_2\right) S_T^2}{C_3 \left(S_w^2 - \frac{S_T^2}{T}\right) + mC_3 \left(S_T^2 - \frac{S_w^2}{M}\right)}}.$$

But t is an integer.

Let us now consider that $S_b^2 - \frac{S_w^2}{M}$ and $S_w^2 - \frac{S_T^2}{T}$ are negative quantities or both are zeros. In such a case, the variance function will be minimum, if t takes the maximum value. The maximum value of t is $t = T$. Also the value of m is to be estimated in such a way that the quantity

$$C_2 \left(S_b^2 - \frac{S_w^2}{M}\right) m + \left(S_w^2 - \frac{S_T^2}{T}\right) \frac{C_1}{m} + mC_3 \left(S_b^2 - \frac{S_w^2}{M}\right) + \frac{1}{mt} C_1 S_T^2$$

is minimum. The term can be written as

$$m \left(S_b^2 - \frac{S_w^2}{M}\right) (C_2 + tC_3) + \frac{C_1}{m} \left(\frac{1}{t} S_T^2 - \frac{S_T^2}{T} + S_w^2\right)$$

and it will be minimum if m takes maximum value.

Let us again consider the $S_b^2 - \frac{S_w^2}{M} > 0$ and $S_w^2 - \frac{S_T^2}{T} < 0$. Under the above conditions the integer value of tm is

$$\sqrt{\frac{C_1 S_T^2}{C_3 \left(S_b^2 - \frac{S_w^2}{M}\right)}}.$$

From this it may be concluded that the value of t will be maximum and the value of m will be minimum.

Chapter 19

Multiphase Sampling

19.1 Introduction

The ratio and regression estimators are based on information of some auxiliary or correlated variable correlated to study variable. The correlated variable may be observed from sample points but this does not help to know the population total or population mean of correlated variable. But to find ratio or regression estimator the population total or population mean of correlated variable is needed.

Let x_1, x_2, \dots, x_N be the values of auxiliary variable x , where $X = \sum_{i=1}^N x_i$ and $\bar{X} = \frac{X}{N}$ are the population total and population mean, respectively. If these values are not available, their estimators \hat{X} or \bar{x} can be found out from a survey which can be performed using smaller survey before conducting the survey using resource available for the survey.

Let there be N units in a population. Consider that n_1 units ($n_1 \leq N$) are randomly selected from N units. These n_1 units are investigated to observe the values of auxiliary variable x . Let these values be x_1, x_2, \dots, x_{n_1} . The estimator of population mean (\bar{X}) and population total (X) are

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i \quad \text{and} \quad \hat{X}_1 = N\bar{x}_1.$$

At this stage a second sample of size n ($n \leq n_1$) is selected from first selected n_1 units to observe the values of the main variable y . Such a sampling when n units are selected at random at second phase from first randomly selected n_1 units is called *double sampling* or *two-phase sampling*.

The sampling method what has been discussed above is first has been proposed by Neyman (1938). If the method of selection is continued at different phases from the first selected units, the sampling is known as multiphase sampling. In this section we shall confine our discussion within double sampling.

Since the sample size n becomes smaller to study the main variable, the estimator may be less efficient. But, if the estimator is obtained using ratio or regression method of estimation, there may be a chance to increase the efficiency of the estimator. In such a case, the double sampling is advantageous.

We have discussed multistage sampling in the previous sections. However, there is difference between multistage sampling and multiphase sampling. For multistage sampling there is no need of population frame. Only frame of last stage sampling units are needed. On the other hand, the frame of population units is needed for two-phase sampling. Since sampling units are selected twice and survey is conducted twice in two-phase sampling, the cost of the survey is increased. Still two-phase sampling is advantageous, if the efficiency of the estimator is increased sufficiently compared to the increased cost.

19.2 Double Sampling for Stratification

Neyman (1938) has proposed double sampling for stratification. Let there be N units in a population and the population units are divided into L strata. But the sizes of strata are not known, even the units which are to be included in h th stratum ($h = 1, 2, \dots, L$) is not known. Therefore, the weight $W_h = N_h/N$ for h th stratum is not known which is needed to estimate the population parameter. The problem can be obviated, if at the first step a random sample of n_1 units are selected. On investigation the selected units can be divided into strata and then the stratum size n_{1h} ($h = 1, 2, \dots, L$) can be found out. This n_{1h} helps us to estimate W_h , where the estimate is $W_{1h} = n_{1h}/n_1$. This W_{1h} is an unbiased estimator of W_h .

At the second step a random sample of size n is to be selected from n_1 units selected at the first step. Let n_h be the size of sampling units selected from h th stratum so that $\sum^L n_h = n$. Let y_{hi} be the value of the study variable recorded from i th unit of h th stratum. The sample mean of the variable from h th stratum is

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}, \quad h = 1, 2, \dots, L.$$

Then the estimate of population mean from double sampling is

$$\bar{y}_{\text{std}} = \sum_{h=1}^L W_{1h} \bar{y}_h.$$

Here the population mean is

$$\bar{Y} = \frac{1}{N} \sum_{h=1}^L \sum_{i=1}^{N_h} y_{hi} = \frac{1}{N} \sum_{h=1}^L N_h \bar{Y}_h = \sum W_h \bar{Y}_h, \quad \text{where } \bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}.$$

The population variance of the observation is

$$S^2 = \frac{1}{N-1} \sum_{h=1}^L \sum_{i=1}^{N_h} (y_{hi} - \bar{Y})^2$$

and the population variance of observations of h th stratum is

$$S_h^2 = \frac{1}{N_h-1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2.$$

Here the problem is to decide the values of n_1 and n_h so that $V(\bar{y}_{\text{std}})$ becomes minimum.

Theorem : In double sampling if first sample of size n_1 is selected randomly and if second sample of size n ($n < n_1$) is a sub-sample of first sample, then \bar{y}_{std} is an unbiased estimator of population mean \bar{Y} . The variance of this estimator is

$$V(\bar{y}_{\text{std}}) = S^2 \left(\frac{N - n_1}{N n_1} \right) + \sum_{h=1}^L \frac{W_h S_h^2}{n_1} \left(\frac{1}{V_h} - 1 \right), \quad \text{where } V_h = \frac{n_h}{n_{1h}}, \quad 0 \leq V_h \leq 1.$$

Proof : It is considered that n_h is the sample size of sub-sample selected from first sample of size n_{1h} . Therefore, $V_h = \frac{n_h}{n_{1h}}, 0 \leq V_h \leq 1$ is to be decided first.

If the sampling procedure is repeated, every time first and second samples are selected. As a result, W_{1h} , n_h and \bar{y}_h become random variables. Since the first sample is a simple random sample of size n_{1h} , $E(W_{1h}) = W_h$. Again, if W_h is considered fixed and if we calculate the mean of all sample means, the $E(\bar{y}_h) = \bar{Y}_h$, since \bar{y}_h is the simple random sample mean from h th stratum. Since first sample is a simple random sample, we can write :

$$E(\bar{y}_{std}) = E_1 \left[E_2 \sum_{h=1}^L W_{1h} \bar{y}_h \right] = E_1 \left[\sum_{h=1}^L W_{1h} \bar{Y}_h \right] = \sum_{h=1}^L W_h \bar{Y}_h = \bar{Y}.$$

Hence, \bar{y}_{std} is an unbiased estimator of \bar{Y} .

To find the variance of \bar{y}_{std} , let us assume that y_{hi} are observed after selecting the first sample of size n_{1h} . Since $W_{1h} = n_{1h}/n_1$, we can write,

$$\sum_{h=1}^L W_{1h} \bar{y}_{1h} = \bar{y}_1.$$

Here \bar{y}_1 is the sample mean of simple random sample of size n_1 . If the sample of size n_1 is selected repeatedly, then

$$V(\bar{y}_1) = V \left[\sum_{h=1}^L W_{1h} \bar{y}_{1h} \right] = \frac{N - n_1}{n_1 N} S^2, \text{ where } S^2 = \frac{1}{N - 1} \sum_{h=1}^L \sum_{i=1}^{N_h} (y_{hi} - \bar{Y})^2.$$

But the double sampling mean \bar{y}_{std} can be written as

$$\bar{y}_{std} = \sum_{h=1}^L W_{1h} \bar{y}_h = \sum_{h=1}^L W_{1h} \bar{y}_{1h} + \sum_{h=1}^L W_{1h} (\bar{y}_h - \bar{y}_{1h}).$$

Again, since \bar{y}_h is the simple random sample mean from first sample, $E_2(\bar{y}_h) = \bar{y}_{1h}$. Here E_2 is used for the expectation in case of second sample. Now,

$$\text{Cov}[\bar{y}_{1h}, (\bar{y}_h - \bar{y}_{1h})] = 0 \text{ and } \text{Cov}[\bar{y}_{1h}, \bar{y}_h] = 0 = V(\bar{y}_{1h}).$$

Again, $V(\bar{y}_h - \bar{y}_{1h}) = V(\bar{y}_h) + V(\bar{y}_{1h})$.

Hence, if W_{1h} is fixed,

$$V_2 \left[\sum W_{1h} (\bar{y}_h - \bar{y}_{1h}) \right] = \sum_{h=1}^L W_{1h}^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{n_h} \right) = \sum_{h=1}^L \frac{W_{1h} S_h^2}{n_1} \left(\frac{1}{V_h} - 1 \right).$$

But the values of W_{1h} are not fixed. Since samples are selected repeatedly, the distribution of W_{1h} can be found out. In that case, if the means of all samples are calculated,

$$V[\bar{y}_{std}] = S^2 \left(\frac{N - n_1}{n_1 N} \right) + \sum_{h=1}^L \frac{W_h S_h^2}{n_1} \left(\frac{1}{V_h} - 1 \right).$$

Corollary : In double sampling for stratification, if first sample is simple random sample and second sample is a simple random sub-sample of first sample, then the variance of \bar{y}_{std} is

$$V(\bar{y}_{std}) \approx \sum_{h=1}^L W_h S_h^2 \left(\frac{1}{n_1 V_h} - \frac{1}{N} \right) + \frac{g_1}{n_1} \sum_{h=1}^L W_h (\bar{Y}_h - \bar{Y}), \text{ where } g_1 = \frac{(N - n_1)}{N - 1}.$$

The above result is obtained by partitioning the total sum of squares as follows. We can write :

$$(N-1)S^2 = \sum_{h=1}^L (N_h - 1)S_h^2 + \sum N_h(\bar{Y}_h - \bar{Y})^2.$$

Multiplying both sides of the above equality by g_1/n_1N , we get

$$\frac{N-n_1}{n_1N}S^2 = S^2 \left(\frac{1}{n_1} - \frac{1}{N} \right) = \frac{g_1}{n_1} \sum \left(W_h - \frac{1}{N} \right) S_h^2 + \frac{g_1}{n_1} \sum W_h(\bar{Y}_h - \bar{Y})^2.$$

Now, putting the value of S^2 in $V(\bar{y}_{std})$, we get

$$V(\bar{y}_{std}) = \sum_{h=1}^L \frac{W_h S_h^2}{n_1} \left(\frac{1}{V_h} - 1 \right) + \frac{g_1}{n_1} \sum_{h=1}^L \left(W_h - \frac{1}{N} \right) S_h^2 + \frac{g_1}{n_1} \sum_{h=1}^L W_h(\bar{Y}_h - \bar{Y})^2.$$

Again, we can write,

$$-\frac{1}{n_1} + \frac{g_1}{n_1} = -\frac{1}{N} + \frac{g_1}{n_1N}.$$

Using this result, we can write,

$$V(\bar{y}_{std}) = \sum_{h=1}^L W_h S_h^2 \left(\frac{1}{n_1 V_h} - \frac{1}{N} \right) + \frac{g_1}{n_1 N} \sum_{h=1}^L (W_h - 1) S_h^2 + \frac{g_1}{n_1} \sum W_h(\bar{Y}_h - \bar{Y})^2.$$

Now, if g_1/n_1N is negligible, we have

$$V(\bar{y}_{std}) \approx \sum_{h=1}^L W_h S_h^2 \left(\frac{1}{n_1 V_h} - \frac{1}{N} \right) + \frac{g_1}{n_1} \sum_{h=1}^L W_h(\bar{Y}_h - \bar{Y})^2.$$

The estimator of this variance is

$$v(\bar{y}_{std}) = \frac{n_1(N-1)}{(n_1-1)N} \left[\sum_{h=1}^L W_{1h} s_h^2 \left(\frac{1}{n_1 V_h} - \frac{1}{N} \right) + \frac{g_1}{n_1} \sum_{h=1}^L s_h^2 \left(\frac{W_{1h}}{N} - \frac{1}{n_1 V_h} \right) + \frac{g_1}{n_1} \sum_{h=1}^L W_{1h} (\bar{y}_h - \bar{y}_{std})^2 \right].$$

According to Rao (1973) this estimator becomes

$$v(\bar{y}_{std}) = \frac{N-1}{N} \sum_{h=1}^L \left(\frac{n_{1h}-1}{n_1-1} - \frac{n_h-1}{N-1} \right) \frac{W_{1h} s_h^2}{n_h} + \frac{N-n_1}{N(n_1-1)} \sum_{h=1}^L W_{1h} (\bar{y}_h - \bar{y}_{std})^2.$$

19.3 Double Sampling for Ratio Estimator

The ratio estimator is defined on the basis of estimator of population ratio $R = \frac{Y}{X}$, where Y is the total of population observations and X is the total of population observations of the auxiliary variable x . If the information of X and \bar{X} is not known, these can be estimated by

selecting a random sample of n_1 observations ($n_1 \leq N$) and investigating the sampling units for the variable x . Let this estimator of X be \bar{x}_1 . The estimator of \bar{X} is

$$\bar{x}_1 = \frac{1}{n_1} \sum x.$$

At the second step, a random sample of n observations ($n \leq n_1$) are selected from the first selected n_1 observations. The information on the study variable y and the auxiliary variable x are observed from the sub-sample of n observations. Let the sample mean of y and x be \bar{y} and \bar{x} , respectively. Then the ratio estimator of population mean \bar{Y} from double sampling is defined by

$$\bar{y}_{rd} = \frac{\bar{y}}{\bar{x}} \bar{x}_1 = \hat{R} \bar{x}_1$$

Theorem : In double sampling if first sample is a simple random sample of size n_1 and if second sample is a sub-sample of first sample selected under SRS scheme, then the ratio estimator \bar{y}_{rd} is a biased estimator of population mean. The relative bias and sampling variance of this estimator are

$$\text{Rel Bias } (\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n_1} \right) (C_x^2 - \rho C_x C_y)$$

$$\text{and } V(\bar{y}_{rd}) = \left(\frac{1}{n_1} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n_1} \right) (S_y^2 + R^2 S_x^2 - 2RS_{yx}),$$

$$\text{where } C_x = \frac{S_x}{\bar{X}}, C_y = \frac{S_y}{\bar{Y}}, S_y^2 = \frac{1}{N-1} \sum (y_i - \bar{Y})^2, S_x^2 = \frac{1}{N-1} \sum (x_i - \bar{X})^2,$$

$$S_{yx} = \frac{1}{N-1} \sum (x_i - \bar{X})(y_i - \bar{Y}).$$

Proof : Let $\bar{y} = \bar{Y}(1 + e)$, $\bar{x} = \bar{X}(1 + e_1)$, $\bar{x}_1 = \bar{X}(1 + e_2)$.

Assume that $E(e) = E(e_1) = E(e_2) = 0$

The ratio estimator is

$$\begin{aligned} \bar{y}_{rd} &= \frac{\bar{y}}{\bar{x}} \bar{x}_1 = \frac{\bar{Y}(1 + e)}{\bar{X}(1 + e_1)} \bar{X}(1 + e_2) = \bar{Y}(1 + e)(1 + e_2)(1 + e_1)^{-1} \\ &= \bar{Y}[1 + (e - e_1 + e_2) + \dots]. \end{aligned}$$

If we ignore the error terms above power 2, we have

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

and \bar{y}_{rd} is unbiased. The bias $(\bar{y}_{rd}) = 0$ and the variance of \bar{y}_{rd} becomes

$$V(\bar{y}_{rd}) \approx \frac{1-f}{n} [S_y^2 + R^2 S_x^2 - 2RS_{yx}].$$

But, if we consider the terms of e , e_1 and e_2 with higher powers, then \bar{y}_{rd} is not an unbiased estimator. We can write,

$$\bar{y}_{rd} = \bar{Y}[1 + (e - e_1 + e_2) + (-ee_1 + ee_2 - e_1e_2 + e_1^2) + \dots].$$

$$\text{Here } E[ee_1] = \frac{\text{Cov}(\bar{y}, \bar{x})}{\bar{Y}\bar{X}} = \frac{\left(\frac{1}{n} - \frac{1}{N}\right) S_{yx}}{\bar{Y}\bar{X}}$$

$$\begin{aligned} E[ee_2] &= \frac{\text{Cov}(\bar{y}, \bar{x}_1)}{\bar{Y}\bar{X}} = \frac{1}{\bar{Y}\bar{X}} [\text{Cov}\{E(\bar{y}/n_1), E(\bar{x}_1/n_1)\} + E\{\text{Cov}(\bar{y}, \bar{x}_1/n_1)\}] \\ &= \frac{1}{\bar{Y}\bar{X}} \text{Cov}(\bar{y}_1, \bar{x}_1) = \frac{1}{\bar{Y}\bar{X}} \left(\frac{1}{n_1} - \frac{1}{N}\right) S_{yx}. \end{aligned}$$

Here \bar{y}_1 is the sample mean of variable y from first sample.

$$E[e_1e_2] = \frac{1}{\bar{X}^2} V(\bar{x}_1) = \frac{\left(\frac{1}{n_1} - \frac{1}{N}\right) S_x^2}{\bar{X}^2} \quad \text{and} \quad E(e_1^2) = V(\bar{x}) = \frac{1}{\bar{X}^2} \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2.$$

Therefore, if e , e_1 and e_2 with powers two are considered, then

$$E[\bar{y}_{rd}] \approx \bar{Y} \left[1 + \left(\frac{1}{n} - \frac{1}{n_1}\right) (C_x^2 - \rho C_x C_y) \right].$$

Hence, the relative bias of \bar{y}_{rd} is

$$\text{Rel bias}(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n_1}\right) (C_x^2 - \rho C_x C_y).$$

Again, if we consider the first powers of e , e_1 and e_2 the variance of \bar{y}_{rd} is

$$\begin{aligned} V(\bar{y}_{rd}) &= E[\bar{y}_{rd} - \bar{Y}]^2 \approx \bar{Y}^2 [e - e_1 + e_2]^2 \\ &= \bar{Y}^2 [E(e^2) + E(e_1^2) + E(e_2^2) - 2E(ee_1) + 2E(ee_2) - 2E(e_1e_2)]. \end{aligned}$$

$$\text{Here } E[e^2] = \frac{1}{\bar{Y}^2} E(\bar{y} - \bar{Y})^2 = \frac{1}{\bar{Y}^2} \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2.$$

$$E[e_1^2] = \frac{1}{\bar{X}^2} E(\bar{x}_1 - \bar{X})^2 = \frac{1}{\bar{X}^2} \left(\frac{1}{n_1} - \frac{1}{N}\right) S_x^2.$$

$$\begin{aligned} V(\bar{y}_{rd}) &= \bar{Y}^2 \left[\left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_y^2}{\bar{Y}^2} + \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_x^2}{\bar{X}^2} + \left(\frac{1}{n_1} - \frac{1}{N}\right) \frac{S_x^2}{\bar{X}^2} - \frac{2}{\bar{Y}\bar{X}} \left(\frac{1}{n} - \frac{1}{N}\right) S_{yx} \right. \\ &\quad \left. + \frac{2}{\bar{Y}\bar{X}} \left(\frac{1}{n_1} - \frac{1}{N}\right) S_{yx} - \frac{2}{\bar{X}^2} \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2 \right] \\ &= \left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n_1}\right) (S_y^2 + R^2 S_x^2 - 2R S_{yx}). \end{aligned}$$

Corollary : In double sampling, if first sample is a simple random sample and if second sample is a simple random sub-sample of first sample, then the unbiased estimator of variance of \bar{y}_{rd} is

$$v(\bar{y}_{rd}) = \left(\frac{1}{n_1} - \frac{1}{N}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n_1}\right) (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{yx}),$$

$$\text{where } s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2, \quad s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2, \quad s_{yx} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}),$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}}.$$

Example 19.1 : To estimate the total mango production in a district the entire district is divided into 4000 units. The units are the areas producing mangoes in large scale. Out of 4000 units 50 units are randomly selected for the survey and during mango season the amount of production of mango and number of mango trees in the selected area are recorded.

Later on a sub-sample of 15 units from the first selected 50 units are selected randomly. The information on the number of mango trees and total mango production in those trees of selected units are given below :

Estimate the total mango production in the district. Also estimate the variance of your estimator.

Sl. No. of Unit	No. of Mango trees x	No. of Mangoes y	Sl. No. of Unit	No. of Mango trees x	No. of Mangoes y
1	30	4048	26	20	4042
2	25	4567	27	30	5618
3	38	5012	28	42	8018
4	46	7018	29	45	9612
5	15	2011	30	19	4011
6	37	4112	31	18	3102
7	27	4445	32	18	2506
8	32	5045	33	25	4002
9	28	4152	34	28	6501
10	25	3047	35	32	4418
11	19	4672	36	42	7018
12	40	7023	37	31	6012
13	29	5660	38	35	7015
14	18	3032	39	40	6008
15	16	6730	40	42	6518
16	25	4912	41	33	6818
17	24	6870	42	27	5012
18	20	6556	43	19	3718
19	42	8011	44	30	6012
20	37	5343	45	33	5802
21	32	4845	46	34	7502
22	29	3912	47	38	8011
23	12	2815	48	42	9008
24	22	5032	49	18	3812
25	20	6872	50	22	4015

The information of sub-sample are as follows :

Sl. No. of Unit	x	y	Sl. No. of Unit	x	y	Sl. No. of Unit	x	y
01	30	4048	11	19	4672	35	32	4418
16	25	4912	44	30	3012	34	28	6501
48	42	9008	14	18	3032	26	20	4042
33	25	4002	22	29	3912	19	42	8011
45	33	5802	10	25	3047	21	32	4845

Solution : We have $n_1 = 0$, $n = 15$, $\bar{x}_1 = 29.02$, $\bar{x} = 28.67$, $\bar{y} = 5084.27$, $\hat{R} = \frac{\bar{y}}{\bar{x}} = 177.34$, $N = 4000$.

The average amount of mangoes produced in a unit is

$$\bar{y}_{rd} = \hat{R}\bar{x}_1 = 177.34 \times 29.02 = 5146.41.$$

Total mango production in the district is

$$\hat{Y}_{rd} = N\bar{Y}_{rd} = 4000 \times 5146.41 = 20585627.$$

The estimate of variance of \hat{Y}_{rd} is

$$v(\hat{Y}_{rd}) = N^2 v(\bar{y}_{rd}) = N^2 \left[\left(\frac{1}{n_1} - \frac{1}{N} \right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n_1} \right) (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}) \right].$$

Here $s_y^2 = 2936653.352$, $s_x^2 = 51.6667$, $s_{yx} = 9780.7381$.

$$\begin{aligned} \therefore v(\hat{Y}_{rd}) &= (4000)^2 \left[\left(\frac{1}{50} - \frac{1}{4000} \right) 2936653.352 + \left(\frac{1}{15} - \frac{1}{50} \right) (2936653.352 \right. \\ &\quad \left. + (177.34)^2 51.6667 - 2 \times 177.34 \times 9780.7381) \right] \\ &= (4000)^2 [57998.9037 + 50983.8825] = (4000)^2 \times 7015.0212. \end{aligned}$$

$$\text{s.e. } (\hat{Y}_{rd}) = \sqrt{v(\hat{Y}_{rd})} = 335022.8935.$$

$$v(\bar{y}_{rd}) = 7015.0212.$$

19.4 Double Sampling for Regression Estimator

The regression estimator also needs the value of population mean of auxiliary variable x . If it is not known, it can be estimated by selecting first a simple random sample of size n_1 from the population of N units and then selecting a sub-sample of size n ($n \leq n_1$) from the first sample at random. Let \bar{X}_1 be the simple random sample mean of auxiliary variable x observed from the first sample. This is an unbiased estimator of \bar{X} . The sample means of y and x from the second sample are \bar{y} and \bar{x} , respectively. Then the linear regression estimator of population mean \bar{Y} is defined by

$$\bar{y}_{1rd} = \bar{y} + b(\bar{x}_1 - \bar{x}),$$

where b is the regression coefficient of y on x observed from second sample.

The second sample from first sample is to be selected in such a way that the size of second sample becomes $n = Vn_1 = n_1/k$. the value of V is a fraction which needs to be decided before selecting the sample.

Theorem : In double sampling, if first sample is selected by SRS scheme and second sample is a sub-sample of the first sample selected at random, then the regression estimator \bar{y}_{1rd} is a biased estimator of population mean \bar{Y} . The bias and variance of \bar{y}_{1rd} are

$$\text{Bias } (\bar{y}_{1rd}) = -B \left(\frac{1}{n} - \frac{1}{n_1} \right) \left[\frac{\mu_{21}}{S_{yx}} - \frac{\mu_{20}}{S_x^2} \right] \quad \text{and} \quad V(\bar{y}_{1rd}) \approx \frac{(1 - \rho^2)}{n} S_y^2 + \frac{\rho^2 S_y^2}{n_1} - \frac{S_y^2}{N}$$

Proof : Given $\bar{y}_{1rd} = \bar{y} + b(\bar{x}_1 - \bar{x}) = \bar{y} - b(\bar{x} - \bar{x}_1)$.

$$E(\bar{y}_{1rd}) = E_1 E_2 (\bar{y}_{1rd} / \bar{x}_1) = E_1 [\bar{y} - E_1(\bar{x} - \bar{x}_1)b] = \bar{Y} - E_1 E_2 \{b(\bar{x} - \bar{x}_1) / \bar{x}_1\}.$$

$$\therefore \text{Bias } (\bar{y}_{1rd}) = -E_1 E_2 \{b(\bar{x} - \bar{x}_1) / \bar{x}_1\}.$$

Let us consider that $\bar{x} = (\bar{X} + e)$, $\bar{x}_1 = \bar{X}(1 + e_1)$, $s_{xy} = S_{xy}(1 + \epsilon_1)$, $s_x^2 = S_x^2(1 + \epsilon_2)$.

Here $E(e) = E(e_1) = E(\epsilon_1) = E(\epsilon_2) = 0$.

$$\begin{aligned} \text{Bias } (\bar{y}_{1rd}) &= -E_1 E_2 \left[B(e - e_1) \left(1 + \frac{\epsilon_1}{S_{xy}} \right) \left(1 + \frac{\epsilon_2}{S_x^2} \right)^{-1} \right] = -E_1 E_2 \{B(e - e_1)(\epsilon_1 - \epsilon_2)\} \\ &= -B \left[\frac{\text{Cov}(s_{xy}, \bar{x}_1) - \text{Cov}(s_{xy}, \bar{x})}{S_x^2} - \frac{\text{Cov}(s_x^2, \bar{x}_1) - \text{Cov}(s_x^2, \bar{x})}{S_x^2} \right] \\ &= -B \left[\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{20}}{S_x^2} \right]. \end{aligned}$$

This is similar to the bias of regression estimator \bar{y}_{1r} . Here $\mu_{21} = E(\bar{x} - \bar{X})(\bar{y} - \bar{Y})$, $\mu_{20} = E(\bar{x} - \bar{X})^2$.

The bias reduces with the increase in the value of n_1 .

To find the variance of \bar{y}_{1rd} , let us use B instead of b , where B is the population regression coefficient of y on x [$B = S_{yx}/S_x^2$]. Due to the use of B , bias of order $1/\sqrt{n}$ will be introduced in the estimator. But, if $1/n$ and $1/n_1$ are neglected, we shall get a value near to $V(\bar{y}_{1rd})$. Now, let us consider

$$\tilde{y}_{1rd} = \bar{y} + B(\bar{x}_1 - \bar{x}) \quad \text{and} \quad u_i = y_i - Bx_i.$$

If the first sample is bigger compared to the second sample, the first sample can be considered a finite population and the second sample is a simple random sample from finite population. Hence,

$$E_2(\tilde{y}_{1rd}) = \bar{Y}_1 \quad \text{and} \quad V_2(\tilde{y}_{1rd}) = \left(\frac{1}{n} - \frac{1}{n_1} \right) S_{1u}^2.$$

Here \bar{Y}_1 is the mean based on the first sample. Then

$$V(\bar{y}_{1rd}) \approx V(\tilde{y}_{1rd}) = V_1(\bar{Y}_1) + E_1 \left(\frac{1}{n} - \frac{1}{n_1} \right) S_{1u}^2.$$

Here S_{1u}^2 is the variance of u_i based on first sample. Therefore,

$$V(\bar{y}_{1rd}) \approx \left(\frac{1}{n_1} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n_1} \right) S_y^2(1 - \rho^2).$$

Here S_{1u}^2 is an unbiased estimator of $S_u^2 = S_y^2(1 - \rho^2)$.

$$\therefore V(\bar{y}_{1rd}) \approx \frac{S_y^2(1 - \rho^2)}{n} + \frac{\rho^2 S_y^2}{n_1} - \frac{S_y^2}{N}$$

The unbiased estimator of this variance is given by

$$v(\bar{y}_{1rd}) = \frac{s_{y \cdot x}^2}{n} + \frac{s_y^2 - s_{y \cdot x}^2}{n_1} - \frac{s_y^2}{N}$$

where $s_{y \cdot x}^2 = \frac{1}{n-2} \left[\sum (y_i - \bar{y})^2 - b^2 \sum (x_i - \bar{x})^2 \right]$ is an unbiased estimator of $S_y^2(1 - \rho^2)$ and $s_y^2 - s_{y \cdot x}^2$ is an unbiased estimator of $\rho^2 S_y^2$.

If the second sample is small in size, then

$$v(\bar{y}_{1rd}) = s_{y \cdot x}^2 \left[\frac{1}{n} + \frac{(\bar{x}_1 - \bar{x})^2}{\sum (x - \bar{x})^2} \right] + \frac{s_y^2 - s_{y \cdot x}^2}{n_1} - \frac{s_y^2}{N}$$

Example 19.2 : Using data of Example 19.1 estimate total mango production in the district and also estimate the standard error of your estimator. Use regression method of estimation.

Solution : We have $n_1 = 50$, $n = 15$, $N = 4000$, $\bar{y} = 5084.27$, $\bar{x} = 28.67$, $\bar{x}_1 = 29.02$, $b = \frac{s_{yx}}{s_x^2} = \frac{9780.7381}{51.6667} = 189.30$, $s_y^2 = 2936653.352$.

$$\begin{aligned} s_{y \cdot x}^2 &= \frac{1}{n-2} \left[\sum (y - \bar{y})^2 - b^2 \sum (x - \bar{x})^2 \right] \\ &= \frac{1}{15-2} [41113146.93 - (189.30)^2 723.333] = 1168682.906. \end{aligned}$$

The regression estimator of mean mango production per study unit is

$$\bar{y}_{1rd} = \bar{y} - b(\bar{x} - \bar{x}_1) = 5084.27 - 189.30(28.67 - 29.02) = 5150.525.$$

Estimate of total mango production in the study area is

$$\hat{Y}_{1rd} = N\bar{y}_{1rd} = 4000 \times 5150.525 = 20602100.$$

The estimate of variance of \hat{Y}_{1rd} is

$$\begin{aligned} v(\hat{Y}_{1rd}) &= N^2 v(\bar{y}_{1rd}), \\ \text{where } v(\hat{y}_{1rd}) &= \frac{s_{y \cdot x}^2}{n} + \frac{s_y^2 - s_{y \cdot x}^2}{n_1} - \frac{s_y^2}{N} \\ &= \frac{1168682.906}{15} + \frac{2936653.352 - 1168682.906}{50} - \frac{2936653.352}{4000} = 112537.4393. \end{aligned}$$

$$\therefore v(\hat{Y}_{1rd}) = (4000)^2 112537.4393.$$

$$\text{s.e.} (\hat{Y}_{1rd}) = \sqrt{v(\hat{Y}_{1rd})} = 1341864.013.$$

19.5 Optimum Allocation in Double Sampling for Ratio Estimator

Let there be N units in a population. Assume that N is large. The problem is to decide about the values of n_1 and n , where n_1 is the sample size of the first sample and n is the sample size of the sub-sample. Since N is large,

$$V(\bar{y}_{rd}) = \frac{S_y^2}{n_1} + \left(\frac{1}{n} - \frac{1}{n_1}\right) (S_y^2 + R^2 S_x^2 - 2RS_{yx}).$$

Consider that the cost function to conduct the survey is

$$C = n_1 C_1 + n C.$$

The problem of optimum allocation is to find the value of n_1 and n on the basis of cost function mentioned above so that $V(\bar{y}_{rd})$ is minimum.

$$\text{Let } V = \frac{1}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2.$$

$$\text{Then } V(\bar{y}_{rd}) = \frac{V}{n} + \frac{1}{n_1} (S_y^2 - V).$$

Let C_0 be the fixed cost for the survey. We need to find the value of n and n_1 so that $\phi = V(\bar{y}_{rd}) + \lambda(C - C_0)$ becomes minimum. The value of n_1 and n are to be found out from the solution of the equations $\frac{\partial \phi}{\partial n_1} = 0$ and $\frac{\partial \phi}{\partial n} = 0$.

Solving these two equations, we get

$$\frac{n\sqrt{C}}{\sqrt{V}} = \frac{n_1\sqrt{C_1}}{\sqrt{S_y^2 - V}} = \frac{C_0}{\sqrt{CV} + \sqrt{C_1(S_y^2 - V)}}.$$

At this stage the minimum value of $V(\bar{y}_{rd})$ is given by

$$V(\bar{y}_{rd})_{\min} = \frac{1}{C_0} [\sqrt{CV} + \sqrt{C_1(S_y^2 - V)}]^2.$$

19.6 Optimum Allocation in Double Sampling for Regression Estimator

The variance of the regression estimator of population mean is

$$V(\bar{y}_{1rd}) \approx \frac{S_y^2(1 - \rho^2)}{n} + \frac{\rho^2 S_y^2}{n_1} - \frac{S_y^2}{N}.$$

Let this variance be V . Consider a cost function

$$C = C_1 n_1 + cn$$

for the survey work. Then we can write,

$$V + \frac{S_y^2}{N} = \frac{S_y^2(1 - \rho^2)}{n} + \frac{\rho^2 S_y^2}{n_1}.$$

Let us put the values of n and n_1 from cost function in variance formula. Using Cauchy-Schwarz inequality and minimizing VC , we get

$$\frac{Cn^2}{S_y^2(1 - \rho^2)} = \frac{C_1 n_1^2}{\rho^2 S_y^2} \quad \text{or,} \quad \frac{n}{n_1} = \left[\frac{C_1}{C} \cdot \frac{1 - \rho^2}{\rho^2} \right]^{\frac{1}{2}}.$$

The minimum variance of \bar{y}_{1rd} is given by

$$V_{\min} = \frac{S_y^2}{C} [\sqrt{C(1-\rho^2)} + \sqrt{C_1\rho^2}]^2 - \frac{S_y^2}{N}.$$

19.7 Double Sampling for Difference Estimator

The difference estimator is defined by

$$\hat{Y}_D = (\bar{y} - \beta\bar{x}) + \beta\bar{X},$$

where $\bar{y} = \frac{1}{n} \sum y$, $\bar{x} = \frac{1}{n} \sum x$, $\bar{X} = \frac{1}{N} \sum x$.

This estimator is defined on the assumption that the study variable y and the auxiliary variable x are linearly related and the regression line of y on x is $y = \alpha + \beta x + e$ with assumption $E(e) = 0$. Then $E(y) = \alpha + \beta x$. The difference estimator is based on the difference of $y - \beta x$.

The difference estimator defined above needs population mean \bar{X} of auxiliary variable x . If it is not known, its estimate can be found out from a preliminary survey. Let n_1 units of N units be selected under SRS scheme to observe the values of x variables. Then a simple random sample of n observations are selected from the first sample of n_1 observations. Let \bar{x}_1 be the sample mean of x variable from the first sample. The sample means of variables y and x from second sample are \bar{y} and \bar{x} , respectively. Then the difference estimator of population mean from double sampling is

$$\bar{y}_{dd} = \bar{y} + \beta(\bar{x}_1 - \bar{x}), \quad \text{where } \beta = \frac{Y}{X}.$$

Theorem : In double sampling, if the first sample is selected under SRS scheme and the second sample is a sub-sample under SRS scheme from first sample, then the difference estimator \bar{y}_{dd} is an unbiased estimator of population mean \bar{Y} . The variance of this estimator is

$$V(\bar{y}_{dd}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n_1}\right) (\beta^2 S_x^2 - 2\beta\rho S_y S_x).$$

Proof : We have $\bar{y}_{dd} = \bar{y} + \beta(\bar{x}_1 - \bar{x})$.

Let us consider that the first sample is of large size and the sample mean \bar{y}_1 is the population mean. Then

$$E_2(\bar{y}_{dd}/\bar{x}_1) = E_2\{[\bar{y} - \beta(\bar{x}_1 - \bar{x})]/\bar{x}_1\} = \bar{y}_1.$$

Hence, $E_1(\bar{y}_1) = \bar{Y}$ and $E(\bar{y}_{dd}) = \bar{Y}$.

$$V(\bar{y}_{dd}) = V_1 E_2(\bar{y}_{dd}/\bar{y}_1) + E_1 V_2(\bar{y}_{dd}/\bar{y}_1).$$

But $V_1 E_2(\bar{y}_{dd}/\bar{y}_1) = \left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2$, where $S_y^2 = \frac{1}{N-1} \sum (y_i - \bar{Y})^2$.

$$\begin{aligned} E_1 V_2(\bar{y}_{dd}/\bar{y}_1) &= E_1 \left(\frac{1}{n} - \frac{1}{n_1}\right) \frac{1}{n_1 - 1} \sum^{n_1} (y_i - \beta x_i - \bar{y}_1 + \beta \bar{x}_1)^2 \\ &= \left(\frac{1}{n} - \frac{1}{n_1}\right) \sum_{i=1}^N \frac{1}{N-1} (y_i - \beta x_i - \bar{Y} - \beta \bar{X})^2 \\ &= \left(\frac{1}{n} - \frac{1}{n_1}\right) [S_y^2 + \beta^2 S_x^2 - 2\beta\rho S_y S_x]. \end{aligned}$$

$$\begin{aligned} \therefore V(\bar{y}_{dd}) &= \left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n_1}\right) (S_y^2 + \beta^2 S_x^2 - 2\beta\rho S_y S_x) \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n_1}\right) (\beta^2 S_x^2 - 2\beta\rho S_y S_x). \end{aligned}$$

Corollary : In double sampling, the estimator of variance of difference estimator \bar{y}_{dd} is

$$v(\bar{y}_{dd}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n_1}\right) s_d^2, \quad \text{where } s_d^2 = \frac{1}{n-1} \sum \{(y_i - \bar{y}) - \beta(x_i - \bar{x})\}^2.$$

Corollary : In double sampling, if the second sample is independent of the first sample, then the difference estimator \bar{y}_{dd} is an unbiased estimator of \bar{Y} . The sampling variance of this estimator is

$$V(\bar{y}_{dd}) = \left(\frac{1}{n} - \frac{1}{N}\right) (S_y^2 + \beta^2 S_x^2 - 2\beta\rho S_y S_x) + \beta^2 \left(\frac{1}{n_1} - \frac{1}{N}\right) S_x^2.$$

The unbiased estimator of this variance is

$$v(\bar{y}_{dd}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_d^2 + \beta^2 \left(\frac{1}{n_1} - \frac{1}{N}\right) s_{1x}^2, \quad \text{where } s_{1x}^2 = \frac{1}{n_1-1} \sum (x_i - \bar{x}_1)^2.$$

For this independent sample at the second step it is assumed that the first sample of size n_1 is selected by an organisation and the second sample of size n is selected by another organisation and the values of y and x are observed by the second organisation. Thus, the first and second samples are independent. The values of x are observed from the first sample also. Since two samples are independent,

$$E(\bar{y} - \beta\bar{x}) = \bar{Y} - \beta\bar{X} \quad \text{and} \quad E(\beta\bar{x}_1) = \beta\bar{X}, \quad \therefore E(\bar{y}_{dd}) = E[\bar{y} + \beta(\bar{x}_1 - \bar{x})] = \bar{Y}.$$

Similarly, we can write,

$$V(\bar{y}_{dd}) = V(\bar{y} - \beta\bar{x}) + V(\beta\bar{x}_1) = \left(\frac{1}{n} - \frac{1}{N}\right) (S_y^2 + \beta^2 S_x^2 - 2\beta\rho S_y S_x) + \beta^2 \left(\frac{1}{n_1} - \frac{1}{N}\right) S_x^2.$$

Corollary : If the cost of double sampling is spent directly for simple random sampling, then the variance of the sample mean is

$$V(\bar{y})_{\text{ran}} = \left(\frac{1}{n_0} - \frac{1}{N}\right) S_y^2, \quad \text{where } n_0 = n + \frac{n_1 C_1}{C}.$$

Here C_1 and C are the costs to collect the information on variables y and x . Then the total cost for the survey is

$$C = C_1 n_1 + cn.$$

At this stage, if double sample is not used and the values of y variables are observed directly, then the sample size should be

$$n_0 = \frac{C_1 n_1 + Cn}{C} = n + \frac{n_1 C_1}{C}.$$

Now, if $\beta S_x / S_y = h$ is considered, then it can be shown that when

$$2\rho - h > \left[h \left(1 - \frac{n}{n_1}\right) \left(1 + \frac{nC}{n_1 C_1}\right) \right]^{-1}$$

the double sampling is more precise than direct simple random sampling.

Chapter 20

Sampling with Varying Probabilities

20.1 Introduction

In simple random sampling every unit is selected with equal probability. This sampling is suitable and preferable if frame is available and if all the units in the population have the same weight. In practice, the different units in the population may have different weights. For example, one may need to estimate the total production of wheat in an area. In the area there are many villages. If each village is considered as a sampling unit, different villages may have different sizes in respect of total land area cultivated for wheat. In that case, if simple random sample is selected with equal probability of selection of every village and if total production of wheat is recorded from each selected village, then the weight of a village producing more wheat from more cultivated land is not considered. As a result, the estimate of total production of wheat may not be efficient. For efficient estimate of total production, the land area cultivated in the village should be considered and village should be selected proportionately to the total number of plots cultivated for wheat. Because, total production of wheat is correlated with total number of plots or total area used for production. In such a situation, if villages are selected with probability proportional to the number of plots cultivated for wheat or probability proportional to total land area cultivated for wheat, then the sampling is called sampling with varying probabilities, where sampling unit is selected with probability proportional to size of (PPS) sampling unit.

The PPS sample can be selected in two ways, viz., (a) PPS sampling with replacement and (b) PPS sampling without replacement. However, if sample is selected with replacement of preceding unit selected from the population, then every unit is selected with equal probability. In this chapter the estimation procedure of both sampling schemes will be discussed.

20.2 Method of Selection of Sampling Units with Varying Probabilities

Let us consider that the frame is available and the population units are identified by serial numbers 1 to N , and n units are selected at random using 'Random Number Table'. But for PPS sampling the units and their sizes or any other characteristic related to the sampling units are also recorded and frame is formed using serial number for the units according to variable under study and the related variable. For example, in estimating total maize production in a locality if village is used as a sampling unit, the list of villages along with land area cultivated or number of plots used for production in each village are to be recorded. Later on the cumulative land area or plot numbers is calculated. However, the sample can be selected without the cumulative value of unit sizes. Thus, PPS sampling can be done in two ways. These are :

- (a) Method of Cumulative Total (b) Lahiri's Method.

(a) Method of cumulative total : Let there be N units in a population. Consider that the size of i -th unit ($i = 1, 2, \dots, N$) is x_i . With first unit the number from 1 to x_1 is to be

assigned. The value from $x_1 + 1$ to $x_1 + x_2$ is to be assigned with 2nd unit, and so on. Here

$$S_N = \sum_{i=1}^N x_i.$$

Now, using 'Random Number Table' a number from 1 to S_N is to be selected. The selected random number is associated with some of the sampling unit in the population. This unit is included in the sample. If every unit is selected in a similar way, the sampling is known as PPS sampling, where the probability of selection of i -th unit is proportional to $x_i (i = 1, 2, \dots, N)$. The process of selection is repeated n -times for a sample of size n and every time the sample is selected by replacing the preceding selected unit in the population. The process of selection is known as PPS sampling with replacement.

Example 20.1 : In a district there are 25 administrative units. The number of families in each unit is recorded and given below :

Sl. No. of administrative unit	No. of families in the unit	C.f. of families
01	200	200
02	150	350
03	180	530
04	50	580
05	200	780
06	300	1080
07	500	1580
08	450	2030
09	700	2730
10	670	3400
11	800	4200
12	500	4700
13	600	5300
14	400	5700
15	355	6055
16	482	6537
17	263	6800
18	350	7150
19	250	7400
20	420	7820
21	500	8320
22	200	8520
23	180	8700
24	230	8930
25	470	9400

Select a PPS sample of 5 administrative units.

Solution : To select the sample the cumulative number of families are calculated. The total number of families in the population is 9400. It is a number of 4 digits. To select the first unit we need to select a random number of 4 digits. One such number is 2315. This number is

related to population unit 9. So, 9th administrative unit is included in the sample. The process is repeated for 4 other units needed in the sample. Below are given the other random numbers selected, the administrative units selected and the number of families of the selected units.

Sl. No. of sample	Random number selected	Administrative unit selected	Number of families in the selected units
1	2315	09	700
2	0554	04	50
3	1487	07	500
4	3897	11	800
5	1174	07	500

The above method of selection is not advantageous if population size is large. In that case, the calculations of cumulative size of unit is time consuming. As an alternative to this method, Lahiri (1951) suggested another method. Let us discuss the Lahiri's method.

(b) Lahiri's method : In this method a random number from 1 to N is selected first. Later on another random number is selected from 1 to the largest size of the units. If the size of i -th unit is large, say M , then a random number from 1 to M is selected. So, we get a pair of random numbers. If this pair coincides with the unit and size of unit, then this unit is included in the sample, otherwise new pair of random numbers is selected. Let us consider that the first 2-digit number [as N is of two digits(25), in case of Example 20.1] is 23 and the second number is 231 [as the size of population unit is 800, which is largest and is of 3 digits] which is a 3-digit random number. So, we have a pair of random numbers (23, 231). But the size of population unit 23 is 180. So, this unit 23 is not included in the sample. The process of selection of pair of random number is continued until a sample of size n is selected. For the given example the selected pair of random numbers are (14, 148), (11, 117), (07, 070) (05, 055) and (09, 092). The sample of $n = 5$ units are shown below.

Sl. No. of sample	Selected limit	Family size
1	14	400
2	11	800
3	07	500
4	09	700
5	05	200

20.3 Method of Estimation in PPS Sampling with Replacement

Let there be N units in a population. We need a sample of size n . Sample is to be selected with PPS sampling scheme with replacement. Let $y_i (i = 1, 2, \dots, N)$ be the value of characteristic to be studied of i -th unit.

Let us consider that the i -th unit is selected with probability

$$p_i = \frac{x_i}{S_N}, \quad \text{where} \quad \sum_{i=1}^N p_i = 1.$$

Consider that n units are selected independently and hence, $y_i, p_i (i = 1, 2, \dots, n)$ are independently and uniformly distributed. Let

$$z_i = \frac{y_i}{p_i} \quad (i = 1, 2, \dots, n).$$

Here z_i values are also independently distributed. The sample mean of z_i is

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}$$

Again, consider that

$$z_{1i} = \frac{y_i}{Np_i} \quad (i = 1, 2, \dots, N) \quad \text{and} \quad \bar{z}_1 = \frac{1}{n} \sum_{i=1}^n z_{1i} = \frac{1}{Nn} \sum_{i=1}^n \frac{y_i}{p_i}$$

Theorem : In PPS sampling with replacement the sample mean \bar{z}_1 is an unbiased estimator of population mean \bar{Y} and the variance of \bar{z}_1 is

$$V(\bar{z}_1) = \frac{1}{n} \sum_{i=1}^N p_i (z_{1i} - \bar{Y})^2 = \frac{1}{n} \sum p_i (z_{1i} - \bar{z}_1)^2.$$

Proof : Given $\bar{z}_1 = \frac{1}{n} \sum_{i=1}^n z_{1i}$, $z_{1i} = \frac{y_i}{Np_i}$

$$\begin{aligned} E(\bar{z}_1) &= \frac{1}{n} \sum E(z_{1i}) = \frac{1}{n} \sum_{i=1}^n \left[\sum_{i=1}^N p_i \frac{y_i}{Np_i} \right] \\ &= \frac{1}{n} \sum_{i=1}^n \bar{Y} = \bar{Y} = \bar{Z}_1, \text{ where } \bar{Z}_1 = \frac{1}{N} \sum_{i=1}^N z_{1i}. \end{aligned}$$

$$\begin{aligned} V(\bar{z}_1) &= E(\bar{z}_1^2) - [E(\bar{z}_1)]^2 \\ &= E \left[\frac{1}{n} \sum_{i=1}^n z_{1i} \right]^2 - \bar{Z}_1^2 \quad (\because E(\bar{z}_1) = \bar{Z}_1) \\ &= \frac{1}{n^2} \left[E \sum_{i=1}^n z_{1i}^2 + E \sum_{i \neq j} z_{1i} z_{1j} \right] - \bar{Z}_1^2. \end{aligned}$$

Again, $E(z_{1i}^2) = \sum_{i=1}^N p_i z_{1i}^2$ and $E(z_{1i} z_{1j}) = E(z_{1i}) E(z_{1j}) = \bar{Z}_1$.

Therefore, $V(\bar{z}_1) = \frac{1}{n} E(z_{1i}^2) + \frac{n-1}{n} E(z_{1i}) E(z_{1j}) - \bar{Z}_1^2 = \frac{1}{n} \left[\sum_{i=1}^N p_i z_{1i}^2 - \bar{Z}_1^2 \right]$.

$$\sigma^2 = \sum_{i=1}^N p_i (z_{1i} - \bar{Z}_1)^2 = \frac{1}{n} \sum_{i=1}^N p_i (z_{1i} - \bar{Z}_1)^2 = \frac{\sigma^2}{n}.$$

Corollary : The estimator of variance of sample mean under PPS sampling with replacement is

$$v(\bar{z}_1) = \frac{s^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{Np_i} - \frac{\hat{Y}}{N} \right)^2, \text{ where } \hat{Y} = N\bar{z}_1 = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{Np_i}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i}{Np_i} - \frac{\hat{Y}}{N} \right)^2.$$

Proof : Let $s^2 = \frac{1}{n-1} \sum (z_{1i} - \bar{z}_1)^2 = \frac{1}{n-1} \left[\sum_{i=1}^n z_{1i}^2 - n\bar{z}_1^2 \right]$

$$E(s^2) = \frac{1}{n-1} \left[\sum E(z_{1i}^2) - nE(\bar{z}_1^2) \right].$$

We know $V(\bar{z}_1) = E(\bar{z}_1^2) - [E(\bar{z}_1)]^2$

$$E(\bar{z}_1^2) = V(\bar{z}_1) + [E(\bar{z}_1)]^2 = \frac{\sigma^2}{n} + \bar{Z}_1^2.$$

$$\begin{aligned} \therefore E(s^2) &= \frac{1}{n-1} \left[\sum_{i=1}^n \left\{ \sum_{i=1}^N p_i z_{1i}^2 - n \left\{ \frac{\sigma^2}{n} + \bar{Z}_1^2 \right\} \right\} \right] \\ &= \frac{1}{n-1} \left[n \left(\sum_{i=1}^N p_i z_{1i}^2 - \bar{Z}_1^2 \right) - \sigma^2 \right] \\ &= \sigma^2, \quad \therefore \sigma^2 = \sum_{i=1}^N p_i z_{1i}^2 - \bar{Z}_1^2. \end{aligned}$$

$\therefore v(\bar{z}_1)$ is an unbiased estimator of $V(\bar{z}_1)$ and

$$v(\bar{z}_1) = \frac{1}{n(n-1)N^2} \left[\sum_{i=1}^n \left(\frac{y_i}{p_i} \right)^2 - n\bar{Y}^2 \right].$$

Theorem : In PPS sampling with replacement the unbiased estimator of population total Y is

$$\bar{z} = \frac{1}{n} \sum z_i = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} \quad \text{and} \quad v(\bar{z}) = \frac{1}{n} \sum_{i=1}^N p_i \left(\frac{y_i}{p_i} - Y \right)^2,$$

Proof : Let $z_i = \frac{y_i}{p_i}$ ($i = 1, 2, \dots, N$); $E(z_i) = \sum_{i=1}^N p_i \frac{y_i}{p_i} = Y$.

$$\therefore \bar{z} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}.$$

$$E(\bar{z}) = \frac{1}{n} \sum_{i=1}^n E \left(\frac{y_i}{p_i} \right) = \frac{1}{n} \sum_{i=1}^n \left[\sum_{i=1}^N p_i \frac{y_i}{p_i} \right] = Y.$$

Again, $V(z) = \sum_{i=1}^N p_i (z_i - Y)^2 = \sum_{i=1}^N p_i \left(\frac{y_i}{p_i} - Y \right)^2$.

$$\begin{aligned} \therefore V(\bar{z}) &= V \left[\frac{1}{n} \sum z_i \right] = \frac{1}{n^2} V \left[\sum_{i=1}^n z_i \right] = \frac{1}{n^2} \left[\sum V(z_i) + \sum_{i \neq j} \text{Cov}(z_i z_j) \right] \\ &= \frac{1}{n} V(z_i) = \frac{1}{n} \sum_{i=1}^N p_i \left(\frac{y_i}{p_i} - Y \right)^2. \end{aligned}$$

Corollary : In PPS sampling with replacement the estimated variance of the estimator of population total is

$$v(\bar{z}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^n \left(\frac{y_i}{p_i} \right)^2 - n\hat{Y}^2 \right].$$

Proof : We know, $\bar{z} = N\bar{z}_1$.

$$\begin{aligned} V(\bar{z}) &= N^2 V(\bar{z}_1) = \frac{N^2}{n(n-1)N^2} \left[\sum_{i=1}^n \left(\frac{y_i}{p_i} \right)^2 - n\hat{Y}^2 \right] \\ &= \frac{1}{n(n-1)} \left[\sum_{i=1}^n \left(\frac{y_i}{p_i} \right)^2 - n\hat{Y}^2 \right]. \end{aligned}$$

Example 20.2 : Estimate average number of families and total of families along with their estimated standard error using the data of example 20.1. Also, find 95% confidence limits of estimated total families.

Solution : The sample observations and the corresponding probability of selection of sample units (p_i) using data of Example 20.1 are as follows :

$y_i :$	700	50	500	800	500
$p_i :$	0.0745	0.0055	0.0532	0.0851	0.0532

The estimate of average family per administrative unit is

$$\bar{z}_1 = \frac{1}{Nn} \sum_{i=1}^n \frac{y_i}{p_i} = \frac{47027.63295}{25 \times 5} = 376.$$

The estimate of total families in the district is

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} = N\bar{z}_1 = 25 \times 376 = 9400 = \hat{Y}.$$

The estimated variance of \bar{z}_1 is

$$\begin{aligned} v(\bar{z}_1) &= \frac{1}{n(n-1)N^2} \left[\sum_{i=1}^n \left(\frac{y_i}{p_i} \right)^2 - n\hat{Y}^2 \right] \\ &= \frac{1}{5(5-1)(25)^2} [442320674.2 - 5 \times (9400)^2] = 41.6539. \end{aligned}$$

$$\text{s.e.}(\bar{z}_1) = \sqrt{v(\bar{z}_1)} = 6.45.$$

Again, $v(\bar{z}) = N^2 v(\bar{z}_1) = (25)^2 41.6539 = 26033.6875$.

$$\text{s.e.}(\bar{z}) = 161.35.$$

95% confidence interval of estimated population total is

$$\hat{Y}_L = \hat{Y} - t_{0.05,4} \text{s.e.}(\hat{Y}) = 9400 - 2.776 \times 161.35 = 8952.09.$$

$$\hat{Y}_U = \hat{Y} + t_{0.05,4} \text{s.e.}(\hat{Y}) = 9400 + 2.776 \times 161.35 = 9847.90.$$

Corollary : In PPS sampling with replacement if $p_i = p = \frac{1}{N}$, then the sampling is equivalent to simple random sampling.

In PPS sampling with replacement the estimator of population mean is

$$\bar{z}_1 = \frac{1}{Nn} \sum \frac{y_i}{p_i} = \frac{1}{Nn} \sum \frac{y_i}{\frac{1}{N}} = \frac{1}{n} \sum y_i \quad \left(\because p_i = p = \frac{1}{N} \right).$$

Here $\bar{z}_1 = \bar{y}$, where $\bar{y} = \frac{1}{n} \sum y_i$ is the estimator of \bar{Y} in simple random sampling.

$$\begin{aligned} \text{Again, } V(\bar{z}_1) &= \frac{1}{n} \sum_{i=1}^N p_i (z_{1i} - \bar{z}_1)^2 = \frac{1}{n} \sum_{i=1}^N \frac{1}{N} \left(\frac{y_i}{N/N} - \frac{1}{N} \sum \frac{y_i}{N/N} \right)^2 \\ &= \frac{1}{Nn} \sum_{i=1}^n (y_i - \bar{Y})^2 = \frac{\sigma^2}{n}, \quad \text{where } \sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{Y})^2. \end{aligned}$$

Example 20.3 : In an union there are 15 villages. The total area (x_i acre) and total water land area (y_i acre) are given. Select a sample of 5 villages by PPS sampling scheme with replacement and estimate total water land area in the union. Also find 95% confidence interval for the total land area under water.

Sl. No. of villages	X_i	Y_i
01	2200	325
02	2550	450
03	2800	228
04	1600	100
05	3310	750
06	2150	280
07	1855	155
08	2604	280
09	3000	400
10	2560	560
11	2400	150
12	2950	180
13	1752	80
14	2880	160
15	2995	162

Solution : Following the method of selection of Lahiri and using random numbers we have pairs of random numbers as follows :

(05,0554), (14,1487), (11, 1174), (07, 0709), (09, 0924).

Therefore, the selected sample is :

Sl. No. of sample	Sl. No. of villages	y_i	$p_i = \frac{x_i}{S_N}$	y_i/p_i
1	05	750	0.08802	8520.7907
2	14	160	0.07658	2089.3183
3	11	150	0.06382	2350.3604
4	07	155	0.04933	3142.1042
5	09	400	0.07977	5014.4164
Total		1615		21116.99

Here S_N = total area of the villages = 37606. The estimate of water land of the villages is

$$\bar{z} = \frac{1}{n} \sum \frac{y_i}{p_i} = \frac{S_N}{n} \sum_{i=1}^n \frac{y_i}{x_i} = \frac{1}{5} \times 21116.99 = 4223.40 = \hat{Y}$$

$$\begin{aligned} v(\bar{z}) &= \frac{1}{n(n-1)} \left[\sum \left(\frac{y_i}{p_i} \right)^2 - n\hat{Y}^2 \right] = \frac{1}{n(n-1)} \left[S_N^2 \sum \left(\frac{y_i}{x_i} \right)^2 - n\hat{Y}^2 \right] \\ &= \frac{1}{5(5-1)} [117506811.6 - 5 \times (4223.40)^2] \\ &= 1416063.69. \end{aligned}$$

$$\text{s.e.}(\bar{z}) = \sqrt{1416063.69} = 1189.98.$$

Therefore, 95% confidence interval of total water land area of the union is

$$\hat{Y}_L = \hat{Y} - t_{0.05,4} \text{s.e.}(\bar{z}) = 4223.40 - 2.776 \times 1189.98 = 920.01.$$

$$\hat{Y}_U = \hat{Y} + t_{0.05,4} \text{s.e.}(\bar{z}) = 4223.40 + 2.776 \times 1189.98 = 7526.78.$$

If simple random sample is considered, then the estimated total water land area is

$$\begin{aligned} \hat{Y}_{\text{ran}} &= N\bar{y}. & \bar{y} &= \frac{1}{n} \sum y_i = \frac{1615}{5} = 323 \\ &= 15 \times 323 = 4845. \end{aligned}$$

$$\begin{aligned} v(\hat{Y})_{\text{ran}} &= \frac{N(N-n)}{n} s^2, & s^2 &= \frac{1}{n-1} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right] \\ &= \frac{15(15-5)}{5} \times 68245, & &= \frac{1}{5} \left[794625 - \frac{(1615)^2}{5} \right] = 68245 \\ &= 2047350. \end{aligned}$$

$$\therefore \text{s.e.}(\hat{Y}) = \sqrt{2047350} = 1430.86.$$

It is seen that the estimated variance of the estimate of population total is more in case of simple random sample than that of PPS sample with replacement. The relative efficiency of PPS sampling with replacement compared to simple random sampling is

$$\text{RE} = \frac{(\hat{Y})_{\text{ran}}}{v(\bar{z})} = \frac{2047350}{1416063.69} = 144.58\%.$$

The gain in precision of PPS sampling with replacement compared to simple random sampling is :

$$\text{Gain in precision} = \frac{v(\hat{Y})_{\text{ran}} - v(\bar{z})}{v(\bar{z})} = \frac{2047350 - 1416063.69}{1416063.69} = 0.4458.$$

20.4 PPS Sampling without Replacement

The selection of sample under PPS sampling scheme without replacement can be done by three methods. These are :

- (a) General Selection Procedure
- (b) Narain's Scheme of Sample Selection
- (c) Sen-Midzuno Method.

(a) **General selection procedure** : In this method the first unit is selected according to the sampling scheme discussed in section 20.2. The second unit is selected assuming that the first selected unit is not present in the population. The third unit is selected assuming that the first and second selected units are not in the population. The procedure is continued until a sample of n observations is selected. Thus, the probability of selection of different units are different.

Let us consider that in a population there are N units. The size of i -th unit ($i = 1, 2, \dots, N$) is x_i and total size of population units are

$$S_N = \sum_{i=1}^N x_i.$$

Now, the first unit is selected under PPS scheme with probability

$$p_i = \frac{x_i}{S_N}, \quad i = 1, 2, \dots, N,$$

if the i -th unit is selected first. As the i -th unit is assumed to be absent in the population before the selection of j -th unit, the probability of selection of second unit under the condition that at first i -th unit is selected is

$$p_{j/i} = \frac{p_j}{1 - p_i}.$$

It is assumed that j -th unit ($j \neq i = 1, 2, \dots, N$) is included in the sample as second unit. This $p_{j/i}$ indicates that the probabilities of selection of the first and second are not equal. Assume that the characteristic of selected units under study are y_1, y_2, \dots, y_n , where p_i ($i = 1, 2, \dots, n$) is the probability of selection of y_i . Let us discuss the process of selection of sample under this scheme by an example.

Example 20.4 : There are 15 higher secondary schools in a police station area. The number of teachers and students in different schools are different. Below are given the numbers of teachers and students of these schools.

Sl. No. of schools	No. of students, x_i	No. of teachers, y_i	$p_i = \frac{x_i}{S_N}$
01	500	15	0.066
02	800	30	0.105
03	400	12	0.053
04	700	25	0.092
05	600	20	0.079
06	300	15	0.039
07	550	18	0.072
08	458	16	0.060
09	570	20	0.075
10	762	28	0.100
11	380	15	0.050
12	410	18	0.054
13	260	12	0.034
14	250	11	0.033
15	660	17	0.088
Total	7600	272	1.000

Select a sample of size $n = 5$ under PPS sampling scheme without replacement. Here $S_N = \sum x_i = 7600$.

Solution : We have $N = 15$. We need $n = 5$. Let us select the first unit according to Lahiri's method of selection. According to random number table the first pair of random number is (05,555). As the number of students in 5th school is $600 > 555$, 5th school can be included in the sample. The second pair of random number is (14, 148) and hence, 14th school is selected in the sample. Here it is assumed that 5th school is not in the population when second school is selected. Thus, the 15th school is becoming 14th school when second school is selected. That is 15th school is included in the sample. Before the selection of next school it is assumed that 5th and 15th school are not in the population. The random number for the selection of third unit in the sample let us consider the pair of random numbers (11, 117). At this stage 12th number school bears serial number 11. Hence, 12th school is included in the sample. The pair of random number for 4th unit to be selected is (07, 070). The 8th school becomes the 7th school as 5th school is selected earlier. So, 8th school is included in the sample. To select 5th unit the pair of random number is (09, 092). The 11th school in the original list becomes 9th school. So, 11th school is included in the sample. Finally, the selected sample is as follows :

Sl. No. of sample observation	Sl. No. of school in the population	No. of teachers, y_i	No. of students, x_i
1	05	20	600
2	15	17	660
3	12	18	410
4	08	16	458
5	11	15	380
Total		86	2508

(b) Narain's scheme of selection : In this scheme, the primary probability of selection of sample is not needed. Let $p'_i (i = 1, 2, \dots, N)$ be an adjusted probability of selection of i -th unit. Narain proposed an inclusion probability π_i to select i -th unit in the sample. This π_i is proportional to p_i , where

$$\pi_i = np_i, \quad i = 1, 2, \dots, N.$$

In this scheme, the sampling becomes sampling without replacement and the probability of selection of second and other units becomes proportional to the adjusted probability p'_i .

Narain (1951) has discussed the method of finding the value of $\pi_i (i = 1, 2, \dots, N)$. Yates and Grundy (1953) also have discussed the method of finding π_i , but the method of finding the value of π_i is labour intensive and the sampling method is not advantageous. Brewer and Undy (1962) have discussed the sampling method when $n = 2$ and they have suggested an iterative procedure to determine the value of π_i . They have also showed that the sample estimate under this scheme is more efficient than the estimate under PPS sample with replacement. The adjusted probability p'_i for this method of sample selection can be obtained by solving the following equations.

Let the probability of inclusion of first unit in the sample be p_{i1} and the probability of inclusion of second unit be p_{i2} , where

$$\begin{aligned}
 p_{i2} &= P(y_i \text{ not included 1st}) \times P(y_i \text{ included second time/it is not included first}). \\
 &= \sum_{j \neq i=1}^N P(y_{j \neq i} \text{ selected first}) \times P(y_i \text{ included second time}/y_{j \neq i} \text{ included first}).
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{j \neq i=1}^N \frac{p_j p_i}{1 - p_j} \\
&= \sum_{j=1}^N \left[\frac{p_j}{1 - p_j} - \frac{p_i}{1 - p_i} \right] p_i = \left[S - \frac{p_i}{1 - p_i} \right] p_i.
\end{aligned}$$

Here $S = \sum_{j=1}^N \frac{p_j}{1 - p_j}$, where $p_{i1} = p_i$.

Again, $\pi_i = p_{i1} + p_{i2} = P$ [y_i included in the sample of size $n = 2$]

$$= p_i \left[S + 1 - \frac{p_i}{1 - p_i} \right].$$

$$\therefore \pi_i = p'_i \left[S' + 1 - \frac{p'_i}{p - p'_i} \right] = n p_i, \text{ where } S' = \sum_{i=1}^N \frac{p'_i}{1 - p'_i}.$$

Solving the equation the value of π_i can be found out. Let π_{ij} be the probability of inclusion of i -th unit at first and at second step j -th ($j \neq i$) unit is included in the sample. Then

$$\sum_{i \neq j} \pi_{ij} = \pi_i.$$

(c) **Sen-Midzuno method** : Midzuno (1952) and Sen (1952) have discussed the method of selection of sample independently. Midzuno (1950) has suggested to select first unit with unequal probability, and the other units are selected with equal probability and without replacement. If there are N units in the population, the next $(n - 1)$ units are selected from $(N - 1)$ units without replacement by simple random sampling scheme. In this method the probability of selection of every unit and pairs of units is

$$\begin{aligned}
\pi_i &= p_{i1} + p_{i2} + \dots + p_{in} \\
&= p_{i1} + P(y_i \text{ not included in the sample first, rather it is included in other steps}) \\
&= p_i + (1 - p_i) \frac{n-1}{N-1} = \frac{N-n}{N-1} p_i + \frac{n-1}{N-1}
\end{aligned}$$

$$\begin{aligned}
\text{and } \pi_{ij} &= p_i \frac{n-1}{N-1} + p_j \frac{n-1}{N-1} + (1 - p_1 - p_2) \frac{(n-1)(n-2)}{(N-1)(N-2)}, \quad i \neq j = 1, 2, \dots, \\
&= \frac{n-1}{N-1} \left[\frac{N-n}{N-2} (p_i + p_j) + \frac{n-2}{N-2} \right].
\end{aligned}$$

$$\text{Similarly, } \pi_{ijk} = \frac{(n-1)(n-2)}{(N-1)(N-2)} \left[\frac{N-n}{N-3} (p_i + p_j + p_k) + \frac{n-3}{N-3} \right].$$

With the same argument,

$$\begin{aligned}
\pi_{ijk\dots q} &= \frac{(n-1)(n-2)(n-3)\dots 1}{(N-1)(N-2)\dots(N-n+1)} (p_i + p_j + \dots + p_q) \\
&= \frac{1}{\binom{N-1}{n-1}} (p_i + p_j + p_k + \dots + p_q).
\end{aligned}$$

Here $y_i, y_j, y_k, \dots, y_q$ are the included sample observations in the sample of size n . It is seen that, if p_i is proportional to population size, then the probability of selection of sample is proportional to the size of all units included in the sample.

20.5 Method of Estimation in PPS Sampling without Replacement

The probability of selection of any unit in case of PPS sampling without replacement is changed. The expected probability of inclusion of any unit is changed when unit is selected in the sample. To avoid this problem a new variable is considered which is correlated to the value of the variable of the unit to be investigated. The expected value of the new variable should be equal to the expected value of the study variable. In such a case, the unit which is selected first time or second time is to be considered. The order of selection is important. Thus, we have two types of estimators. These are (a) Ordered estimators, and (b) Un-ordered estimators. Das (1951), Sukhatme (1953) and Des Raj (1956) have discussed the ordered estimator. Horvitz and Thompson (1952), Murthy (1957) and Basu (1958) have discussed un-ordered estimator and they have shown that un-ordered estimator is more efficient than ordered estimator.

(a) Ordered estimator : Des Raj (1956) has proposed ordered estimator for $n = 2$. The estimator of population mean and population total are as follows :

Let y_1 and y_2 be the values of the units selected first and second times, respectively. Of course, y_1 and y_2 may not be the values of 1st and 2nd units in the population. Let p_1 be the probability of selection of y_1 and p_2 be the probability of selection of y_2 . Let us define two new variables

$$z_1 = \frac{y_1}{Np_1} \quad \text{and} \quad z_2 = \frac{1}{N} \left[y_1 + y_2 \frac{1 - p_1}{p_2} \right].$$

Then
$$\bar{z} = \frac{1}{2}(z_1 + z_2) = \frac{1}{2N} \left[(1 + p_1) \frac{y_1}{p_1} + (1 - p_1) \frac{y_2}{p_2} \right]$$

and
$$z = \frac{1}{2} \left[(1 + p_1) \frac{y_1}{p_1} + (1 - p_1) \frac{y_2}{p_2} \right],$$
 where z is the estimator of population total.

Theorem : The sample mean \bar{z} is the unbiased estimator of population mean \bar{Y} in PPS sampling without replacement, and variance of \bar{z} is

$$V(\bar{z}) = \frac{1}{4N^2} \left\{ \left(2 - \sum_{i=1}^N p_i^2 \right) \sum_{i=1}^N \frac{y_i^2}{p_i} - \sum_{i=1}^N y_i^2 + 2 \left(\sum_{i=1}^N y_i \right) \left(\sum_{i=1}^N p_i y_i \right) \right\} - \frac{\bar{Y}^2}{2}.$$

Proof : The expected value of z_1 is

$$E(z_1) = \sum_{i=1}^N \frac{y_i}{Np_i} p_i = \frac{1}{N} \sum_{i=1}^N y_i = \bar{Y}.$$

Similarly, $E(z_2)$, if y_1 is selected first, is

$$E(z_2) = E \left[\frac{y_2(1 - p_1)}{p_2} / y_1 \right] = \sum_{j=1}^N \frac{y_j(1 - p_j)}{p_j} \frac{p_j}{1 - p_j} = N\bar{Y} - y_1.$$

Here $E(z_2)$ is calculated on the assumption that y_1 is selected first and y_1 is not in the population when second unit is selected. The above summation is for the other values of y_i except y_1 .

Therefore, $E[z_2/y_1]$ does not depend on y_1 , $\therefore E(z_2) = \bar{Y}$.

$$\therefore E(\bar{z}) = E\left[\frac{z_1 + z_2}{2}\right] = \frac{1}{2}[E(\bar{z}_1) + E(\bar{z}_2)] = \bar{Y}.$$

Therefore, \bar{z} is an unbiased estimator of \bar{Y} .

$$\begin{aligned} V(\bar{z}) &= E(\bar{z}^2) - [E(\bar{z})]^2 = E\left\{\frac{1}{2N}\left[(1+p_1)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2}\right]\right\}^2 - \bar{Y}^2 \\ &= \frac{1}{4N^2} \sum_{i=1}^N \left\{ (1+p_i)\frac{y_i}{p_i} + \frac{(1-p_j)y_j}{p_j} \right\}^2 \frac{p_i p_j}{1-p_i} - \bar{Y}^2. \end{aligned}$$

On simplification, we have

$$V(\bar{z}) = \frac{1}{4N^2} \left\{ \left(2 - \sum_{i=1}^N p_i^2\right) \sum \frac{y_i^2}{p_i} - \sum_{i=1}^N y_i^2 + 2 \left(\sum_{i=1}^N y_i\right) \left(\sum_{i=1}^N p_i y_i\right) \right\} - \frac{\bar{Y}^2}{2}.$$

Corollary : In case of PPS sampling without replacement, the estimate of population total is z , where

$$z = N\bar{z}_1 = \frac{1}{2} \left[\frac{(1+p_1)y_1}{p_1} + \frac{(1-p_1)y_2}{p_2} \right]$$

and the variance of z is

$$\begin{aligned} V(z) &= V(N\bar{z}) = N^2 V(\bar{z}) \\ &= \frac{1}{4} \left\{ \left(2 - \sum_{i=1}^N p_i^2\right) \sum_{i=1}^N \frac{y_i^2}{p_i} - \sum_{i=1}^N y_i^2 + 2 \left(\sum_{i=1}^N y_i\right) \left(\sum_{i=1}^N p_i y_i\right) \right\} - \frac{Y^2}{2}. \end{aligned}$$

Theorem : In case of PPS sampling without replacement, the estimator of variance of the estimator of population mean \bar{Y} is

$$v(\bar{z}) = \frac{(1-p_1)^2}{4N^2} \left[\frac{y_1}{p_1} - \frac{y_2}{p_2} \right]^2.$$

Proof : $V(\bar{z}) = E(\bar{z}^2) - \bar{Y}^2$

Therefore, $v(\bar{z}) = \bar{z}^2 - \bar{Y}^2$ is the estimator of $V(\bar{z})$.

We know $E[z_2/y_1] = \bar{Y}$ and $E(z_1) = \bar{Y}$.

But $E(z_1 z_2) = E[z_1 \{E(z_2/y_1)\}] = \bar{Y} E(z_1) = \bar{Y}^2$.

$\therefore z_1 z_2$ is the unbiased estimator of \bar{Y}^2 .

$$\begin{aligned} \therefore v(\bar{z}) &= \bar{z}^2 - z_1 z_2 = \frac{(z_1 + z_2)^2}{4} - z_1 z_2 = \frac{(z_1 - z_2)^2}{4} \\ &= \frac{1}{4} \left[\frac{y_1}{N p_1} - \frac{1}{N} \left(y_1 + y_2 \frac{(1-p_1)}{p_2} \right) \right]^2 \\ &= \frac{1}{4N^2} (1-p_1)^2 \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2 \end{aligned}$$

Corollary : In case of PPS sampling without replacement, the estimator of variance of the estimator of population total z is

$$v(z) = \frac{1}{4}(1 - p_1)^2 \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2.$$

Ordered estimator when n is general : Let y_1, y_2, \dots, y_n be the values of study variable observed from n sample units. Assume that p_i is the probability of selection of i -th unit and hence it is the probability of $y_i (i = 1, 2, \dots, n)$. Let us define

$$z_i = \frac{1}{N}(y_1 + y_2 + \dots + y_{i-1} + y_i) \left(\frac{1 - p_1 - p_2 - \dots - p_{i-1}}{p_i} \right)$$

and $z_1 = \frac{y_1}{Np_1}$.

Then $E(z_1) = \bar{Y}$ and $E(z_i/y_1, y_2, \dots, y_{i-1}) = \bar{Y} (i = 2, 3, \dots, n)$.

This means that the above expectation is independent of y_1, y_2, \dots, y_{i-1} .

$\therefore E(z_i) = \bar{Y}; i = 1, 2, \dots, n$.

Theorem : In PPS sampling without replacement, the estimate of population mean when n is general is

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$$

and the estimated variance of \bar{z} is

$$v(\bar{z}) = \bar{z}^2 - \frac{1}{n(n-1)} \sum_{i \neq j} z_i z_j = \frac{1}{n(n-1)} \sum_{i=1}^n (z_i - \bar{z})^2.$$

Proof : $E(\bar{z}) = \frac{1}{n} \sum_{i=1}^n E(z_i) = \frac{1}{n} \sum_{i=1}^n \bar{Y} = \bar{Y} \quad [\because E(\bar{z}) = \bar{Y}]$.

Now, $V(\bar{z}) = E(\bar{z}^2) - [E(\bar{z})]^2 = E(\bar{z}^2) - \bar{Y}^2$.

Therefore, the estimator of $V(\bar{z})$ is $\bar{z}^2 - \bar{Y}^2$.

But, if $i < j$,

$$E(z_i z_j) = E\{z_i [E(z_j/y_1, y_2, \dots, y_{j-1})]\} = E(z_i \bar{Y}) = \bar{Y}^2.$$

$\therefore E \left[\frac{1}{n(n-1)} \sum_{i \neq j} z_i z_j \right] = \bar{Y}^2$.

Therefore, $v(\bar{z}) = \bar{z}^2 - \frac{1}{n(n-1)} \sum_{i \neq j} z_i z_j = \frac{1}{n(n-1)} \sum_{j=1}^n (z_j - \bar{z})^2$.

Corollary : In PPS sampling without replacement, the estimator of variance of the estimated total is

$$v(\hat{Y}) = \frac{N^2}{n(n-1)} \sum_{i=1}^n (z_i - \bar{z})^2.$$

Example 20.5 : Using data of Example 20.4 estimate (i) population mean of teachers per school, (ii) estimate population total of teachers. Also, estimate the variances of your estimates. Use Des Raj's (a) ordered estimator for $n = 2$, (b) ordered estimator for general $n(n = 5)$. Estimate 95% confidence interval for population total.

Solution : (a) Using Lahiri's method of selection the following two units are selected, where y_1 (5th unit in the population and y_2 (17th unit in the population) are the two observations ($n = 2$) in the sample. Here the unit values and the corresponding probabilities are :

$$x_1 = 600, y_1 = 20, p_1 = 0.079;$$

$$x_2 = 660, y_2 = 17, p_2 = 0.088.$$

(i) Estimate of population mean is

$$\begin{aligned}\bar{z} &= \frac{1}{2N} \left[(1 + p_1) \frac{y_1}{p_1} + (1 - p_1) \frac{y_2}{p_2} \right] \\ &= \frac{1}{2 \times 15} \left[(1 + 0.079) \frac{20}{0.079} + (1 - 0.079) \frac{17}{0.088} \right] = 15.\end{aligned}$$

(ii) Estimate of population total is $z = N\bar{z} = 15 \times 15 = 225$.

The estimates of variance of \bar{z} is

$$v(\bar{z}) = \frac{(1 - p_1)^2}{4N^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2 = \frac{(1 - 0.079)^2}{4(15)^2} \left[\frac{20}{0.079} - \frac{17}{0.088} \right]^2 = 3.39.$$

The estimates of variance of z is

$$v(z) = N^2 v(\bar{z}) = (15)^2 3.39 = 762.75.$$

(b) For $n = 5$. The selected sample observations [according to Lahiri's method] are

$$x_1 = 600, x_2 = 660, x_3 = 410, x_4 = 458, x_5 = 380;$$

$$y_1 = 20, y_2 = 17, y_3 = 18, y_4 = 16, y_5 = 15;$$

$$p_1 = 0.079, p_2 = 0.088, p_3 = 0.054, p_4 = 0.060, p_5 = 0.050.$$

$$\text{Now, } z_1 = \frac{y_1}{Np_1} = \frac{20}{15 \times 0.079} = 16.88.$$

$$z_2 = \frac{1}{N} \left[y_1 + y_2 \frac{1 - p_1}{p_2} \right] = \frac{1}{15} \left[20 + 17 \times \frac{1 - 0.079}{0.088} \right] = 13.19.$$

$$z_3 = \frac{1}{N} \left[y_1 + y_2 + y_3 \frac{(1 - p_1 - p_2)}{p_3} \right] = \frac{1}{15} \left[20 + 17 + \frac{18(1 - 0.079 - 0.088)}{0.054} \right] = 20.98.$$

$$\begin{aligned}z_4 &= \frac{1}{N} \left[y_1 + y_2 + y_3 + y_4 \frac{1 - p_1 - p_2 - p_3}{p_4} \right] \\ &= \frac{1}{15} \left[20 + 17 + 18 + \frac{16(1 - 0.079 - 0.088 - 0.054)}{0.060} \right] = 17.52.\end{aligned}$$

$$\begin{aligned}z_5 &= \frac{1}{N} \left[y_1 + y_2 + y_3 + y_4 + y_5 \frac{1 - p_1 - p_2 - p_3 - p_4}{p_5} \right] \\ &= \frac{1}{15} \left[20 + 17 + 18 + 16 + \frac{15(1 - 0.079 - 0.088 - 0.054 - 0.060)}{0.050} \right] = 19.11.\end{aligned}$$

$$\therefore \text{(i) } \bar{z} = \frac{1}{n} \sum z_i = \frac{1}{5} [16.88 + 13.19 + 20.98 + 17.521 + 19.11] = 17.536 \approx 18$$

and (ii) $z = N\bar{z} = 15 \times 18 = 270$.

$$\begin{aligned} v(\bar{z}) &= \frac{1}{n(n-1)} \sum (z_i - \bar{z})^2 = \frac{1}{n-1} \left[\sum z_i^2 - \frac{(\sum z_i)^2}{N} \right] \\ &= \frac{1}{5(5-1)} [1571.2134 - 1537.5565] = 1.6828 \end{aligned}$$

and $v(z) = N^2 v(\bar{z}) = (15)^2 \times 1.6828 = 378.63$.

95% confidence interval for population total (Y) is

$$\hat{Y}_L = z - t_{0.05,4} \text{ s.e. } (z) = 270 - 2.776 \times 19.46 = 216.$$

$$\hat{Y}_U = z + t_{0.05,4} \text{ s.e. } (z) = 270 + 2.776 \times 19.46 = 324.$$

(b) Unordered Estimator : Two unordered estimators are discussed here. These are (i) Horvitz-Thompson Estimator, and (ii) Murthy's Estimator.

For the estimation procedure the order of selected unit is not considered. Let us consider that there are N units in a finite population and n sample units are selected under PPS scheme without replacement. The n units can be ordered in $n! = M$ ways. Therefore, an unordered sample of size n can be considered equivalent to M -ordered sample. Let \bar{z}_{ij} be the estimate of any parameter of population. This estimator is obtained from j -th order of i -th sample [$i = 1, 2, \dots, \binom{N}{n}$; $j = 1, 2, \dots, M$]. Consider that p_{ij} is the probability of selection of j -th order of i -th sample. Then

$$p_i = \sum_{j=1}^M p_{ij}.$$

Let \bar{z}_1 be an unordered estimator, where

$$\bar{z}_1 = \sum_{j=1}^M \bar{z}_{ij} p'_{ij}, \quad \text{where } p'_{ij} = \frac{p_{ij}}{p_i}.$$

$$\text{Now, } E(\bar{z}_1) = E \left[\sum_{j=1}^M \bar{z}_{ij} p'_{ij} \right] = \sum_{i=1}^{\binom{N}{n}} \left[\sum_{j=1}^M \bar{z}_{ij} p'_{ij} \right] p_i = \sum_{i=1}^{\binom{N}{n}} \sum_{j=1}^M \bar{z}_{ij} \cdot p_{ij} = E[\bar{z}_{ij}].$$

\therefore unordered estimator is unbiased.

The variance of the ordered estimator \bar{z}_{ij} is

$$V(\bar{z}_{ij}) = E[\bar{z}_{ij}^2] - [E(\bar{z}_{ij})]^2 = \sum_{i=1}^{\binom{N}{n}} \sum_{j=1}^M \bar{z}_{ij}^2 p_{ij} - \left[\sum_{i=1}^{\binom{N}{n}} \sum_{j=1}^M \bar{z}_{ij} p_{ij} \right]^2$$

The variance of the unordered estimator \bar{z}_1 is

$$V(\bar{z}_1) = E(\bar{z}_1^2) - [E(\bar{z}_1)]^2 = \sum_{i=1}^{\binom{N}{n}} \left[\sum_{j=1}^M \bar{z}_{ij} p'_{ij} \right]^2 p_i - \left[\sum_{i=1}^{\binom{N}{n}} \sum_{j=1}^M \bar{z}_{ij} p_{ij} \right]^2$$

$$\begin{aligned}
\text{Then } V(\bar{z}_{ij}) - V(\bar{z}_1) &= \sum_{i=1}^{\binom{N}{n}} \sum_{j=1}^M \bar{z}_{ij}^2 p_{ij} - \sum_{i=1}^{\binom{N}{n}} \left[\sum_{j=1}^M \bar{z}_{ij} p_{ij} \right]^2 \\
&= \sum_i \sum_j \bar{z}_{ij}^2 p'_{ij} p_i - \sum_i \left[\sum_j \bar{z}_{ij} p'_{ij} \right]^2 p_i \\
&= \sum_i p_i \left[\sum_j \bar{z}_{ij}^2 p'_{ij} - \left\{ \sum_j \bar{z}_{ij} p'_{ij} \right\}^2 \right] \\
&= \sum_i p_i \left[\sum_j p'_{ij} (\bar{z}_{ij} - \bar{z}_1)^2 \right] \\
&= \sum_i \sum_j p_{ij} (\bar{z}_{ij} - \bar{z}_1)^2 \\
&= \text{a +ve term.}
\end{aligned}$$

$$\therefore V(\bar{z}_{ij}) > V(\bar{z}_1).$$

This implies that the unordered estimator \bar{z}_1 is more efficient than ordered estimator \bar{z}_{ij} . This is true for all estimators. However, $V(\bar{z}_{ij}) = V(\bar{z}_1)$, if all ordered sample is equivalent to any unordered sample.

It is seen that

$$V(\bar{z}_1) = V(\bar{z}_{ij}) - \sum_{i=1}^{\binom{N}{n}} \sum_{j=1}^M p_{ij} (\bar{z}_{ij} - \bar{z}_1)^2.$$

From the above result, it can be written :

$$\begin{aligned}
\text{Estimate } [V(\bar{z}_1)] &= \text{Estimate } [V(\bar{z}_{ij})] - \text{Estimate} \left[\sum_i \sum_j p_{ij} (\bar{z}_{ij} - \bar{z}_1)^2 \right] \\
&= \sum_{j=1}^M p'_{ij} \text{Estimate} [V(\bar{z}_{ij})] - \sum_{j=1}^M p'_{ij} (\bar{z}_{ij} - \bar{z}_1)^2.
\end{aligned}$$

Now, to obtain the unbiased estimator of $V(\bar{z}_1)$ let us consider a sample of size $n = 2$ and consider the estimate of Des Raj (1956) for \bar{z}_{ij} .

Let y_i and y_j be the sample observations in a sample of size $n = 2$ and these are selected under PPS sampling scheme without replacement. Consider that p_i is the probability of selection of y_i and p_j is the probability of selection of y_j . If these observations are ordered, then the total number of order becomes $M = 2! = 2$. Consider that y_i is selected first and y_j is selected second. Then the Des Raj's estimator is

$$\bar{z}_{ij} = \frac{1}{2N} \left[(1 + p_i) \frac{y_i}{p_i} + (1 - p_i) \frac{y_j}{p_j} \right].$$

But, if y_j is selected first, then

$$\bar{z}_{i2} = \frac{1}{2N} \left[(1 + p_j) \frac{y_j}{p_j} + (1 - p_j) \frac{y_i}{p_i} \right].$$

Again, let us consider that y_i is selected first and y_j is selected second. Then the probability of selection of these observations is

$$p_{i1} = \frac{p_i p_j}{1 - p_i}.$$

If y_j is selected first and y_i is selected second, then

$$p_{i2} = \frac{p_i p_j}{1 - p_j}.$$

Therefore, the probability of selection of 2 units in the sample is

$$p_i = \sum_{j=1}^M p_{ij} = p_{i1} + p_{i2} = p_i p_j \left(\frac{1}{1 - p_i} + \frac{1}{1 - p_j} \right) = \frac{p_i p_j (2 - p_i - p_j)}{(1 - p_i)(1 - p_j)}.$$

$$\therefore p'_{i1} = \frac{p_{i1}}{p_i} = \frac{1 - p_j}{2 - p_i - p_j} \quad \text{and} \quad p'_{i2} = \frac{p_{i2}}{p_i} = \frac{1 - p_i}{2 - p_i - p_j}.$$

Using these probabilities the unordered estimator of population mean is

$$\begin{aligned} \bar{z}_1 &= \sum_{j=1}^2 z_{ij} p'_{ij} = \frac{1}{2N} \left[\left\{ (1 + p_i) \frac{y_i}{p_i} + (1 - p_i) \frac{y_i}{p_j} \right\} \frac{1 - p_j}{2 - p_i - p_j} \right. \\ &\quad \left. + \left\{ (1 + p_j) \frac{y_j}{p_j} + (1 - p_j) \frac{y_i}{p_i} \right\} \frac{1 - p_i}{2 - p_i - p_j} \right] \\ &= \frac{(1 - p_j) \frac{y_i}{p_i} + (1 - p_i) \frac{y_j}{p_j}}{N(2 - p_i - p_j)}. \end{aligned}$$

This unordered estimator is proposed with respect to Des Raj's (1956) estimator. Here \bar{z}_{i1} and \bar{z}_{i2} are equivalent to the estimator of \bar{z}_1 of Des Raj (1956) for $n = 2$. Therefore, the unbiased estimators of variance of \bar{z}_{i1} and \bar{z}_{i2} are

$$v(\bar{z}_{i1}) = \frac{1}{4N^2} (1 - p_i)^2 \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \quad \text{and} \quad v(\bar{z}_{i2}) = \frac{1}{4N^2} (1 - p_j)^2 \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2.$$

On simplification the unbiased estimator of \bar{z}_1 is

$$v(\bar{z}_1) = \frac{(1 - p_i - p_j)(1 - p_i)(1 - p_j)}{N^2(2 - p_i - p_j)^2} \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2.$$

The estimator of population total Y is

$$\hat{Y} = \frac{(1 - p_j) \frac{y_i}{p_i} + (1 - p_i) \frac{y_j}{p_j}}{(2 - p_i - p_j)}$$

and the estimator of variance of \hat{Y} is

$$v(\hat{Y}) = N^2 v(\bar{z}_1) = \frac{(1 - p_i - p_j)(1 - p_i)(1 - p_j)}{(2 - p_i - p_j)^2} \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2.$$

Example 20.6 : Using the data of Example 20.5 find unordered estimate of average number of teachers per school. Also, estimate the variance of your estimator. Estimate total number of teachers in the schools along with estimated variance of your estimator.

Solution : We have [Example 20.5]

$$x_1 = 600, y_1 = 20, p_1 = 0.079, N = 15;$$

$$x_2 = 660, y_2 = 17, p_2 = 0.088.$$

The estimate of average teachers per school is

$$\bar{z}_1 = \frac{(1 - p_j) \frac{y_1}{p_1} + (1 - p_i) \frac{y_2}{p_2}}{N(2 - p_1 - p_2)}$$

$$\bar{z}_1 = \frac{(1 - 0.088) \frac{20}{0.079} + (1 - 0.079) \frac{17}{0.088}}{15(2 - 0.079 - 0.088)} = 15.$$

The estimate of total teachers is $\hat{Y} = N\bar{z}_1 = 15 \times 15 = 225$.

The estimate of variance of \bar{z}_1 is

$$\begin{aligned} v(\bar{z}_1) &= \frac{(1 - p_i - p_j)(1 - p_i)(1 - p_j)}{N^2(2 - p_i - p_j)^2} \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \\ &= \frac{(1 - p_1 - p_2)(1 - p_1)(1 - p_2)}{N^2(2 - p_1 - p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2, \quad i = 1, j = 2 \\ &= \frac{(1 - 0.079 - 0.088)(1 - 0.079)(1 - 0.088)}{(15)^2(2 - 0.079 - 0.088)^2} \left(\frac{20}{0.079} - \frac{17}{0.088} \right)^2 = 3.33. \end{aligned}$$

The estimate of variance of \hat{Y} is

$$v(\hat{Y}) = N^2 v(\bar{z}_1) = (15)^2 \times 3.33 = 749.25.$$

Horvitz-Thompson estimator : We have considered unordered estimator \bar{z}_1 for $n = 2$ observations. This calculation of \bar{z}_1 will be lengthy if n exceeds 2. To avoid this problem Horvitz and Thompson (1952) have suggested an estimator.

Let there be N units in a population and n units be selected from N units under PPS scheme without replacement. The values of study variables in the population are y_1, y_2, \dots, y_N . For the estimation the sample observations are selected in such a way that every unit is selected with some defined probability distribution. The probability distribution may be or may not be dependent on the elementary probability of selection of the first unit.

Let us consider a variable the value of which is a_i when i -th unit is included in the sample, where

$$a_i = 1, \text{ if } y_i \text{ is included in the sample,}$$

$$a_i = 0, \text{ if } y_i \text{ is not included in the sample.}$$

Also, consider a constant C_i which is associated with the inclusion of i -th unit in the sample. Let the probability of selection of i -th unit in the sample be p_i , where $p_i = \frac{x_i}{S_N}$ ($i = 1, 2, \dots, N$). Then the probability of inclusion of i -th unit in the sample is

$$\begin{aligned} \pi_i &= p_i + \sum_{j \neq i} \frac{p_i p_j}{(1 - p_j)} = p_i \left[1 + \sum_{j \neq i} \frac{p_j}{1 - p_j} \right] \\ &= p_i \left[S + 1 - \frac{p_i}{1 - p_i} \right], \quad S = \sum_{i=1}^N \frac{p_i}{1 - p_i}. \end{aligned}$$

Also, consider that the probability of inclusion of i -th and j -th unit in the sample is

$$\pi_{ij} = \frac{p_i p_j}{1 - p_i} + \frac{p_i p_j}{1 - p_j} = p_i p_j \left[\frac{1}{1 - p_i} + \frac{1}{1 - p_j} \right].$$

From the above discussion, it is seen that, if $n = 1$, $a_i = 1$ or 0 . This a_i follows binomial distribution, where $E(a_i) = \pi_i$ and $V(a_i) = \pi_i(1 - \pi_i)$. If i -th and j -th units are different and if these units are included in the sample, then $a_i a_j = 1$ and $\text{Cov}(a_i, a_j) = \pi_{ij} - \pi_i \pi_j$.

Let $z_i = \frac{n}{N} \frac{y_i}{\pi_i}$, $i = 1, 2, \dots, N$.

Then $\bar{z} = \frac{1}{n} \sum z_i = \frac{1}{n} \sum \frac{n}{N} \frac{y_i}{\pi_i} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{\pi_i} = \frac{1}{N} \sum_{i=1}^N a_i \frac{y_i}{\pi_i}$.

Now, $E(\bar{z}) = \frac{1}{N} \sum_{i=1}^N E(a_i) \frac{y_i}{\pi_i} = \frac{1}{N} \sum_{i=1}^N y_i = \bar{Y}$.

Therefore, \bar{z} is an unbiased estimator of \bar{Y} and $\hat{Y} = N\bar{z}$ is an unbiased estimator of population total Y .

Theorem : The variance of estimator of population mean when sample is selected under PPS scheme without replacement is

$$V(\bar{z}) = \frac{1}{N^2} \left[\sum_{i=1}^N (1 - \pi_i) \frac{y_i^2}{\pi_i} + \sum_i \sum_{j \neq i}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j \right].$$

Proof : We have $z_i = \frac{n}{N} \frac{y_i}{\pi_i}$, $i = 1, 2, \dots, N$ and $\bar{z} = \frac{1}{n} \sum z_i = \frac{1}{N} \sum_{i=1}^N a_i \frac{y_i}{\pi_i}$.

$$\begin{aligned} V(\bar{z}) &= V \left[\frac{1}{N} \sum a_i \frac{y_i}{\pi_i} \right] \\ &= \frac{1}{N^2} \left[\sum_{i=1}^N \frac{y_i^2}{\pi_i^2} V(a_i) + \sum_{i=1}^N \sum_{j \neq i=1}^N \frac{y_i y_j}{\pi_i \pi_j} \text{Cov}(a_i, a_j) \right] \\ &= \frac{1}{N^2} \left[\sum_{i=1}^N \pi_i (1 - \pi_i) \frac{y_i^2}{\pi_i^2} + \sum_{j \neq i=1}^N \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j) \right] \\ &= \frac{1}{N^2} \left[\sum_{i=1}^N (1 - \pi_i) \frac{y_i^2}{\pi_i} + \sum_i \sum_j \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j) \right]. \end{aligned}$$

This variance depends on π_i, π_j and π_{ij} . The values of π_i, π_j and π_{ij} are to be selected on the basis of sampling scheme.

Corollary : In PPS sampling without replacement the estimated variance of the unordered estimator of population mean is

$$v(\bar{z}) = \frac{1}{N^2} \left\{ \sum_{i=1}^n \frac{1 - \pi_i}{\pi_i^2} y_i^2 + \sum_{i \neq j}^n \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j \right\}.$$

Corollary : In PPS sampling without replacement the estimate of population total is

$$\hat{Y} = N\bar{z}$$

and its variance is

$$V(\hat{Y}) = \sum_{i=1}^N (1 - \pi_i) \frac{y_i^2}{\pi_i} + \sum_i \sum_{i \neq j=1} \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j).$$

The unbiased estimator of $V(\hat{Y})$ is

$$v(\hat{Y}) = \sum_{i=1}^n (1 - \pi_i) \frac{y_i^2}{\pi_i^2} + \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} y_i y_j.$$

Example 20.7 : Using the information of Example 20.5 estimate the average number of teachers per school along with its estimated variance. Use Horvitz-Thompson estimator.

Solution : We have [from Example 20.5]

$$x_1 = 600, y_1 = 20, p_1 = 0.079; N = 15$$

$$x_2 = 660, y_2 = 17, p_2 = 0.088; n = 2.$$

$$S = \sum_{i=1}^{15} \frac{p_i}{1 - p_i} = 1.0808.$$

$$\text{Then } \pi_1 = p_1 \left[S + 1 - \frac{p_1}{1 - p_1} \right] = 0.079 \left[1.0808 + 1 - \frac{0.079}{1 - 0.079} \right] = 0.1576.$$

$$\pi_2 = p_2 \left[S + 1 - \frac{p_2}{1 - p_2} \right] = 0.088 \left[1.0808 + 1 - \frac{0.088}{1 - 0.088} \right] = 0.1746.$$

$$\pi_{12} = p_1 p_2 \left[\frac{1}{1 - p_1} - \frac{1}{1 - p_2} \right] = 0.079 \times 0.088 \left[\frac{1}{1 - 0.079} - \frac{1}{0.088} \right] = 0.0152.$$

$$\text{Now, } \bar{z} = \frac{1}{N} \left(\frac{y_1}{\pi_1} + \frac{y_2}{\pi_2} \right) = \frac{1}{15} \left(\frac{20}{0.1576} + \frac{17}{0.1746} \right) = 15.$$

The estimate of variance of \bar{z} is

$$\begin{aligned} v(\bar{z}) &= \frac{1}{N^2} \left[\sum_{i=1}^n \frac{1 - \pi_i}{\pi_i^2} y_i^2 + \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} y_i y_j \right] \\ &= \frac{1}{(15)^2} \left[\frac{1 - 0.1576}{(0.1576)^2} (20)^2 + \frac{1 - 0.1746}{(0.1746)^2} (17)^2 \right. \\ &\quad \left. + \frac{2(0.0152 - 0.1576 \times 0.1746) \times 20 \times 17}{0.0152 \times 0.1576 \times 0.1746} \right] = 6.07. \end{aligned}$$

This variance is more than the variance of the unordered estimator \bar{z}_1 given in Example 20.6. This problem of higher variance can be obviated, if π_1 , π_2 and π_{12} are calculated according to the suggestion of Sen-Midzuno, where

$$\pi_1 = \frac{N - n}{N - 1} p_1 + \frac{n - 1}{N - 1} = \frac{15 - 2}{15 - 1} \times 0.079 + \frac{2 - 1}{15 - 1} = 0.1448.$$

$$\pi_2 = \frac{N-n}{N-1}p_2 + \frac{n-1}{N-1} = \frac{15-2}{15-1} \times 0.088 + \frac{2-1}{15-1} = 0.1531.$$

$$\begin{aligned} \pi_{12} &= \frac{n-1}{N-1} \left[\frac{N-n}{N-2}(p_1 + p_2) + \frac{n-2}{N-2} \right] \\ &= \frac{2-1}{15-1} \left[\frac{15-2}{15-2}(0.079 + 0.088) + \frac{2-2}{15-2} \right] = 0.0119. \end{aligned}$$

Now, the estimated variance of Horvitz-Thompson estimator is

$$\begin{aligned} v(\bar{z}) &= \frac{1}{N^2} \left[\sum_{i=1}^n \frac{(1-\pi_i)y_i^2}{\pi_i^2} + \sum_{i \neq j} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}\pi_i\pi_j} y_i y_j \right] \\ &= \frac{1}{(15)^2} \left[\frac{(1-0.1448)(20)^2}{(0.1448)^2} + \frac{(1-0.1531)(17)^2}{(0.1531)^2} \right. \\ &\quad \left. + 2 \times \frac{0.0119 - 0.1448 \times 0.1531 \times 20 \times 17}{0.0119 \times 0.1448 \times 0.1531} \right] = 1.28. \end{aligned}$$

It is seen that Horvitz-Thompson estimator is more efficient when inclusion probability is calculated according to the suggestion of Sen-Midzuno.

Murthy's Unordered Estimator : Let there be N units in a population. We need a sample of size $n(\leq N)$. There will be $\binom{N}{n}$ unordered sample. Each unordered sample can be ordered in M ways. Let us consider the case of selection of a sample, where p_l is the probability of selection of l -th sample [$l = 1, 2, \dots, \binom{N}{n}$]. Let p_{il} be the probability of i -th ordered sample corresponding to l -th unordered sample. Then

$$p_l = \sum_{i=1}^M p_{il}.$$

Let $\hat{\theta}_{il}$ be the estimator of population parameter θ , where $\hat{\theta}_{il}$ is obtained from i -th ordered sample corresponding to l -th unordered sample. The unordered estimator of θ is

$$\hat{\theta} = \sum_{i=1}^M \hat{\theta}_{il} p'_{il}, \quad \text{where } p'_{il} = \frac{p_{il}}{p_i}.$$

Theorem : In PPS sampling without replacement the unbiased estimator of parameter θ is $\hat{\theta}$ and the variance of $\hat{\theta}$ is

$$V(\hat{\theta}) = \sum_{i=1}^{\binom{N}{n}} p_i \left[\left(\sum_{i=1}^M \hat{\theta}_{il} p'_{il} \right)^2 - \left(\sum_{l=1}^{\binom{N}{n}} \sum_{i=1}^M \hat{\theta}_{il} p'_{il} \right)^2 \right], \quad \text{where } \hat{\theta} = \sum_{i=1}^M \hat{\theta}_{il} p'_{il}.$$

Proof : Given $\hat{\theta} = \sum_{i=1}^M \hat{\theta}_{il} p'_{il}$.

$$E(\hat{\theta}) = \sum_{i=1}^M E(\hat{\theta}_{il}) \frac{p_{il}}{p_i} = \theta \sum_{i=1}^M \frac{p_{il}}{p_l} = \theta.$$

$$\begin{aligned} V(\hat{\theta}) &= E(\hat{\theta}^2) - [E(\hat{\theta})]^2 = E\left(\sum_{i=1}^M \theta_{il} p'_{il}\right)^2 - \left(E \sum_{i=1}^M \hat{\theta}_{il} p'_{il}\right)^2 \\ &= \sum_l p_l \left(\sum_i \hat{\theta}_{il} p'_{il}\right)^2 - \left(\sum \sum \hat{\theta}_{il} \cdot p_{il}\right)^2, \end{aligned}$$

The unbiased estimator of this variance is

$$v(\hat{\theta}) = \left(\frac{1}{p_l}\right)^2 \sum_{i=1}^n \sum_{j>1}^n \{p_i P(l_{ij}) - p_{il} p_{ij}\} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

Here $P(l_{ij})$ is the conditional probability of selection of i -th and j -th unit in l -th sample.

Let us, now, discuss the unordered estimator of Murthy for $n = 2$. Assume that y_i and y_j are selected with preliminary probabilities p_i and p_j . These two observations can be ordered in $M = 2$ ways. Then the ordered estimator of Murthy is

$$\hat{\theta}_{11} = \frac{1}{2} \left[(1 + p_1) \frac{y_1}{p_1} + (1 - p_1) \frac{y_2}{p_2} \right],$$

if y_i is included first in the sample.

if y_j is included in the sample first, then

$$\hat{\theta}_{12} = \frac{1}{2} \left[(1 + p_2) \frac{y_2}{p_2} + (1 - p_2) \frac{y_1}{p_1} \right], \quad i = 1, 2.$$

The i -th and j -th units will be included in the samples first and second, respectively with probability,

$$p_{1i} = \frac{p_i p_j}{1 - p_i} = p_1 p_2 (1 - p_1).$$

Similarly, the probability of j -th and i -th units to be included in the samples first and second, respectively is

$$p_{2i} = \frac{p_1 p_2}{1 - p_2}.$$

Therefore, the unordered estimator of θ is

$$\hat{\theta} = \sum_{i=1}^2 \theta_{il} p'_{il} = \sum \hat{\theta}_{il} p_{il} / p_l.$$

$$\text{Here } p_l = \sum_{i=1}^2 p_{il} = p_{1l} + p_{2l} = p_1 p_2 \left[\frac{1}{1 - p_1} + \frac{1}{1 - p_2} \right].$$

$$\begin{aligned} \therefore \hat{\theta} &= \frac{1}{p_l} \left[\frac{1}{2} \left\{ (1 + p_1) \frac{y_1}{p_1} + (1 - p_1) \frac{y_2}{p_2} \right\} \frac{p_1 p_2}{1 - p_1} + \frac{1}{2} \left\{ (1 + p_2) \frac{y_2}{p_2} + (1 - p_2) \frac{y_1}{p_1} \right\} \frac{p_1 p_2}{1 - p_2} \right] \\ &= \left[(1 - p_2) \frac{y_1}{p_1} + (1 - p_1) \frac{y_2}{p_2} \right] / (2 - p_1 - p_2) \end{aligned}$$

The variance of $\hat{\theta}$ is

$$V(\hat{\theta}) = \sum_{i=1}^2 p_1 p_2 \frac{(1 - p_1 - p_2)}{(2 - p_1 - p_2)} \left(\frac{y_i - y_j}{p_i p_j} \right)^2.$$

The unbiased estimator of $V(\hat{\theta})$ is

$$v(\hat{\theta}) = \frac{(1 - p_1)(1 - p_2)(1 - p_1 - p_2)}{(2 - p_1 - p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2.$$

This variance is always positive.

Example 20.8 : Using the data of Example 20.7 let us estimate the average number of teachers per school along with the estimated variance following the suggestion of Murthy.

Solution : We have $N = 15$, $n = 2$. The information of sample observations are

$$x_1 = 600, y_1 = 20. p_1 = 0.079$$

$$x_2 = 660, y_2 = 17. p_2 = 0.088.$$

Here
$$S = \sum_{i=1}^{15} \frac{p_i}{1 - p_i} = 1.0808.$$

Now,
$$\hat{\theta} = \left[(1 - p_2) \frac{y_1}{p_1} + (1 - p_1) \frac{y_2}{p_2} \right] / (2 - p_1 - p_2)$$

$$= \left[(1 - 0.088) \frac{20}{0.079} + (1 - 0.079) \frac{17}{0.088} \right] / (2 - 0.079 - 0.088)$$

$$= 223.$$

Thus, the average number of teacher per school is

$$\hat{\mu} = \frac{\hat{\theta}}{N} = \frac{223}{15} \approx 15.$$

Here $\hat{\theta}$ is the estimator of total teachers in the school.

The estimate of variance of $\hat{\theta}$ is

$$v(\hat{\theta}) = \frac{(1 - p_1)(1 - p_2)(1 - p_1 - p_2)}{(2 - p_1 - p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2$$

$$= \frac{(1 - 0.079)(1 - 0.088)(1 - 0.079 - 0.088)}{(2 - 0.079 - 0.088)^2} \left[\frac{20}{0.079} - \frac{17}{0.088} \right]^2$$

$$= 749.25.$$

$$\therefore v(\hat{\mu}) = \frac{v(\hat{\theta})}{N^2} = \frac{749.25}{225} = 3.33.$$

If $n = 2$, then the estimator of population mean according to Murthy's unordered estimator is

$$\hat{\mu} = \frac{1}{N(2 - p_1 - p_2)} \left[(1 - p_2) \frac{y_1}{p_1} + (1 - p_1) \frac{y_2}{p_2} \right]$$

and the unbiased estimator of $V(\hat{\mu})$ is

$$v(\hat{\mu}) = \frac{(1 - p_1)(1 - p_2)(1 - p_1 - p_2)}{N^2(2 - p_1 - p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2.$$

20.6 Other Methods of Sampling with Varying Probabilities

The methods of sampling with varying probabilities discussed in the previous sections are not suitable for practical application. The complexity arises either in selection procedure or in estimation, specially in the estimation of variance of the estimator, when $n = 2$. In some cases, the complexity arises in calculating modified probability based on preliminary probability of selection of unit. The problem increases, if the value of n increases.

To avoid the problem Stevens (1958) has discussed a method. He has suggested to make k clusters of the population units. The clustering is done, according to him, by considering an auxiliary variable x so that population units are measured depending on the values of x and then population units are clustered into k clusters. After that n clusters are selected without replacement. The clusters are selected in such a way that the probability of selection of a cluster is proportional to total size of the population units. If a cluster is selected t times, then t units are selected from that cluster and each unit is selected with equal probability without replacement. Let us consider that i -th cluster ($i = 1, 2, \dots, k$) has N_i units and the size of each unit is x_i . Let us consider that the preliminary probability of selection of every unit of N_i units is $p_i = x_i/X$, where

$$X = \sum_{i=1}^k N_i x_i.$$

This X is the total size of population units.

Let the value of j -th unit of i -th cluster is y_{ij} . Then according to Stevens (1958) the estimate of population mean \bar{Y} is

$$\hat{\mu} = \frac{1}{Nn} \sum \frac{y_{ij}}{p_i}.$$

The sum is for all the units included in the sample. Cochran (1963) has discussed the method of clustering in detail. He suggested to make n clusters and from each cluster one unit is selected with probability proportional to the size of units in the cluster. Let us consider that x_{ij} is the size of j -th unit in i -th cluster and $X_i = \sum_j x_{ij}$. This X_i is the total units in i -th cluster. The probability of selection of j -th unit from i -th cluster is x_{ij}/X_i .

As one unit is selected from each cluster, the value of i -th unit selected from i -th cluster is assumed y_i and the size of this unit is x_i . Then the estimator of population mean is

$$\hat{\mu} = \frac{1}{N} \sum \frac{y_i}{x_i/X_i} = \frac{1}{N} \sum_{i=1}^n \frac{X_i y_i}{x_i}.$$

Here, if $X_1 = X_2 = \dots = X_n$, then the probability of selection of y_i is maintained and this probability is p_i . But, the estimator of variance of $\hat{\mu}$ is not available, if the sample is selected according to the method of selection discussed in the section.

Rao, Hartley and Cochran (1962) have proposed another method of sample selection. They have suggested to make n clusters with the N units in the population. Let there be N_i units in i -th cluster ($i = 1, 2, \dots, n$) and y_i be the value of the variable under study of i -th unit and p_i be the probability of selection of i -th unit. Then the estimator of population mean is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{p_i/\pi_i} = \frac{1}{N} \sum_{i=1}^n \frac{\pi_i y_i}{p_i}, \quad \text{where } \pi_i = \sum_{j=1}^{N_i} p_{ij}.$$

Theorem : In PPS sampling under random clustering of population units the Rao-Hartley-Cochran estimator $\hat{\mu}$ is unbiased estimator of population mean and the variance of this estimator is

$$V(\hat{\mu}) = \frac{\sum_{i=1}^n N_i^2 - N}{N^2(N-1)} \left[\sum_{j=1}^{N_i} \frac{y_j^2}{p_j} - Y^2 \right],$$

where $\hat{\mu} = \frac{1}{N} \sum_{i=1}^n \frac{\pi_i y_i}{p_i}$ and Y is the population total.

Proof : If $\hat{\mu}$ is the estimator of population mean μ , then $E(\hat{\mu}) = \mu$. In calculating expectation, we have to consider two expectation, one for the selection of j -th unit from i -th cluster. The second one is considered for the inclusion of N_1, N_2, \dots, N_n observations in n clusters.

Thus, $E(\hat{\mu}) = E_1 E_2(\hat{\mu})$, where E_1 is used for the N units to be clustered into n and E_2 is used for the selection of a cluster.

$$\begin{aligned} E(\hat{\mu}) &= E_1 \frac{1}{N} \sum_{i=1}^n E_2 \left(\frac{\pi_i y_i}{p_i} \right) = \frac{1}{N} \sum_{i=1}^n E_1 \sum \frac{p_j}{\pi_i} \frac{\pi_i y_i}{p_i} \\ &= \frac{1}{N} \sum_{i=1}^n E_1 Y_i, \quad Y_i \text{ is the total of the variables in } i\text{-th cluster.} \\ &= \frac{1}{N} \sum_{i=1}^n Y_i \\ &= \bar{Y} \quad \left[\because \sum_{i=1}^n Y_i = Y, E(Y) = Y \right]. \end{aligned}$$

$\therefore \hat{\mu}$ is the unbiased estimator of \bar{Y} . Again,

$$\begin{aligned} V(\hat{\mu}) &= E_1 V_2(\hat{\mu}) + V_1 E_2(\hat{\mu}) = E_1 V_2(\hat{\mu}) \quad [\because E_2(\hat{\mu}) = \mu, V(\mu) = 0] \\ &= \frac{1}{N^2} \left[\sum_{j=1}^{N_i} \frac{p_j}{\pi_i} \left(\frac{y_j}{p_j/\pi_i} - Y_j \right)^2 \right]. \end{aligned}$$

Here Y_j is the total value of N_j units in j -th cluster.

Again, the probability that a pair of values of a random cluster will fall in i -th cluster is $N_i(N_i - 1)/N(N - 1)$. Replacing this probability and on simplification, we have

$$V(\hat{\mu}) = \frac{\left(\sum_{i=1}^n N_i^2 - N \right)}{N^3(N-1)} \left[\sum_{j=1}^{N_i} \frac{y_j}{p_j} - Y^2 \right].$$

If $N_1 = N_2 = \dots = N_n = \frac{N}{n}$, then

$$V(\hat{\mu}) = \frac{1}{N^2} \left[\frac{1}{n} \sum_{i=1}^n p_i \{ 1 - (n-1)/N \} \left(\frac{y_i}{p_i} - Y \right)^2 \right].$$

The unbiased estimator of $V(\hat{\mu})$ is

$$v(\hat{\mu}) = \frac{1}{N^2} \left[\frac{\left(\sum_{i=1}^n N_i^2 - N \right)}{\left(N^2 - \sum_{i=1}^n N_i^2 \right)} \left\{ \sum \pi_i \left(\frac{y_i}{p_i} - \hat{\mu} \right)^2 \right\} \right]$$

This process of random clustering is different than that of random clustering suggested by Cochran (1963). Cochran (1963) has proposed that the values of N_i should be equal, as far as possible.

The method of sample selection according to the method discussed above can be applied if $n = 2$. Then

$$\hat{\mu} = \frac{1}{N} \left[\frac{y_1}{p_1/\pi_1} + \frac{y_2}{p_2/\pi_2} \right]$$

Here y_1 and y_2 are the values of study variable of two units of two clusters and p_1 and p_2 are the probabilities of two values, respectively. π_1 and π_2 are the sum of the preliminary probabilities of those two units selected from the clusters.

The $\hat{\mu}$ is also unbiased, because

$$E(\hat{\mu}/N_1, N_2) = \frac{1}{N} \left[\sum^{N_1} y_1 + \sum^{N_2} y_2 \right] = \bar{Y}$$

$$\begin{aligned} \text{and } V(\hat{\mu}) &= \frac{1}{N^2} \left[V \left(\frac{y_1}{p_1/\pi_1} \right) + V \left(\frac{y_2}{p_2/\pi_2} \right) \right] \\ &= \frac{1}{N^2} E \left\{ \sum^{N_1} \left(\frac{y_1}{p_1/\pi_1} \right)^2 \frac{p_1}{\pi_1} - N_1^2 \bar{Y}_{N_1}^2 + \sum^{N_2} \left(\frac{y_2}{p_2/\pi_2} \right)^2 \frac{p_2}{\pi_2} - N_2^2 \bar{Y}_{N_2}^2 \right\} \end{aligned}$$

Here N_1 and N_2 are the population size of those two clusters from which the two units are selected. The population mean of these clusters are \bar{Y}_{N_1} and \bar{Y}_{N_2} .

$$V(\hat{\mu}) = \frac{1}{N^2} E \left\{ \pi_1 \sum^{N_1} \frac{y_1^2}{p_1} + \pi_2 \sum^{N_2} \frac{y_2^2}{p_2} - N_1^2 \bar{Y}_{N_1}^2 - N_2^2 \bar{Y}_{N_2}^2 \right\}$$

$$\begin{aligned} \text{But } E \left(\pi \sum^{N_1} \frac{y_1^2}{p_1} \right) &= E \left\{ \left[\sum^{N_1} p_1 \right] \sum^{N_1} \frac{y_1^2}{p_1} \right\} = E \left[\sum^{N_1} y_1^2 + \sum_{1 \neq 1'}^{N_1} \frac{y_1^2}{p_1} p_1' \right] \\ &= \frac{N_1}{N} \sum^{N_1} y_1^2 + \frac{N_1(N_1 - 1)}{N(N - 1)} \sum_{1 \neq 1'} \frac{y_1^2}{p_1} p_1' \\ &= \frac{N_1 N_2}{N(N - 1)} \sum^{N_1} y_1^2 + \frac{N_1(N_1 - 1)}{N(N - 1)} \sum^{N_1} \frac{y_1^2}{p_1} \end{aligned}$$

$$\text{Similarly, } E \left[\pi_2 \sum^{N_2} \frac{y_2^2}{p_2} \right] = \frac{N_1 N_2}{N(N - 1)} \sum^{N_2} y_2^2 + \frac{N_2(N_2 - 1)}{N(N - 1)} \sum^{N_2} \frac{y_2^2}{p_2}$$

Again, $E[\bar{Y}_{N_1}^2] = \left(\frac{1}{N_1} - \frac{1}{N}\right) S^2 + \bar{Y}_N^2$

$$E[\bar{Y}_{N_2}^2] = \left(\frac{1}{N_2} - \frac{1}{N}\right) S^2 + \bar{Y}_N^2, \text{ where } S^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2.$$

$$\therefore V(\hat{\mu}) = \frac{N_1^2 + N_2^2 - N}{N(N-1)} \left[\sum_{i=1}^N \frac{y_i^2}{N^2 p_i} - \bar{Y}_N^2 \right] = \frac{2(N_1^2 + N_2^2 - N)}{N(N-1)} \frac{1}{2} \sum_{i=1}^N p_i \left(\frac{y_i}{N p_i} - \bar{Y}_N \right)^2.$$

If $N_1 = N_2 = \frac{N}{2}$, then $V(\hat{\mu})$ becomes less. If N is an even number, then $N_1 = N_2 = \frac{N}{2}$ and

$$V(\hat{\mu}) = \left(1 - \frac{1}{N-1}\right) \frac{1}{2} \sum_{i=1}^N p_i \left(\frac{y_i}{N p_i} - \bar{Y}_N \right)^2.$$

If N is an odd number, let $N_1 = \frac{N-1}{2}$ and $N_2 = \frac{N+1}{2}$, then

$$V(\hat{\mu}) = \left(1 - \frac{1}{N}\right) \frac{1}{2} \sum_{i=1}^N p_i \left(\frac{y_i}{N p_i} - \bar{Y}_N \right)^2.$$

The estimator of $V(\hat{\mu})$ is

$$v(\hat{\mu}) = \left(1 - \frac{1}{R}\right) \left[\pi_1 \left(\frac{y_1}{N p_1} - \hat{\mu} \right)^2 + \pi_{1/4} \left(\frac{y_{1/2}}{N p_2} - \hat{\mu} \right)^2 \right].$$

Here $R = \frac{N}{2}$, if N is an even number
 $= \frac{N+1}{2}$, if N is an odd number.

20.7 Sampling Procedures where Inclusion Probability is Proportional to Size (π PS Sampling)

Durbin's π PS sampling technique : According to Durbin (1967) the i -th unit is selected with probability proportional to p_i , where $p_i = \frac{x_i}{S_N}$ is the probability of selection of i -th unit.

Here x_i is the size of i -th unit and $S_N = \sum_{i=1}^N x_i$. Durbin has proposed the inclusion probability for $n = 2$. According to him the first unit is selected with probability p_i , where p_i is the preliminary probability of selection of i -th unit. The probability of selection of the second unit is conditional probability under the condition that the first unit has already been selected. Let the second selected unit is j -th unit and the first selected one is the i -th unit in the population. Then the conditional probability of selection of j -th unit under the condition that i -th unit is selected first is given by

$$p_{j/i} = p_i \left(\frac{1}{1 - 2p_i} + \frac{1}{1 - 2p_j} \right) \left(1 + \sum_{k=1}^N \frac{p_k}{1 - 2p_k} \right)^{-1}$$

Here $j \neq i$. The inclusion of probability of i -th unit to be included in the sample is $\pi_i = 2p_i$ ($i = 1, 2, 3, \dots, N$).

The inclusion probability of j -th unit to be selected second is

$$\pi_{ij} = 2p_i p_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) \left(1 + \sum_{k=1}^N \frac{p_k}{1-2p_k} \right)^{-1}$$

Using this inclusion probability of selection of i -th and j -th units the Horvitz-Thompson estimator can be found out.

Hanurav's π PS sampling technique : In this method two units are selected under PPS sampling scheme with replacement. If the selected units are different, then these two units are considered the sampling units in the sample. If the unit is the same one, then another two units are selected. In that case, i -th unit will be selected with probability proportional to p_i^2 . At this stage, if two units are different, the selected two units will be the sample units. If same unit is selected, then another two units are to be selected. The i -th unit is selected with probability proportional to p_i^4 . The probability of inclusion of units in the sample is

$$\pi_i = 2p_i$$

$$\pi_{ij} = 2p_i p_j \left(1 + \sum_{k=1}^{\infty} W_k \right), \text{ where } W_k = \frac{(p_i p_j) 2^k - 1}{S(1)S(2) \cdots S(k)}, S(t) = \sum_{j=1}^N p_j 2^t.$$

Using this inclusion probability, Horvitz-Thompson estimator of population mean can be obtained. However, the estimator of variance of this estimator is found out by the method of Yates and Grundy (1953), where

$$v(\hat{\mu}) = \frac{1}{N^2} \frac{\pi_1 \pi_2 - \pi_{12}}{\pi_{12}} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2$$

Example 20.9 : There are 30 villages in a police station area. The number of ever married couples (x_i) and the number of adopter couples (y_i) of the villages are shown below :

Sl. No. of villages	x_i	y_i	$p_i = \frac{x_i}{S_N}$	Sl. No. of villages	x_i	y_i	$p_i = \frac{x_i}{S_N}$
01	540	324	0.021	16	522	218	0.020
02	1008	570	0.039	17	1130	609	0.044
03	842	480	0.033	18	1368	700	0.054
04	932	408	0.036	19	940	500	0.037
05	424	201	0.017	20	1220	480	0.048
06	672	302	0.038	21	1167	640	0.046
07	580	300	0.023	22	452	200	0.018
08	1050	678	0.041	23	545	250	0.021
09	1250	542	0.049	24	980	544	0.038
10	1300	705	0.051	25	792	432	0.031
11	482	168	0.019	26	978	500	0.038
12	550	300	0.022	27	1250	750	0.049
13	672	360	0.038	28	1380	800	0.054
14	480	168	0.019	29	690	300	0.027
15	578	300	0.023	30	785	305	0.031
Total	11360		0.469		14199		0.556

Here $S_N = \sum^N x_i = 25559$, $N = 30$.

Let us make random clusters of villages and select a sample of size $n = 2$ to estimate average number of adopter couples per village using Rao-Hartley-Cochran estimator. Also, estimate variance of your estimator.

Solution : We have $N = 30$ (N is an even number) and hence, we can consider $N_1 = N_2 = \frac{N}{2} = 15$. Therefore, the first 15 villages can be considered as cluster-1 and second 15 villages can be considered as cluster-2. We need to select one unit from cluster-1 and one unit from cluster-2. Using Lahiri's method of sample selection and using the random number given in Appendix, we have pair of observations to select one unit from cluster-1, which is (01, 0314). So the first unit in the population is selected in the sample. The second pair of random numbers is (03, 0674). The corresponding unit in the population is 18. So, village 18 is selected in the sample. The sample information are

$$x_1 = 540, y_1 = 324, p_1 = 0.021, N_1 = 15, N = 30.$$

and $x_2 = 1368, y_2 = 700, P_2 = 0.054, N_2 = 15, n = 2$.

$$\pi_1 = \sum_{j=1}^{15} p_{1j} = 0.469, \pi_2 = \sum_{j=1}^{15} p_{2j} = 0.556.$$

The estimate of average adopter couples per village is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^2 \frac{\pi_i y_i}{p_i} = \frac{1}{30} \left[\frac{0.469 \times 324}{0.021} + \frac{0.556 \times 700}{0.054} \right] = 241.2 \approx 241.$$

The estimated variance of $\hat{\mu}$ is

$$\begin{aligned} v(\hat{\mu}) &= \left(1 - \frac{2}{N}\right) \left[\pi_1 \left(\frac{y_1}{Np_1} - \hat{\mu}\right)^2 + \pi_2 \left(\frac{y_2}{Np_2} - \hat{\mu}\right)^2 \right] \\ &= \left(1 - \frac{2}{30}\right) \left[0.469 \left(\frac{324}{30 \times 0.021} - 241\right)^2 + 0.556 \left(\frac{700}{30 \times 0.054} - 241\right)^2 \right] \\ &= \left(1 - \frac{2}{30}\right) [0.469(514.286 - 241)^2 + 0.556(432.099 - 241)^2] \\ &= \left(1 - \frac{2}{30}\right) [35027.376 + 20283.436] \\ &= 0.933 \times 55310.812 = 51604.9876. \end{aligned}$$

Chapter 21

Non-Sampling Error

21.1 Introduction

A general idea about non-sampling error is discussed in chapter 11. The objective of sampling is to estimate the population parameter using the observations recorded from sample survey. The population characteristic can also be calculated if data related to the characteristic are recorded from census. The data recorded from census or sample survey are assumed to be free of error. The errors crept in the survey data distort the estimate of the population characteristic. The data are collected by investigator using some measuring devices. Data can also be provided by the sampling unit himself. But willingly or unwillingly some error may be crept in the data by the enumerator. The error which creeps in the data during its collection is termed as non-sampling error.

The non-sampling error can be classified in two classes, viz., (a) Errors of reporting or Response errors, and (b) Non-response errors. The response error can again be classified into several classes. These are (i) Errors in sample selection, (ii) Failure to collect information from some sampling units, and (iii) Errors in pre-analysis of data. The failure of collection of data can, again, be divided into 2 classes. These are : (1) Failure to identify the sampling unit, and (2) Refuse to provide information.

Besides non-sampling errors, the sampling errors are also discussed. It is seen that the sampling error is reduced if sample size is increased. But with the increase in sample size the non-sampling errors are increased or probability of non-sampling error is increased.

Because of limitation of resources in terms of trained man power, time and money involved in the survey the probability of non-sampling error increases. As a result, the estimate of population parameter may be free of sampling error when data are collected through census but it is not free of non-sampling error.

The non-sampling error what creeps in the data may be increased in such an extent that the analytical results may provide wrong or distorted information about population characteristic. However, the estimation procedure may be modified to reduce the impact of non-sampling error. As a part of modification the resources available for the survey are to be utilized in such a way that the non-sampling error is reduced. The method of reduction of non-sampling error has been discussed by many authors. The important works in reducing non-sampling errors are proposed by Mahalonobis (1940, 1944, 1946), Deming (1944, 1950), Hansen et al. (1946, 1951), Birnbaum and Sirken (1950), Sukhatme and Seth (1952), Durbin (1954), Lahiri (1958), Moser (1958), Mahalanobis and Lahiri (1961), Zarkovich (1963, 1965, 1966), Kish (1965) and Singh et al. (1974). Hansen et al. have discussed the problems of non-sampling errors in census and sample survey data. Sukhatme and Seth have proposed some model to measure the amount of non-sampling errors. In this chapter the impact of non-sampling errors and adjustment in the estimator will be discussed.

21.2 Effects of Non-Response

Let us discuss the non-sampling error due to the failure of collection of information from some units. Let there be N units in a population. The units are divided into two strata. Let there be N_1 units in stratum-1 and N_2 units in stratum-2. Assume that if sample is selected from stratum-1, then information from sampling units will be available. The information will not be available if sample is selected from stratum-2. Further assumption is that the enumerator will try at least 3 times to collect information, so that number of units in stratum-2 is reduced.

Consider that the simple random sample is selected from this population and sample is selected only from stratum-1. Let \bar{y}_1 be the sample mean when sample is selected from stratum-1 and \bar{Y}_1 be the population mean of this stratum. The population mean of stratum-2 is \bar{Y}_2 . Let $w_1 = \frac{N_1}{N}$ and $w_2 = \frac{N_2}{N}$, where w_2 is the ratio of population units from which information are not available. Then the bias due to sampling is

$$E(\bar{y}_1) - \bar{Y} = \bar{Y}_1 - \bar{Y} = \bar{Y}_1 - (w_1\bar{Y}_1 + w_2\bar{Y}_2) = w_2(\bar{Y}_1 - \bar{Y}_2).$$

Here \bar{Y} is the population mean. As nothing is available to estimate \bar{Y}_2 , the information of bias are not available. However, if the information on \bar{Y}_2 are not available, then the information on bias of estimator are not available. However, if the information of \bar{Y}_2 are available from any other sources, then a limit of bias can be calculated.

The limit of bias can easily be calculated if the variable under study is qualitative. In that case, let P be the population proportions of units possessing a particular characteristic. The corresponding proportions are P_1 and P_2 for population units of stratum-1 and stratum-2, where $0 \leq P_2 \leq 1$. If w_2 is known, then using the limit $0 \leq P_2 \leq 1$, a confidence interval can be estimated for P . Let us consider that n units are included in the sample and $n_1 (< n)$ units are selected from stratum-1 and information are collected from these n_1 units. If n_1 is big, then 95% confidence limits of P_1 is given by

$$p_1 \pm 2\sqrt{\frac{p_1q_1}{n_1}}.$$

Here p_1 is an estimator of P_1 and $q_1 = 1 - p_1$. Now, assuming $P_2 = 0$ and $P_2 = 1$ the 95% confidence limits can be obtained by

$$\hat{P}_L = w_1 \left(p_1 - 2\sqrt{\frac{p_1q_1}{n_1}} \right) + w_2(0)$$

$$\text{and } \hat{P}_U = w_2 \left[p_1 + 2\sqrt{\frac{p_1q_1}{n}} \right] + w_2(1).$$

A formula for calculating n is proposed by Birnbaum and Sirken (1950a, 1950b). According to them, if w_2 is known, then a formula can be proposed to calculate the value of n , where

$$n = z_\alpha^2 PQ/d^2.$$

Here z_α is the 5% value of z and d is the value of error. Let $P = 0.5$, then

$$n = \frac{(1.96)^2}{4d^2}, \text{ where 5\% value of } z, \text{ i.e., } z_{0.05} = 1.96.$$

Birnbaum and Sirken have shown that, if

$$n = \frac{z_\alpha^2}{4d(d - w_2)w_1} - 1,$$

then the error in the estimator is confined to d . However, if $w_2 > d$, then the value of n is not available.

21.3 Technique for Adjustment of Non-Responses

In the previous section it is shown that the bias in the estimate of population mean is

$$E(\bar{y}_1) - \bar{Y} = w_2(\bar{Y}_1 - \bar{Y}_2).$$

This bias will be negligible, if w_2 or $\bar{Y}_1 - \bar{Y}_2$ is very small and if $\bar{Y}_1 - \bar{Y}_2$ becomes insignificant for a standard value of w_2 .

Hansen and Hurwitz (1940) have discussed a method to remove the bias due to non-response. According to them, from a sample of size $n (\leq N)$ the information are available from $n_1 (< n)$ units. Then the number of units from which information are not available are $n_2 = n - n_1$. In such a situation a sample of size $n'_2 (< n_2)$ is to be selected without replacement. Then N_1 and N_2 are to be estimated on the basis of n_1 and n_2 , where the estimators are

$$\hat{N}_1 = \frac{n_1}{n} N \quad \text{and} \quad \hat{N}_2 = \frac{n_2}{n} N.$$

The n'_2 units are to be investigated personally to collect information. These collected information and the information collected from n_1 units are to be utilized to have a combined estimate of \bar{Y} , where the combined estimator is

$$\bar{y}_w = \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}'_2), \quad \text{where} \quad \bar{y}'_2 = \frac{1}{n'_2} \sum y.$$

Theorem : In case of non-response error, if a sub-sample of n'_2 units is selected without replacement, then the combined sample estimator \bar{y}_w becomes unbiased estimator of \bar{Y} and the variance of \bar{y}_w is

$$V(\bar{y}_w) = (1-f) \frac{S^2}{n} + \frac{k-1}{n} w_2 S_2^2,$$

where $S_2^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (y_i - \bar{Y}_2)^2$, $S^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2$, $k = \frac{n_2}{n'_2}$.

Proof : We have $\bar{y}_w = \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}'_2)$

$$E(\bar{y}_w) = E_1 E_2 (\bar{y}_w / n_1, n_2) = E_1 E_2 \left(\frac{n_1\bar{y}_1}{n} / n_1 \right) + E_1 E_2 \left(\frac{n_2\bar{y}'_2}{n} / n_2 \right).$$

But $E_1 E_2 \left(\frac{n_1\bar{y}_1}{n} / n_1 \right) = E_1 \left[\frac{n_1}{n} E_2 (\bar{y}_1 / n_1) \right] = E_1 \left(\frac{n_1}{n} \bar{y}_1 \right) = \frac{N_1}{N} \bar{Y}_1$

and $E_1 E_2 \left(\frac{n_2\bar{y}'_2}{n} / n_2 \right) = E_1 \left[\frac{n_2}{n} E (\bar{y}'_2 / n_2) \right] = E_1 \left(\frac{n_2}{n} \bar{y}_2 \right) = \frac{N_2}{N} \bar{Y}_2.$

$$\therefore E(\bar{y}_w) = \frac{N_1}{N} \bar{Y}_1 + \frac{N_2}{N} \bar{Y}_2 = \frac{1}{N} [N_1 \bar{Y}_1 + N_2 \bar{Y}_2] = \bar{Y}$$

$$V(\bar{y}_w) = V_1 E_2 (\bar{y}_w) + E_1 V_2 (\bar{y}_w) = V_1 (\bar{y}) + E_1 [V_2 (\bar{y}_w / n_1, n_2)]$$

But $V_1 (\bar{y}) = \frac{N-n}{Nn} S^2 = (1-f) \frac{S^2}{n}$

and $V_2(\bar{y}_w/n_1, n_2) = V_2\left(\frac{n_2 \bar{y}'_2/n_1}{n}\right) = \frac{n_2^2}{n^2} \left[\frac{1}{n'_2} - \frac{1}{n_2} \right] S_2^2 = \frac{n_2}{n^2} [k - 1] S_2^2.$

$\therefore E_1 V_2(\bar{y}_w/n_1, n_2) = \frac{k - 1}{n} E_1\left(\frac{n_2}{n} S_2^2\right) = \frac{k - 1}{n} \frac{N_2}{N} S_2^2.$

$\therefore V(\bar{y}_w) = [1 - f] \frac{S^2}{n} + \frac{k - 1}{n} \frac{N_2}{N} S_2^2 = (1 - f) \frac{S^2}{n} + \frac{k - 1}{n} w_2 S_2^2$
 $= (1 - f) \frac{S^2}{n}, \text{ if } k = \frac{n_2}{n'_2} = 1.$

Further information regarding the sample selection from non-response group and the estimate of population mean are studied by E1-Badry (1956) and Foradori (1961).

Hansen and Hurwitz have discussed the value of n'_2 . For this a cost function is considered, where the cost function is

$$C' = C_0 + Cn + C_1 n_1 + C_2 n'_2,$$

where C_0 is the overall cost of survey, C is the cost of inclusion of every unit in the first sample, C_1 is the cost of collection of data from n_1 units, and C_2 is the cost of collection of data from n'_2 units. The value of n'_2 is to be selected in such a way that the value of C' is minimum. Assume that the value of C' is different for different samples. Then a cost function of average cost of survey can be considered.

Given $C' = C_0 + Cn + C_1 n_1 + C_2 n'_2 = C_0 + n \left(C + C_1 \frac{n_1}{n} + C_2 \frac{n'_2}{n} \right)$

$$E(C') = C_0 + n \left(C + C_1 \frac{N_1}{N} + C_2 \frac{N_2}{n} \right) = C_0 + n \left(C + C_1 w_1 + C_2 \frac{w_2}{k} \right).$$

Let us consider that $V(\bar{y}_w) = V$ and a specific value of V is V_0 . Then the value of n is to be selected in such a way that for fixed cost V_0 the value of $E(C')$ is minimum.

Let $E(C') = C$. Then n and k are to be estimated in such a way that

$$\phi = C + \lambda(V - V_0)$$

becomes minimum. Here λ is the Lagrange's multiplier. By minimizing ϕ , we have

$$n(\text{opt}) = \frac{S^2 + (k - 1)w_2^2 S_2^2}{\left(V_0 + \frac{S^2}{N}\right)} \text{ and } k(\text{opt}) = \left[\frac{C_2(S^2 - w_2 S_2^2)}{(C + C_1 w_1) S_2^2} \right]^{\frac{1}{2}}.$$

If k (opt) is known, the value of n'_2 can be estimated.

In practice, the survey can be started by selecting a sample of size n (opt) without replacement. Then from the non-response units another sample of size n'_2 can be selected without replacement and these n'_2 units are to be investigated personally. To study more regarding adjustment of non-response Kish and Hess (1959) and Srinath (1971) can be discussed.

21.4 Call Backs and its Effects

The causes of non-response can be classified into 4 classes. These are (1) Non-coverage, (2) Not-at-homes, (3) Unable to answer, (4) The hard core. Due to these 4 causes of non-response and hence, non-sampling error creeps in the the estimate. The error due to non-response can be reduced by repeated visits to the unit. However, the error due to non-coverage

cannot be reduced by repeated visits. The technique of reduction of non-sampling error due to non-coverage has been discussed by Kish and Hess (1958), and Woolsey (1956). Stephan and McCarthy (1958), Durbin (1954), and Durbin and Stuart (1954) have proposed the technique of reduction of non-sampling error by repeated visits to the sampling unit. They have also discussed the costs involved in the repeated samples.

Deming (1953) has considered a model to examine the analytical results after repeated visit to the unit.

Let a population can be classified into r classes. The classification is done on the basis of probability of availability of a sampling unit at home.

Let w_{ij} = probability of availability of a unit of j -th class during i -th visit ($j = 1, 2, \dots, r$).

p_j = proportion of a population unit to be included in j -th class.

μ_j = mean of units of j -th class.

σ_j^2 = variance of units of j -th class.

\bar{y}_{ij} = the average of units of j -th class during i -th visit or before the i -th visit.

Let $w_{ij} > 0$ [for all values of j] and $E(\bar{y}_{ij}) = \mu_j$. Then the population mean of all units is

$$\bar{\mu} = \sum_j^r p_j \mu_j.$$

Let us now observe those sampling units which are not available during r visits. These units can be considered as the units in $(r + 1)$ -th class. Some of the units may be available in the first visit and data are collected from those units during first visit. The data may be collected during second visit from some of the units, during third visit from some of the units, and so on. If finite population correction is neglected, then the units of $(r + 1)$ -th class will follow multi-dimensional distribution. The value of this multi-dimensional variable is

$$w_{i1}p_1 + w_{i2}p_2 + \dots + w_{ir}p_r + \left(1 - \sum w_{ij}p_j\right) n_0,$$

where n_0 is the preliminary sample size.

Let us consider that the data are collect from n_i units during i -th visit. Then the distribution of n_i units will be binomial distribution, where n_0 units are observed and probability of success is $\sum w_{ij}p_j$. Therefore, $E(n_i) = n_0 \sum w_{ij}p_j$ = expected number of units visited during i -th visit.

For a fixed value of n_i , we have

$$E\left(\frac{n_{ij}}{n_i}\right) = \frac{n_i w_{ij} p_j}{\sum w_{ij} p_j}, \quad j = 1, 2, \dots, r.$$

Here n_{ij} follows multivariate distribution and its probability is

$$\frac{w_{ij} p_j}{\sum w_{ij} p_j}.$$

Let \bar{y}_i be the mean of sample observations observed during i -th visit, where

$$E(\bar{y}_i/n_i) = E\left[\frac{\sum n_{ij} \bar{y}_{ij}}{n_i}\right] = \frac{\sum_j n_i w_{ij} p_j \mu_j}{n_i \sum_j w_{ij} p_j} = \frac{\sum w_{ij} p_j \mu_j}{\sum w_{ij} p_j} = \bar{\mu}_i.$$

But this conditional expectation does not depend on the values of n_i and hence, $E(\bar{y}_i) = \bar{\mu}_i$.

The bias of \bar{y}_i is $(\bar{\mu}_i - \bar{\mu})$.

The conditional variance of \bar{y}_i for any value of n_i is

$$V(\bar{y}_i/n_i) = \frac{\sum_{j=1}^r w_{ij} p_j [\sigma_j^2 + (\mu_j - \bar{\mu}_i)^2]}{n_i \sum w_{ij} p_j}$$

If n_i is replaced by $E(n_i)$ and $1/n_i^2$ is considered negligible, then the conditional variance of \bar{y}_i is found out. Then after i -th visit the mean square error of \bar{y}_i is found out, where

$$\text{MSE}(\bar{y}_i/i) = V(\bar{y}_i/i) + (\bar{\mu}_i - \bar{\mu})^2.$$

21.5 Politz-Simmons Technique

Hansen and Hurwitz (1946) have proposed an adjustment in the estimator if there is error in the estimator due to non-response. Their suggestion is mainly applied when data are collected through mail questionnaire. But non-response may occur even in case of data collection by the investigator by face to face conversation. The sources of non-response are : (a) the sampling unit may be not-at-home during survey, (b) the sampling unit may not be identified, etc. The investigator may try repeatedly to collect information, because the estimator will be biased due to the non-availability of information from some sampling units. But Politz-Simmons (1946, 1950) have proposed a method by which the bias due to non-response can be adjusted without repeated investigation of the sampling units.

The Politz-Simmons method can be explained in the following way :

- (a) The sampling units are to be investigated during randomly selected time, specially at night.
- (b) Each unit is to be investigated once.
- (c) The information regarding the stay of sampling units at home before 5 days of survey date. This information is collected by any means of communications with the units.

Let us consider that the i -th unit was present at home $(r_i - 1)$ times before the survey. Then r_i is the random variable, the values of which are 1, 2, 3, 4, 5, 6. The proportion of i -th unit to be present at home during the period is $r_i/6$. At this stage the value of the study variable is to be collected for every segment of 6 segments of the time and the total value of each segment is to be weighted by $6/r_i$ to estimate the population total.

Let there be N units in the population. Let us consider the value of the study variable of i -th unit is y_i ($i = 1, 2, \dots, N$) and the total value of y_i for the units who are present at home during j -time is Y_j . Here j -time is the time of collection of data. Then the estimator of population total is

$$\hat{Y} = 6 \sum_{j=1}^6 Y_j/j = 6 \sum_{i=1}^n \frac{y_i}{r_i} = 6 \sum_{i=1}^N \frac{\delta_i y_i}{r_i},$$

where $\delta_i = 1$, if i -th unit is found at home
 $= 0$, otherwise.

Let the probability of availability of i -th unit during evening is p_i . Then

$$E(\hat{Y}) = 6 \sum_{i=1}^N y_i E\left(\frac{\delta_i}{r_i}\right).$$

$$\text{Again, } E\left(\frac{\delta_i}{r_i}\right) = E\left[\delta_i E\left(\frac{1}{r_i/\delta_i}\right)\right] = P(\delta_i = 1)E\left[\frac{1}{r_i/\delta_i} = 1\right].$$

Now, if i -th unit is at home, then the probability of $r_i = h(1, 2, \dots, 6)$ is equal to the probability of that i -th unit is at home in preceding $(h-1)$ times. Then

$$E\left(\frac{\delta_i}{r_i}\right) = p_i \sum_{h=1}^6 \frac{1}{h} \binom{5}{h-1} p_i^{h-1} (1-p_i)^{5-(h-1)} = \frac{1}{6} [1 - (1-p_i)^6].$$

Replacing the value of $E\left(\frac{\delta_i}{r_i}\right)$ in $E(\hat{Y})$, we have

$$E(\hat{Y}) = \sum_{i=1}^N y_i [1 - (1-p_i)^6] = \sum_{i=1}^N y_i - \sum_{i=1}^N y_i (1-p_i)^6.$$

It indicates that \hat{Y} is not unbiased. The bias will be less, if the units which are not-at-home are very small in number. This indicates that the survey should be conducted during a time in which most of the units are available at home.

The mean square error of \hat{Y} is

$$\text{MSE}(\hat{Y}) = V(\hat{Y}) + (\text{Bias})^2 = E(\hat{Y}^2) - \left[\sum_{i=1}^N y_i \{1 - (1-p_i)^6\} \right]^2 + \left\{ \sum_{i=1}^N y_i [1 - (1-p_i)^6] \right\}^2.$$

$$E(\hat{Y}^2) = E\left\{ 6 \sum \frac{\delta_i y_i}{r_i} \right\}^2 = 36 \left\{ \sum_{i=1}^N y_i^2 E\left(\frac{\delta_i}{r_i}\right)^2 + \sum_{i \neq k=1}^N y_i y_k E\left(\frac{\delta_i}{r_i} \frac{\delta_k}{r_k}\right) \right\}.$$

$$\text{Again, } E\left(\frac{\delta_i}{r_i}\right)^2 = E\left\{ \delta_i^2 E\left(\frac{1}{r_i/\delta_i^2}\right) \right\} = p_i E\left[\frac{1}{r_i^2/\delta_i} = 1\right] = \frac{1}{6} \sum_{h=1}^6 \frac{1}{h} \binom{6}{h} p_i^h (1-p_i)^{6-h}.$$

$$E\left[\frac{\delta_i}{r_i} \frac{\delta_k}{r_k}\right] = E\left(\frac{\delta_i}{r_i}\right) E\left(\frac{\delta_k}{r_k}\right) \text{ as } i \neq k \text{ and hence, } \frac{\delta_i}{r_i} \text{ and } \frac{\delta_k}{r_k} \text{ are independent.}$$

$$\therefore E\left[\frac{\delta_i}{r_i} \frac{\delta_k}{r_k}\right] = \frac{1}{36} [1 - (1-p_i)^6][1 - (1-p_k)^6].$$

$$\text{Then } E(\hat{Y}^2) = 6 \sum_{i=1}^N y_i^2 \sum_{h=1}^6 \frac{1}{h} \binom{6}{h} p_i^h (1-p_i)^{6-h} + \sum_{i \neq k=1}^N y_i y_k [1 - (1-p_i)^6][1 - (1-p_k)^6].$$

$$\therefore \text{MSE}(\hat{Y}) = \sum_{i=1}^N y_i^2 \left\{ 6 \sum_{h=1}^6 \frac{1}{h} \binom{6}{h} p_i^h (1-p_i)^{6-h} - [1 - (1-p_i)^6]^2 \right\} + \left[\sum_{i=1}^N y_i (1-p_i)^6 \right]^2.$$

If the bias of \hat{Y} is negligible, then

$$\text{MSE}(\hat{Y}) = V(\hat{Y}) = \sum_{i=1}^N y_i^2 \left\{ \sum_{h=1}^6 \frac{1}{h} \binom{6}{h} p_i^h (1-p_i)^{6-h} - [1 - (1-p_i)^6]^2 \right\}.$$

The unbiased estimate of MSE (\hat{Y}) is also available. For this, let us consider that out of 6 times j times a unit is at home. Let the number of such units be n_j . Let the value of i -th unit of j -th group be y_{ij} . Then the estimator of P_i of each unit of j -th group is $\hat{P}_j = j/6; j = 1, 2, \dots, 6$. Then the estimator of MSE (\hat{Y}) is

$$v(\hat{Y}) = \sum_{j=1}^6 A_j \sum_{i=1}^{n_j} y_{ij}^2,$$

$$\text{where } A_j = \frac{6}{j} \left[6 \sum_{h=1}^6 \frac{1}{h} \binom{6}{h} p_i^h (1-p_i)^{6-h} - \{1 - (1-p_i)^6\}^2 \right].$$

This Politz-Simmons technique is fruitful if repeated visits are not possible or repeated visits are not justified. Moreover, the technique is suitable for a situation where the study variable value is changed over time. The technique will be less efficient if all the units can be visited and information for all units are available during main survey.

21.6 Response Errors

The data are collected, usually, by the investigator by face to face conversation. In some cases the information are provided by the sampling units himself, specially in case of mail questionnaire. If sampling unit is not human being, information are collected by the investigator. Whatever be the mode of collection of information there may be a chance of misreporting of the information. The misreported information or distorted information are the sources of response errors. The sources can be classified as follows :

(a) Wrong information deliberately provided by the respondent : This case arises if data are related to income and expenditure of the respondent. During family budget inquiry the response error arises. In opinion poll, specially in political opinion survey, the sampling unit tries to avoid the answer which may go against the government or against the influential person.

(b) Difficulty in understanding questionnaire : Some questions in the questionnaire may not be understood by the respondents or the respondents may not know the exact answer to some questions. In such cases, answer may be avoided or misinformation may be provided. The proper answer to the question regarding health may not be known to a respondent. He may give distorted answer.

(c) Error due to investigator : An investigator may not collect the information properly, if he is not well trained for conducting the survey. Distorted information may be recorded due to misunderstanding of measuring devices. Question may be asked wrongly or distorted answer may be recorded deliberately. In case of question related to opinion survey, the investigator may try to influence the respondent according to his/her belief. The investigator may record the information by himself if a sampling unit is not available during survey.

Beside these, there are many other sources of errors. Whatever be the type of error or whatever be the direction of error, it affects the estimate of the population parameter. Thus, during analysis the errors should not be avoided. Modified analytical techniques should be applied to reduce the impacts of the errors. Different authors have discussed the sources of errors and they have proposed the analytical techniques to handle such errors. The important works in this context are due to Deming (1944), Marks and Mouldin (1950), Marks et al. (1953), Mahalanobis (1944), Sukhatme and Seth (1952), Hansen et al. (1951, 1953, 1961, 1964), Kish

and Lansing (1954) and Madow (1965). In this section the estimation procedure in presence of response error will be discussed.

Let there be N units in a population and $n(\leq N)$ units are selected by simple random sampling scheme to estimate the parameter of the population. Let us assume that the exact value of the study variable of i -th unit be $x_i (i = 1, 2, \dots, N)$. Also, assume that the collected information of the study variable of i -th unit by j -th investigator during k -th visit be $y_{ijk} (i = 1, 2, \dots, n; j = 1, 2, \dots, m \text{ and } k = 0, 1, 2, \dots, n_{ij})$. There are m investigator and j -th investigators have collected information l_{ij} times from i -th sampling unit.

Here the difference between x_i and y_{ijk} is the error due to response. In the error, there may be the impact of investigator as well as the interaction of investigator and sampling unit. The error may be due to some other random causes. Thus, the value y_{ijk} can be expressed as a linear function of the impacts of sources of errors, where the model is :

$$y_{ijk} = x_i + \alpha_j + \beta_{ij} + e_{ijk},$$

where x_i = real value of study variable of i -th unit,

α_j = impact of j -th investigator,

β_{ij} = interaction of j -th investigator with i -th unit,

e_{ijk} = random error.

Assume that $E[e_{ijk}/i, j] = 0$ and $E[\beta_{ij}/j] = 0$.

In most cases, the data are collected once. For such type of data the model is

$$y_{ij} = x_i + \alpha_j + \epsilon_{ij}.$$

Assume that $E[\epsilon_{ij}] = 0$

and $E[\epsilon_{ij}, \epsilon_{i'j'}] = 0$, if $i \neq i', j \neq j'$
 $= S_\epsilon^2$, if $i = i', j = j'$.

Let us discuss the analysis of data using the latter model. For analytical process of former model the works of Sukhatme and Seth (1952), Zasepa (1961) and Hansen et al. (1953, 1961, 1964) may be studied.

To estimate the population mean let us consider the calculation of sample mean based on latter model. For this, let

$$l_{ij} = 0 \text{ or } 1, l_{i.} = \sum_{j=1}^m l_{ij}, l_{.j} = \sum_{i=1}^n l_{ij}, l_{..} = \sum_{i=1}^n \sum_{j=1}^m l_{ij}$$

$\bar{y}_{.j}$ = mean of $l_{.j}$ observation collected by j -th investigator.

$\bar{y}_{..}$ = mean of $l_{..}$ observations.

Then $\bar{y}_{.j}$ can be written as

$$\bar{y}_{.j} = \frac{1}{l_{.j}} \sum_{i=1}^n x_i l_{ij} + \alpha_j + \frac{1}{l_{.j}} \sum_{i=1}^n \epsilon_{ij} l_{ij}$$

and $\bar{y}_{..} = \frac{1}{l_{..}} \sum_{i=1}^n x_i l_{i.} + \frac{1}{l_{..}} \sum_{j=1}^m \alpha_j l_{.j} + \frac{1}{l_{..}} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij} l_{ij}$.

Let us investigate whether these two sample means are the unbiased estimators of the population mean. Let us assume that there are M investigators in the population. All of them

may be employed for the survey or $m(\leq M)$ of them can be selected at random for the survey. Again, the sampling units can be allocated to the randomly selected investigator at random. Let us assume that

- (i) m investigators are selected at random from M investigators,
- (ii) the sampling units are allocated at random to the investigators,
- (iii) every investigator has collected information from same number of sampling units, i.e.,

$$l_i = \frac{l_{..}}{n} = p \text{ (say).}$$

Thus, we can have

$$\bar{y}_{.j} = \frac{1}{\bar{l}} \sum_{i=1}^n x_i l_{ij} + \alpha_j + \frac{1}{\bar{l}} \sum_{i=1}^n \epsilon_{ij} l_{ij}, \text{ where } \bar{l} = \frac{l_{..}}{nm}$$

$$\text{and } \bar{y}_{..} = \frac{1}{n} \sum_{i=1}^n x_i + \frac{1}{m} \sum_{j=1}^m \alpha_j + \frac{1}{l_{..}} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij} l_{ij}.$$

$$\text{Then } E(\bar{y}_{.j}) = \frac{1}{N} \sum_{i=1}^N x_i + \frac{1}{M} \sum_{j=1}^M \alpha_j = \mu + \bar{\alpha}$$

$$\text{and } E(\bar{y}_{..}) = \frac{1}{N} \sum_{i=1}^N x_i + \frac{1}{M} \sum_{j=1}^M \alpha_j = \mu + \bar{\alpha}.$$

Here μ is the population mean and $\bar{\alpha}$ is population mean of the bias of investigator.

It is seen that the sample mean $\bar{y}_{..}$ is not the unbiased estimator of the population mean. The estimator will be unbiased if $\bar{\alpha} = 0$. This will be zero, if the response error is random. That means on addition of positive and negative values of α_j we shall get the sum as zero. However, even if α_j is random its sum may not be zero. Hence, $\bar{y}_{..}$ is not unbiased. The estimator is to be found out so that $\bar{\alpha}$ becomes minimum.

Let us now calculate the variance of the estimator. By definition, we have

$$\begin{aligned} V(\bar{y}_{.j}) &= E[\bar{y}_{.j} - E(\bar{y}_{.j})]^2 = E \left[\frac{1}{\bar{l}} \sum_{i=1}^n x_i l_{ij} - \mu + \alpha_j - \bar{\alpha} + \frac{1}{\bar{l}} \sum_{i=1}^n \epsilon_{ij} l_{ij} \right]^2 \\ &= E \left\{ \frac{1}{\bar{l}} \sum_{i=1}^n x_i l_{ij} - \mu \right\}^2 + E(\alpha_j - \bar{\alpha})^2 + E \left\{ \frac{1}{\bar{l}} \sum_{i=1}^n \epsilon_{ij} l_{ij} \right\}^2. \end{aligned}$$

$$\because E[\text{cross-product term}] = 0$$

The mean $\frac{1}{\bar{l}} \sum x_i l_{ij}$ is the mean of \bar{l} units and these \bar{l} units are randomly investigated by j -th investigator. Again, \bar{l} units are selected from n units and hence, these \bar{l} units can be considered as simple random sample unit selected from N units. Then

$$E \left\{ \frac{1}{\bar{l}} \sum_{i=1}^n x_i l_{ij} - \mu \right\}^2 = \left(\frac{1}{\bar{l}} - \frac{1}{N} \right) S_x^2, \text{ where } S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2.$$

It has already been assumed that the j -th investigator is selected at random from M investigators and α_j is the bias of j -th investigator. Therefore, the variance of the random variable α_j is

$$E(\alpha_j - \bar{\alpha})^2 = \left(1 - \frac{1}{M}\right) S_\alpha^2, \quad \text{where } S_\alpha^2 = \frac{1}{M-1} \sum_{j=1}^M (\alpha_j - \bar{\alpha})^2.$$

Again, ϵ_{ij} is the error of data collection recorded from \bar{l} units investigated by j -th investigator and mean of ϵ_{ij} is

$$\frac{1}{\bar{l}} \sum_i \epsilon_{ij} l_{ij}.$$

It is known that $E(\epsilon_{ij}) = 0$ and $V(\epsilon_{ij}) = S_\epsilon^2$.

Hence,
$$E \left\{ \frac{1}{\bar{l}} \sum_{i=1}^n \epsilon_{ij} l_{ij} \right\}^2 = \frac{S_\epsilon^2}{\bar{l}}$$

and
$$V(\bar{y}_{.j}) = \frac{N - \bar{l}}{N\bar{l}} S_x^2 + \frac{M-1}{M} S_\alpha^2 + \frac{S_\epsilon^2}{\bar{l}}.$$

If M and N are big, $\frac{1}{M}$ and $\frac{1}{N}$ may be neglected.

Then
$$V(\bar{y}_{.j}) = \frac{1}{\bar{l}} (S_x^2 + S_\epsilon^2) + S_\alpha^2.$$

Similarly, $V(\bar{y}_{..}) = E[\bar{y}_{..} - E(\bar{y}_{..})]^2$

$$\begin{aligned} &= E \left[\frac{1}{n} \sum x_i - \mu + \frac{1}{m} \sum \alpha_j - \bar{\alpha} + \frac{1}{\bar{l}} \sum_i^n \sum_j^m \epsilon_{ij} l_{ij} \right]^2 \\ &= E \left\{ \frac{1}{n} \sum_i^n x_i - \mu \right\}^2 + E \left\{ \frac{1}{m} \sum_j^m \alpha_j - \bar{\alpha} \right\}^2 + E \left\{ \frac{1}{\bar{l}} \sum \sum \epsilon_{ij} l_{ij} \right\}^2. \end{aligned}$$

But x_1, x_2, \dots, x_n are the values of the variable observed from n sampling units selected from N units by simple random sampling method. So,

$$E \left\{ \frac{1}{n} \sum x_i - \mu \right\}^2 = \frac{N-n}{Nm} S_x^2.$$

It has already been mentioned that m investigators are selected from M investigators at random and α_j is the bias of j -th investigator. Therefore,

$$E \left\{ \frac{1}{m} \sum_{j=1}^m (\alpha_j - \bar{\alpha}) \right\}^2 = \frac{M-m}{mM} S_\alpha^2.$$

Again, $\frac{1}{\bar{l}} \sum_i^n \sum_j^m \epsilon_{ij} l_{ij}$ is the mean of $l = m\bar{l}$ observations ϵ_{ij} .

Therefore,
$$E \left\{ \frac{1}{l} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij} l_{ij} \right\}^2 = \frac{S_{\epsilon}^2}{l}.$$

$$\therefore V(\bar{y}_{..}) = \frac{N-n}{Nn} S_x^2 + \frac{M-m}{mM} S_{\alpha}^2 + \frac{S_{\epsilon}^2}{l}.$$

Neglecting $\frac{1}{N}$ and $\frac{1}{M}$, we have

$$V(\bar{y}_{..}) = \frac{S_x^2}{n} + \frac{S_{\alpha}^2}{m} + \frac{S_{\epsilon}^2}{l}$$

$$V(\bar{y}_{..}) = \frac{S_x^2 + S_{\epsilon}^2}{n} + \frac{S_{\alpha}^2}{m}, \text{ if } p = 1 \text{ and } l = np = n.$$

It can also be written as

$$V(\bar{y}_{..}) = \frac{S_x^2 + S_{\alpha}^2 + S_{\epsilon}^2}{n} + S_{\alpha}^2 \left(\frac{1}{m} - \frac{1}{n} \right) = \frac{S_y^2}{n} + S_{\alpha}^2 \left(\frac{1}{m} - \frac{1}{n} \right).$$

This result is true if population is infinite and M is big and data are collected from any unit once. In that case,

$$S_y^2 = S_x^2 + S_{\alpha}^2 + S_{\epsilon}^2.$$

Hansen et al. (1951) have given another form of $V(\bar{y}_{..})$, where

$$V(\bar{y}_{..}) = \frac{S_y^2}{n} \left\{ 1 + \rho \left(\frac{n}{m} - 1 \right) \right\}.$$

Here ρ is the correlation coefficient of data collected from different units by the same investigator. If N and M are sufficiently large,

$$\rho S_y^2 = S_{\alpha}^2.$$

The above $V(\bar{y}_{..})$ formula indicates that, if the number of investigators are less, then any investigator will collect data from more units. In that case, even if ρ is small, the $V(\bar{y}_{..})$ will be increased. If $\rho = 0$, then

$$V(\bar{y}_{..}) = \frac{S_y^2}{n},$$

which is the usual variance of $\bar{y}_{..}$. Again, if α_j is same for all investigators or $m = n$, $\rho = 0$. But in practice, $\rho > 0$ as there is a tendency among the investigators to record high or low value of any characteristic.

The variance of the estimator $\bar{y}_{..}$ can be estimated by

$$v(\bar{y}) = \frac{s_u^2}{n} + \frac{n-m}{m-1} \frac{1}{n} (s_u^2 - s_{eo}^2).$$

Here
$$s_u^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y}_{..})^2$$

$$s_{eo}^2 = \frac{1}{l-m} \sum_{i=1}^l \sum_{j=1}^m (y_{ij} - \bar{y}_{.j})^2.$$

Example 21.1 : In a police station area there are 1000 jute growing farmers. Thirty farmers are selected by simple random sampling scheme. The data on production of jute from the selected farmers are collected by two investigators. The data on jute production (y in quintal) of selected farmers are collected by asking questions to the head of household and are given below. Estimate the average production of jute per farmer and also estimate the variance of your estimator.

Sl. No.	Jute production y quintal		Sl. No.	Jute production y quintal	
	1st investigator y_{i1}	2nd investigator y_{i2}		1st investigator y_{i1}	2nd investigator y_{i2}
01	10.5	10.0	16	30.0	32.0
02	12.8	13.0	17	22.0	20.0
03	15.0	15.8	18	24.0	26.0
04	20.5	19.0	19	18.5	20.0
05	21.5	20.5	20	16.7	16.0
06	16.0	17.0	21	12.0	13.5
07	25.0	24.0	22	11.5	11.0
08	14.8	14.0	23	5.8	6.8
09	10.5	9.0	24	20.0	22.0
10	15.0	16.0	25	24.0	25.0
11	18.0	18.5	26	20.0	19.0
12	20.5	20.0	27	10.2	10.5
13	25.8	25.0	28	9.5	8.0
14	10.0	9.5	29	8.0	4.0
15	11.8	11.0	30	12.5	14.0
Total	$y_{.1} = 247.7$	$y_{.2} = 242.3$	Total	$y_{.1} = 244.7$	$y_{.2} = 247.8$

Solution : Here $N = 1000$, $n = 30$, $m = 2$, $l_{ij} = 1$.

Total number of observations, $l_{..} = \sum_{i=1}^n \sum_{j=1}^2 l_{ij} = 60 = l$,

$$l_{i.} = \sum_{j=1}^2 l_{ij} = 2, \quad l_{.j} = \sum_{i=1}^{30} l_{ij} = 30, \quad \bar{l} = \frac{l_{..}}{n \times m} = \frac{60}{30 \times 2} = 1.$$

Total of observations, $\sum_{i=1}^{30} \sum_{j=1}^2 y_{ij} = 982.6 = y_{...}$

$$y_{.1} = \sum_{i=1}^{30} y_{i1} = 492.4, \quad y_{.2} = \sum_{i=1}^{30} y_{i2} = 490.2.$$

$$y_{i.} = y_{i1} + y_{i2}, \quad \bar{y}_{i.} = \frac{1}{l_{i.}} \times y_{i.}, \quad i = 1, 2, \dots, 30.$$

y_i	20.5	25.8	30.8	39.5	42.0	33.0	49.0	28.8	19.5	31.0	36.5	40.5
\bar{y}_i	10.25	12.9	15.4	19.75	21.0	16.5	24.5	14.4	9.75	15.5	18.25	20.25
y_i	50.8	19.5	22.8	62.0	42.0	50.0	38.5	32.7	25.5	22.5	12.6	42.0
\bar{y}_i	25.4	9.75	11.4	31.0	21.0	25.0	19.25	16.35	12.75	11.25	6.3	21.0
y_i	49.0	39.0	20.7	17.5	12.0	26.5	24.5	19.5	10.35	8.75	6.0	13.25

Assume that $m = 2$ investigators are selected from M investigators, where M is large. Then, the average production of jute per farmer is

$$\bar{y}_{..} = \frac{1}{l..} \sum_{i=1}^{30} \sum_{j=1}^2 y_{ij} = \frac{982.6}{60} = 16.38.$$

$$\text{Total sum of squares} = \sum \sum y_{ij}^2 - \frac{y_{..}^2}{l..} = 17989 - \frac{(982.6)^2}{60} = 1897.287.$$

$$\text{Sum of squares (investigators)} = \frac{\sum y_j^2}{n} - \frac{y_{..}^2}{l..} = \frac{482753.8}{30} - \frac{(982.6)^2}{60} = 0.0807.$$

$$s_u^2 = \frac{1}{n-1} \left[\sum \bar{y}_i^2 - n \bar{y}_{..}^2 \right] = \frac{1}{30} [9158.4625 - 30 \times (16.38)^2] = 1109.3305.$$

$$s_{eo}^2 = \frac{1}{l-m} \left[\sum \sum y_{ij}^2 - l \sum \bar{y}_j^2 \right], \bar{y}_1 = 15.41, \bar{y}_2 = 16.24$$

$$= \frac{1}{60-2} [17989 - 1 \times 536.2837] = 300.9089.$$

Therefore, the estimate of $V(\bar{y}_{..})$ is

$$v(\bar{y}_{..}) = \frac{s_u^2}{n} + \frac{n-m}{m-1} \frac{1}{n} (s_u^2 - s_{eo}^2) = \frac{1109.3305}{30} + \frac{30-2}{2-1} \times \frac{1}{30} (1109.3305 - 300.9089)$$

$$= 36.9768 + 754.5268 = 791.5036.$$

21.7 Determination of Optimum Number of Investigators

The variance formula of $\bar{y}_{..}$ indicates that, if a large number of investigators are employed, the $V(\bar{y}_{..})$ will be small. But, in practice, there are some problems if a large number of investigators are employed. The cost of the survey will be increased with the increase in number of investigators. There will be possibility of employment of inefficient investigators when its number will be increased and due to inefficient investigators there will be chances of higher response error leading to inefficient estimator. Thus, the number of investigators is to be employed in such a way that the cost of the survey will be fixed but variance of the estimator will be minimum.

Let there be a simple random sample of n units and the sample units are allocated randomly to m investigators. Let us consider that cost function of the survey is

$$C = C_1 n + C_2 m + C_3 \sqrt{nm},$$

where C_1 = cost of data collection from each unit,

C_2 = cost of employment of an investigator,

C_3 = proportional cost to the cost of travel to each unit.

Let us consider that the cost of survey is fixed and it is C_0 . The values of n and m are to be found out so that for fixed cost the variance of the estimator is minimum. Let ϕ be a function which indicates the minimum variance of \bar{y} , subject to fixed cost, where

$$\phi = \frac{S_y^2}{n} + S_\alpha^2 \left[\frac{1}{m} - \frac{1}{n} \right] + \lambda [C_1 n + C_2 m + C_3 \sqrt{nm} - C_0].$$

Here λ is the Lagrange's multiplier. Now, the values of λ , n and m are to be found out from the equations :

$$\frac{\partial \phi}{\partial \lambda} = 0, \quad \frac{\partial \phi}{\partial n} = 0 \quad \text{and} \quad \frac{\partial \phi}{\partial m} = 0.$$

From the solution of these 3 equations, we have

$$\frac{m^2}{n^2} + \frac{C_3}{2C_2} \left(\frac{m}{n} \right)^{3/2} - \frac{C_3}{2C_2} \frac{S_\alpha^2}{(S_y^2 - S_\alpha^2)} - \frac{C_1 S_\alpha^2}{C_2 (S_y^2 - S_\alpha^2)} = 0.$$

This equation has one positive real root and one negative real root. Moreover, it is difficult to have a simple formula of root. As a result, we have to find the root by trial and error method. Accordingly, if $C_1 = 0$, we have

$$\left(\frac{m}{n} \right)^{3/2} + \frac{C_3}{2C_1} \frac{m}{m} - \frac{C_3}{2C_1} \frac{S_\alpha^2}{(S_y^2 - S_\alpha^2)} = 0.$$

At this stage also it is difficult to have simple formula for m/n .

Again, let $C_3 = 0$, then we have

$$\frac{m^2}{n^2} = \frac{C_1}{C_2} \frac{S_\alpha^2}{S_y^2 - S_\alpha^2}.$$

Let $C_0 = C_1 n + C_2 m$, then

$$n = \frac{C_0}{C_1 + C_2 \sqrt{\frac{C_1}{C_2} \frac{S_\alpha^2}{S_y^2 - S_\alpha^2}}} \quad \text{and} \quad m = C_0 \sqrt{\frac{C_1}{C_2} \frac{S_\alpha^2}{S_y^2 - S_\alpha^2}}.$$

This last result indicates that the number of investigators will be increased with the increase in S_α^2 compared to S_y^2 . Again, if number of investigators is increased, the response error is also increased and the value of α will be increased.

Exercise

1. Distinguish population and sample. What is meant by sampling? Discuss, in short, the different sampling techniques.
2. What is sampling? Write down its advantages, disadvantages and uses.
3. Distinguish between census and sample survey. Explain the sources of errors in sample survey.
4. What is sample survey? What is the objective of sample survey? Explain the sources of errors in sample survey.
5. Define population, sample, frame, sampling error and non-sampling error.

Prepare a questionnaire to conduct the following survey :

- (i) To study the causes and effects of dropout in primary education.
- (ii) To study the impacts and reasons of non-adopting family planning methods.

- (iii) To estimate the number of diabetic and heart patients in a rural area.
- (iv) To study the housing problem in rural area.
- (v) To study the socioeconomic condition of industrial workers in an industrial belt.

6. Distinguish between

- (a) Precision and efficiency of estimator
- (b) Biased and unbiased estimators
- (c) Standard error and standard deviation.

What is meant by survey? How does sample survey differ from census? Explain the merits and demerits of sample survey.

7. Write down the objectives of the sample survey. Discuss the principles of sample survey. Explain the sources of errors in sample survey.
8. What do you mean by simple random sampling? How does it differ from random sampling. Explain the method of selection of a simple random sample.

Show that sample mean from a simple random sample is an unbiased estimator of population mean. Find the variance of your estimator.

9. Explain simple random sampling with and without replacement with examples. Find an unbiased estimator of variance of simple random sample mean.

The following data represent the number of ever-born children of some couples in a rural area.

Select a simple random sample of 25% couples and estimate the average ever-born children per couple. Also find 95% confidence interval for the average ever-born children per couple.

Sl.No. of couple	Ever born children	Sl.No. of couple	Ever born children	Sl.No. of couple	Ever born children
1	5	21	6	41	6
2	3	22	7	42	2
3	0	23	4	43	2
4	4	24	5	44	1
5	2	25	2	45	1
6	2	26	3	46	2
7	1	27	3	47	2
8	2	28	0	48	3
9	4	29	1	49	4
10	3	30	2	50	4
11	0	31	3	51	2
12	1	32	2	52	3
13	1	33	2	53	4
14	2	34	4	54	4
15	5	35	5	55	2
16	4	36	4	56	4
17	3	37	3	57	6
18	3	38	3	58	5
19	2	39	4	59	2
20	2	40	5	60	3

10. What is meant by simple random sampling? Explain the method of estimation of population proportion along with its estimated standard error.

How would you decide about the sample size to estimate a parameter for a pre-determined precision?

In a sample survey on family planning activities it is observed that 55% out of 500 ever married couples have adopted at least one of the family planning methods. Find 95% confidence interval for the population proportion of adopter couples.

If the population size $N = 10000$, find the sample size to estimate population proportion of adopter couples with 95% confidence and 5% precision.

11. Explain the advantages and disadvantages of simple random sampling. Distinguish between random sampling and simple random sampling. How does random sampling differ from judgement sampling? Let there be N units in a population, two observations of which are y_1 and y_2 . Let a sample of size n be selected without replacement from the population apart from these two observations. Consider that $y_1 + y_2 + N\bar{y}'$ is the estimator of population total and $N\bar{y}$ is the usual estimator of population total. Compare the efficiency of these two estimators.

In an area there are $N = 1000$ farmers involved in white revolution. A simple random sample of $n = 25$ has been selected and milk production (y kg) per day of these farmers are recorded and given below :

y : 15.5, 12.8, 15.0, 20.0, 10.5, 16.0, 14.0, 10.0, 10.0, 11.0, 12.4, 13.0, 20.0, 24.0, 8.5, 12.0, 15.0, 16.0, 17.5, 20.0, 16.5, 17.8, 11.5, 16.2, 18.5.

Estimate total milk production of the farmers in the study area. Find 95% confidence interval for the total milk production.

12. Define simple random sampling. Write down the different steps to select a simple random sample of size n to estimate the socioeconomic condition of people living in a slum area.

Show that in case of simple random sampling sample variance is an unbiased estimator of population variance. Also show that for a simple random sample of size n ,

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

13. Define stratified random sampling. How does it differ from simple random sampling? Explain the necessity of stratified sampling.

What is meant by allocation problem in stratified sampling? What are the different methods of allocation of sample sizes?

Show that, for a stratified sample the variance of stratified random sample mean is minimum, if $n_h \propto N_h S_h$. Also find the minimum variance of \bar{y}_{st} .

14. Write down the advantages and disadvantages of stratified sampling. Explain a practical situation where stratified random sampling can be applied profitably.

For stratified random sampling when $\bar{y} = \frac{1}{n} \sum_h n_h \bar{y}_h$ becomes an unbiased estimator of population mean, find the variance of this estimator.

Show that stratified random sampling is more efficient than simple random sampling.

15. Discuss the methods of deciding strata and strata sizes. Find the effect of error in strata sizes. Show that, for a fixed total cost of survey $V(\bar{y}_{st})$ is minimum, if $n_h \propto N_h S_h / \sqrt{C_h}$. Find the formula to find n . Also find the minimum variance of \bar{y}_{st} .
16. Discuss the necessity of stratified random sampling. Show that, if $1/N_h$ is negligible, then

$$V_{\text{ran}} \geq V_{\text{prop}} \geq V_{\text{opt}}$$

Distinguish between quota sampling and stratification after sampling.

17. Discuss the method of construction of strata. Suggest a stratified random sampling scheme to estimate the total fish production in an area. Also suggest the estimator along with its sampling variance.
18. What do you mean by stratified random sampling? Explain the practical situation where we need stratified random sampling. Show that stratified sampling is more precise than simple random sampling.

What is meant by allocation problem? Explain different methods of allocation.

19. Define strata, stratified sampling. Suggest estimator of population total in case of stratified sampling. Find sampling variance of your estimator.

What is meant by optimum allocation? Find the variance of \bar{y}_{st} in case of optimum allocation.

20. To estimate the proportion of adopter females in a district a survey is conducted. The survey result is shown below :

Level of education of female	No. of couples in the area	Sample number of couples	Proportion of adopter couples in the sample
Illiterate	640	150	0.35
Primary	648	145	0.48
Secondary	344	92	0.45
Higher	198	46	0.58

Estimate the proportion of adopter couples in the area. Also estimate the standard error of your estimator. Find the values of n_h for Neyman allocation using sample information. Do you think that the estimator will be affected due to the allocation of n_h using sample information?

21. What are the different methods of estimation. Compare each method of estimation in case of simple random sampling.
22. What do you mean by ratio and regression methods of estimation? Under what conditions a ratio estimator will be unbiased? Find the relative bias of ratio estimator.
23. Define regression estimator. How does it differ from ratio estimator? Show that regression estimator is more efficient than ratio estimator.
24. What is the difference between unbiased ratio estimator and unbiased type ratio estimator? Suggest an unbiased ratio estimator of population total. How would you estimate the standard error of your estimator?
25. The variables y and x are related by $y_i = \alpha + \beta x_i + \epsilon_i$.

Find an estimator of population total assuming the above linear model of y and x . Also suggest an estimator of variance of your estimator.

26. If \hat{R} is distributed normally for large sample size, find $V(\hat{R})$.

If in a simple random sample the value of b is known, then

$$V(\bar{y}_{1r}) = \frac{1-f}{n} \frac{1}{N-1} \sum \{(y_i - \bar{y}) - b(x_i - \bar{x})\}^2, \text{ where } \bar{y}_{1r} = \bar{y} + b(\bar{X} - \bar{x}).$$

27. Discuss the importance of using auxiliary variable in estimating parameter. Show that ratio estimator is BLUE. Also show that ratio estimator is more precise than simple estimator.
28. In an administrative unit, there are 150 villages. Thirty villages are randomly selected from these 150 villages. The number of cows (y) in the selected villages are recorded. The number of cows in those villages during agriculture census (x) are also recorded. The information are given in the next page :

Sl.No. of village	x	y	Sl.No. of village	x	y	Sl.No. of village	x	y
01	400	415	11	125	120	21	900	905
02	180	202	12	240	230	22	870	950
03	275	320	13	180	201	23	660	650
04	240	220	14	95	90	24	225	228
05	126	150	15	115	108	25	278	315
06	97	120	16	120	128	26	570	460
07	140	125	17	172	168	27	300	350
08	275	270	18	290	295	28	230	260
09	180	210	19	315	310	29	500	525
10	200	200	20	440	460	30	610	580

The total number of cows in the study area during agriculture census was recorded as 60000. Estimate the total number of cows in the study area. Find 95% confidence interval for your estimator.

29. Distinguish between regression and ratio estimator. Find bias and mean square error of ratio estimator. Under what conditions ratio estimator becomes unbiased?

Estimate total number of cows by regression method of estimation. Use data of Example 28. Compare ratio estimator and regression estimator.

30. Give a comparative study of ratio estimator, regression estimator and simple estimator. Prove that, if population regression coefficient B is known, then regression estimator is unbiased and variance of this estimator becomes minimum.
31. What is meant by systematic sampling? Write down the merits and demerits of this sampling scheme. Show that, systematic sampling is more precise than simple random sampling and stratified random sampling.
32. Explain the method of selecting systematic sample. Show that, stratified random sampling is more precise than systematic sample when there is a linear trend in population observations.
33. Discuss the method of selecting systematic sample from different populations. How would you estimate the variance of systematic sample mean?

Show that, if $S_w^2 > S^2$, systematic sampling is more precise than simple random sampling.

34. Explain linear and circular systematic sampling. Give a comparative study of these two sampling schemes.

$$\text{Show that } \text{Cov}(\bar{y}_n, \bar{x}_n) = \frac{N-n}{nN} \frac{1}{N-1} \sum (x_i - \bar{x}_n)(\bar{y}_i - \bar{y}_n),$$

where \bar{x}_n and \bar{y}_n are the sample means of n observations selected from N units of the variables x and y , respectively.

35. The following data represent the amount of cultivable land (x hectare) and amount of land for jute (y hectare) cultivation in some villages.

Sl.No. of village	x	y	Sl.No. of village	x	y	Sl.No. of village	x	y	Sl.No. of village	x	y
01	1600	218	16	875	115	31	672	220	46	603	121
02	1508	108	17	1348	546	32	812	183	47	882	208
03	592	12	18	1550	212	33	970	178	48	472	115
04	2001	316	19	975	178	34	664	172	49	1692	442
05	508	112	20	1664	352	35	360	82	50	1308	348
06	2408	15	21	510	108	36	472	115	51	1550	550
07	618	211	22	692	207	37	698	220	52	1012	392
08	1542	312	23	428	12	38	890	165	53	888	144
09	651	178	24	1005	118	39	1507	312	54	907	268
10	750	121	25	572	119	40	1175	476	55	765	119
11	1548	225	26	648	212	41	225	52	56	908	178
12	1672	348	27	1565	365	42	365	88	57	555	126
13	1807	478	28	972	182	43	750	155	58	845	321
14	592	18	29	1015	362	44	872	178	59	932	122
15	672	211	30	908	184	45	987	384	60	1068	328

- (i) Select a systematic sample of 25% observations and show that systematic sampling is more precise than simple random sampling and stratified random sampling.
- (ii) Find ratio and regression estimator of total cultivable land for jute from a simple random sample of $n = 10$ villages. Also estimate the standard error of your estimators.
- (iii) If these 60 villages are considered a simple random sample of villages from 300 villages, then select a second sample of 15 villages and find ratio estimator for double sampling to estimate the total cultivable land for jute. Also estimate the standard error of your estimator.
36. Define cluster sampling. What is the need of this sampling technique? Write down the advantages and disadvantages of this sampling procedure. Suggest an estimator of population total under cluster sampling. Also find its sampling variance.
37. What are the reasons of cluster sampling? Compare this sampling technique with simple random sampling. When this sampling technique and simple random sampling become equally precise?
38. Discuss the advantages of cluster sampling. Consider that the clusters are of unequal sizes. In such a situation suggest an estimator of population total along with its sampling variance.

39. Distinguish between cluster sampling and two-stage sampling. Mention some practical situations where two-stage sampling can be used profitably. Discuss the advantages of these two sampling schemes compared to simple random sampling. If intra-class correlation coefficient is positive, then how would you plan cluster sampling?
40. What do you mean by multi-stage sampling? What is the need of this sampling scheme? Explain it with special reference to two-stage sampling.

If n clusters are selected at random from N clusters and if m ultimate units are selected from M ultimate units, then suggest an estimator of population total. How would you estimate the variance of your estimator?

41. In a district there are 1800 villages. The district is divided into 16 administrative units. Four administrative units are randomly selected and from the selected units some villages are randomly selected. The number of families in each village and number of cows in the villages are recorded and given below :

Sl.No. of Unit	No. of Villages in unit M_i	No. of Selected villages m_i	No. of families in selected villages x_{ij}	No. of cows in selected villages y_{ij}
1	18	5	28, 132, 96, 50, 62	12, 110, 45, 15, 20
2	10	3	62, 75, 44	15, 48, 12
3	7	2	70, 82	50, 62
4	15	4	52, 48, 110, 92	25, 12, 68, 88

Estimate the number of cows in the district. Also estimate the standard error of your estimate.

Estimate the number of cows by ratio method of estimation and estimate the variance of your estimate.

42. Define two-stage and three-stage samplings. Mention some practical situations where these sampling techniques can be used.

Suggest estimator of population total in two-stage sampling when clusters are of unequal size. Find the sampling variance of your estimator.

43. Explain three-stage sampling with example. To estimate the cost of living index number of industrial workers in a state, suggest a two-stage sample design to select the workers.

How would you estimate the population characteristic from such a sampling design? Also find the sampling variance of your estimator.

44. What is the difference between two-stage sampling and two-phase sampling? Explain the necessity of two-phase sampling.

How would you estimate the population mean under double sampling scheme? Is your estimator unbiased?

45. Show that ratio estimator and regression estimator in case of double sampling are biased. Find the relative bias of ratio estimator in case of double sampling. Also find the variance of such an estimator.

46. Explain double sampling for difference estimator along with the method of estimation.

47. Explain the rules for double sampling. Mention some practical populations where double sampling can be used.

Find the conditions under which double sampling is more precise than simple random sampling for same cost.

48. Define double sampling. How would you construct strata using double sampling scheme? Suggest estimator of population mean in such stratified sampling. Also find the sampling variance of your estimator.
49. How does double sampling differ from two-stage sampling? Explain the necessity of double sampling.

From a double sample the following information are recorded : $n_1 = 300$, $n = 87$, $\sum(y - \bar{y})^2 = 17284$, $\sum(x - \bar{x})^2 = 3248$, $\sum(x - \bar{x})(y - \bar{y}) = 5114$.

Estimate the variance of the regression estimator of population mean.

References

- Armitage, P. (1947): A comparison of stratified with unrestricted random sampling from a finite population. *Biometrika*, 34, 273-280.
- Avadhani, M. S. and Sukhatme, B. V. (1973): Controlled sampling with equal probabilities and with replacement. *Int. Stat. Rev.*, 41, 175-183.
- Bose, Chameli (1943): Note on the sampling error in the method of double sampling, *Sankhya*, 6, 330.
- Bowley, A. L. (1926): Measurement of precision attained in sampling. *Bull. Inter. Statist. Inst.*, 22, 1-62.
- Cochran, W. G. (1977): *Sampling Technique*, John Wiley and Sons Inc., New York.
- Cox, D. R. (1952): Estimation by double sampling, *Biometrika*, 19, 217-227.
- Dalenius, T. (1950): The problem of optimum stratification. *Sk and Akt.*, 33, 203-213.
- Delury, D. B. (1947): On the estimation of biological populations, *Biometrics*, 3, 145-167.
- Deming, W. E. (1950): *Some Theory of Sampling*, John Wiley and Sons Inc., New York.
- Des, Raj (1968): *Sampling Theory*. McGraw-Hill Book Company, New York.
- Evans, W. D. (1951): On stratification and optimum allocations. *Jour. Amer. Stats. Assoc.*, 46, 95-104.
- Ghosh, M. N. (1947): Survey of public opinion. *Calcutta Statist. Assoc. Bull. No. 1*, 13-18.
- Godambe, V. P. (1955): A unified theory of sampling from finite population. *Jour. Roy. Stat. Soc. B.*, 17, 269-278.
- Hajek, J. (1958): Some contributions to the theory of probability sampling. *Bull. Int. Stat. Inst.*, 36(3), 127-134.
- Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953): *Sample Survey Methods and Theory*, Vol. 1, Methods and Applications, Vol. II, Theory, John Wiley and Sons, New York.
- Hartley, H. O. and Rao, J. N. K. (1962): Sampling with Unequal probabilities and without replacement. *Ann. Math. Stats.*, 33, 350-374.
- Horvitz, D. G. and Thompson, D. J. (1952): A generalisation of sampling without replacement from a finite universe. *Jour. Amer. Stats. Assoc.*, 47, 663-685.

- Kish, L. (1957): Survey Sampling, John Wiley and Sons, New York.
- Lahiri, D. B. (1954): Technical paper on some aspects of the development of the sample design. *Sankhya*, 17, 332-362.
- Madow, L. H. (1950): On the use of country as a primary sampling unit for state estimates, *Jour. Amer. Stats. Assoc.*, 45, 30-47.
- Mahalanobis, P. C. (1944): On large scale sample surveys, *Phil. Trans. R. Soc.*, 231, 329-451.
- Mahalanobis, P. C. (1952): Some aspects of the design of sample surveys, *Sankhya*, 12, 1-7.
- Mahalanobis, P. C. and Lahiri, D. B. (1961): Analysis of errors in census and surveys with special reference to experience in India, *Bull. Int. Stat. Inst.*, 38, 401-433.
- Midzuno, H. (1952): Report of the survey design for agricultural production estimates in Ryuka Islands. *Ann. Inst. Statist. Math.*, 3, 109-121.
- Moser, C. A. (1958): *Survey Methods in Social Investigation*, Heinemann, London.
- Murthy, M. N. (1967): *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta, India.
- Neyman, J. (1958): Contribution to the theory of Sampling human population. *Jour. Amer. Stats. Assoc.*, 33, 101-116.
- Raj, D. (1964): Ratio estimation in sampling with equal and unequal probabilities, *Jour. Indian Soc. Agril. Stat.*, 6.
- Rao, J. N. K. (1968): Some small sample results in ratio and regression estimation. *Jour. Indian Stats. Assoc.*, 6, 160-168.
- Rao, J. N. K. and Singh M. P. (1973): On the choice of estimator in survey sampling. *Aust. Jour. Stat.*, 15, 95-104.
- Sukhatme, B. V. (1962): Some ratio-type estimators in two-phase sampling, *Jour. Amer. Stats. Assoc.*, 57, 628-632.
- Sukhatme, P. V. and Sukhatme, B. V. (1971): *Sampling Theory of Surveys with Applications*, Iowa State University Press. Amer, Iowa, USA.
- Warner, S. R. (1971): The linear randomized response model. *Jour. Amer. Stat. Assoc.*, 66, 884-888.
- Williams, W. H. (1962): On the variance of an estimator with post stratification. *Jour. Amer. Stat. Assoc.*, 57, 622-627.
- Yates, F. (1960): *Sampling Methods for Censuses and Surveys*. Charles Griffin and Co., London.
- Yates, F. and Grundy, P. M. (1953): Selection without replacement from within strata with probability proportional to size. *Jour. Roy. Stat. Soc. B.*, 15, 253-261.
- Zarkovich, S. S. (1965): *Sampling Methods and Census*. F.A.O., Rome.

Appendix—1

Table of Random Digits

1	2	3	4	5	6	7	8	9	10	11
1	51449	39284	85527	67168	91284	19954	91166	70918	85957	19492
2	16144	56830	67507	97275	25982	69294	32841	20861	83114	12531
3	48145	48230	99481	13050	81818	25282	66466	24461	97021	21072
4	83780	48351	85422	42978	26088	17869	94245	26622	48318	73850
5	95329	38482	93510	39170	63683	40587	80451	43058	81923	97072
6	11179	69004	34273	36062	26234	58601	47159	82248	95968	99722
7	94631	52413	31524	02316	27611	15888	13525	43809	40014	30667
8	64275	10294	35027	25604	65695	36014	17988	02734	31732	29911
9	72125	19232	10782	30615	42005	90419	32447	53688	36125	28456
10	16463	42028	27927	48403	88963	79615	41218	43290	53618	68082
11	10036	66273	69506	19610	01479	92338	55140	81097	73071	61544
12	85356	51400	88502	98267	73943	25828	38219	13268	09016	77465
13	84076	82087	55053	75370	71030	92275	55497	97123	40919	57479
14	76731	39755	78537	51937	11680	78820	50082	56068	36908	55399
15	19032	73472	79399	05549	14772	32746	38841	45524	13535	03113
16	72791	59040	61529	74437	74482	76619	05232	28616	98690	24011
17	11553	00135	28306	65571	34465	47423	39198	54456	95283	54637
18	71405	70352	76763	64002	62461	41982	15933	46942	36941	93412
19	17594	10116	55483	96219	85493	96955	89180	59690	82170	77643
20	09584	23476	09243	65568	89128	36747	63692	09986	47687	46448
21	81677	62634	52794	01466	85938	14565	79993	44956	82254	65223
22	45849	01177	13773	43523	69825	03222	58458	77463	58521	07273
23	97252	92257	90419	01241	52516	66293	14536	23870	78402	41759
24	26232	77422	76289	57587	42831	87047	20092	92676	12017	43554
25	87799	33602	01931	66913	63008	03745	93939	07178	70003	18158
26	46120	62298	69126	07862	76731	58527	39342	42749	57050	91725
27	53292	55652	11834	47581	25682	64085	26587	92289	41853	38354
28	81606	56009	06021	98392	40450	87721	50917	16978	39472	23505
29	67819	47314	96988	89931	49395	37071	72658	53947	11996	64631
30	50458	20350	87362	83996	86422	58694	71813	97695	28804	58523
31	59772	27000	97805	25042	09916	77569	71347	62667	09330	02152
32	94752	91056	08939	93410	59204	04644	44336	55570	21106	76588
33	01885	82054	45944	55398	55487	56455	56940	68787	36591	29914
34	85190	91941	86714	76593	77199	39724	99548	13827	84961	76740
35	97747	67607	14549	08215	95408	46381	12449	03672	40325	77312
36	43318	84469	26047	86003	34786	38931	34846	28711	42833	93019
37	47874	71365	76603	57440	49514	17335	71969	58055	99136	73589
38	24259	48079	71198	95859	94212	55402	93392	31965	94622	11673
39	31947	64805	34133	03245	24546	48934	41730	47831	26531	02230
40	37911	93224	87153	54541	57529	38299	65659	00202	07054	40168
41	82714	15799	93126	74180	94171	97117	31431	00323	62793	11995
42	82927	37884	74411	45887	36713	52339	68421	35968	67714	05883
43	65934	21782	35804	36676	35404	69987	52268	19894	81977	87764
44	56953	04356	68903	21369	35901	86797	83901	68581	02397	55359
45	16278	17165	67843	49349	90163	97337	35003	34915	91485	33814
46	96339	95028	48468	12279	81039	56531	10759	19579	00015	22829
47	84110	49661	13988	75909	35580	18426	29038	79111	56049	96451
48	49017	60748	03412	09880	94091	90052	43596	21424	16584	67970
49	43560	05552	54344	69418	01327	07771	25364	77373	34841	75927
50	25206	15177	63049	12464	16149	18759	96184	15968	89446	07168

Appendix—3

Critical Values for Chi-Squared Tests

Degrees of Freedom	10% Level	5% Level	1% Level	0.1% Level
1	2.706	3.841	6.635	10.828
2	4.605	5.991	9.210	13.816
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.467
5	9.236	11.071	15.086	20.515
6	10.645	12.592	16.812	22.458
7	12.017	14.067	18.475	24.322
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.528
14	21.064	23.685	29.141	36.123
15	22.307	24.996	30.578	37.697
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.790
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.820
20	28.412	31.410	37.566	45.315
21	29.615	32.671	38.932	46.797
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.620
26	35.563	38.885	45.642	54.052
27	36.741	40.113	46.963	55.476
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.703
31	41.422	44.985	52.191	61.098
32	42.585	46.194	53.486	62.487
33	43.745	47.400	54.776	63.870
34	44.903	48.602	56.061	65.247
35	46.059	48.802	57.342	66.619
36	47.212	50.998	58.619	67.985
37	48.363	52.192	59.893	69.346
38	49.513	53.384	61.162	70.703
39	50.660	54.572	62.428	72.055
40	51.805	55.758	63.691	73.402
41	52.949	56.942	64.950	74.745
42	54.090	58.124	66.206	76.084
43	55.230	59.304	67.459	77.419
44	56.369	60.481	68.710	78.749
45	57.505	61.656	69.957	80.077
46	58.641	62.830	71.201	81.400
47	59.774	64.001	72.443	82.720
48	60.907	65.171	73.683	84.037
49	62.038	66.339	74.919	85.351
50	63.167	67.505	76.154	86.661

Appendix—3 (Continued)

Degrees of Freedom	10% Level	5% Level	1% Level	0.1% Level
51	64.295	68.669	77.386	87.968
52	65.422	69.832	78.616	89.272
53	66.548	70.993	79.843	90.573
54	67.673	72.153	81.069	91.872
55	68.796	73.311	82.292	93.167
56	69.919	74.468	83.513	94.461
57	71.040	75.624	84.733	95.751
58	72.160	76.778	85.950	97.039
59	73.279	77.931	87.166	98.324
60	74.397	79.082	88.379	99.607
61	75.514	80.232	89.591	100.888
62	76.630	81.381	90.802	102.166
63	77.745	82.529	92.010	103.442
64	78.860	83.675	93.217	104.716
65	79.973	84.821	94.422	105.988
66	81.085	85.965	95.626	107.258
67	82.197	87.108	96.828	108.526
68	83.308	88.250	98.028	109.791
69	84.418	89.391	99.228	111.055
70	85.527	90.531	100.425	112.317
71	86.635	91.670	101.621	113.577
72	87.743	92.808	102.816	114.835
73	88.850	93.945	104.010	116.091
74	89.956	95.081	105.202	117.346
75	91.061	96.217	106.393	118.599
76	92.166	97.351	107.583	119.850
77	93.270	98.484	108.771	121.100
78	94.374	99.617	109.958	122.348
79	95.476	100.749	111.144	123.594
80	96.578	101.879	112.329	124.839
81	97.680	103.010	113.512	126.083
82	98.780	104.139	114.695	127.324
83	99.880	105.267	115.876	127.565
84	100.980	106.395	117.057	129.804
85	102.079	107.522	118.236	131.041
86	103.177	108.648	119.414	132.277
87	104.275	109.773	120.591	133.512
88	105.372	110.898	121.767	134.745
89	106.469	112.022	122.942	135.978
90	107.565	113.145	124.116	137.208
91	108.661	114.268	125.289	138.438
92	109.756	115.390	126.462	139.666
93	110.850	116.511	127.633	140.893
94	111.944	117.632	128.803	142.119
95	113.038	118.752	129.973	143.344
96	114.131	119.871	131.141	144.567
97	115.223	120.990	132.309	145.789
98	116.315	122.108	133.476	147.010
99	117.407	123.225	134.642	148.230
100	118.498	124.342	135.807	149.449

Appendix—4

The *t* table

Confidence level :		80%	90%	95%	98%	99%	99.8%	99.9%
Two-sided								
One-sided		90%	95%	97.5%	99%	99.5%	99.9%	99.95%
Hypothesis Test level :								
Two-sided		0.20	0.10	0.05	0.02	0.01	0.002	0.001
One sided		0.10	0.05	0.025	0.01	0.005	0.001	0.0005

For one Sample : <i>n</i>	In general : Degree of freedom	Critical values						
2	1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
3	2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
4	3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
5	4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
6	5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
7	6	1.440	1.947	2.447	3.143	3.707	5.208	5.959
8	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
9	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
10	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
11	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
12	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
13	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
14	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
15	14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
16	15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
17	16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
18	17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
19	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
20	19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
21	20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
22	21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
23	22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
24	23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
25	24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
26	25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
27	26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
28	27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
29	28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
30	29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
31	30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
32	31	1.309	1.696	2.040	2.453	2.744	3.375	3.633
33	32	1.309	1.694	2.037	2.449	2.738	3.365	3.622
34	33	1.308	1.692	2.035	2.445	2.733	3.356	3.611
35	34	1.307	1.691	2.032	2.441	2.728	3.348	3.601
36	35	1.306	1.690	2.030	2.438	2.724	3.340	3.591
37	36	1.306	1.688	2.028	2.434	2.719	3.333	3.582
38	37	1.305	1.687	2.026	2.431	2.715	3.326	3.574
39	38	1.304	1.686	2.024	2.429	2.712	3.319	3.566
40	39	1.304	1.685	2.023	2.426	2.708	3.313	3.558
	Infinity	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Appendix—5

F Distribution

Values of F Exceeded with Probabilities of 5 and 1 per cent

		df (numerator)																								∞
		1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	248	249	250	251	252	253	254	254	254	254	
	4.052	4.999	5.403	5.625	5.764	5.859	5.928	5.981	6.022	6.056	6.082	6.106	6.142	6.169	6.208	6.234	6.261	6.286	6.302	6.323	6.334	6.352	6.361	6.366	6.366	
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.49	19.50	19.50	
	98.49	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.49	99.50	99.50	99.50	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.57	8.56	8.45	8.54	8.53	
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.14	26.12	26.12	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63	5.63	
	21.00	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37	14.24	14.15	14.02	13.93	13.83	13.74	13.69	13.61	13.57	13.52	13.48	13.46	13.46	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36	4.36	
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.29	10.15	10.05	9.96	9.89	9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.17	9.13	9.07	9.04	9.02	9.02	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	3.99	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67	3.67	3.67	
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.94	6.90	6.88	6.88	
7	5.59	4.74	4.34	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23	3.23	
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47	6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5.65	5.65	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93	2.93	
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67	5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86	4.86	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71	2.71	
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11	5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31	4.31	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54	2.54	
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71	4.60	4.52	4.41	4.33	4.25	4.17	4.12	4.05	4.01	3.96	3.93	3.91	3.91	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40	2.40	
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40	4.29	4.21	4.10	4.02	3.94	3.86	3.80	3.74	3.70	3.66	3.62	3.60	3.60	
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30	2.30	
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16	4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36	3.36	
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21	2.21	
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.85	3.78	3.67	3.59	3.51	3.42	3.37	3.30	3.27	3.21	3.18	3.16	3.16	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13	2.13	
	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00	3.00	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07	2.07	
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87	2.87	
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01	2.01	
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.88	2.86	2.80	2.77	2.75	2.75	
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96	1.96	
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65	2.65	
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92	1.92	
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37	3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57	2.57	

Contd.

Appendix-5 Contd.

19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.49
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.84
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23	3.13	3.05	2.94	2.86	2.77	2.69	2.63	2.56	2.53	2.47	2.42
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.81
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.32	3.17	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.78
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02	2.94	2.85	2.75	2.67	2.58	2.46	2.42	2.37	2.33	2.31
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.76
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.53	2.48	2.41	2.37	2.32	2.28
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.04	1.98	1.94	1.83	1.86	1.82	1.80	1.76	1.74
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.71
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99	1.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96	1.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.65
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59
	7.44	5.29	4.42	3.93	3.61	3.33	3.21	3.08	2.97	2.89	2.82	2.76	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56
	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.76	2.72	2.62	2.54	2.43	2.35	2.26	2.17	2.12	2.04	2.00	1.94	1.90
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54
	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75	2.69	2.59	2.51	2.40	2.32	2.22	2.14	2.08	2.00	1.97	1.90	1.86
40	4.07	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.51
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84

Contd.

Appendix—5 Contd.

Values of F Exceeded with Probabilities of 5 and 1 Per cent

	df (numerator)																∞							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24		30	40	50	75	100	200	500
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64	2.54	2.46	2.46	2.35	2.26	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62	2.52	2.44	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60	2.50	2.42	2.42	2.30	2.22	2.13	2.04	1.98	1.90	1.86	1.80	1.76	1.72
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58	2.48	2.40	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.89	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56	2.46	2.39	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.76	1.71	1.68
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
7.15	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.69	2.62	2.56	2.46	2.39	2.39	2.25	2.15	2.06	1.96	1.90	1.82	1.78	1.71	1.66	1.64
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.40	2.32	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.79	2.70	2.61	2.54	2.47	2.37	2.30	2.30	2.18	2.09	2.00	1.90	1.84	1.76	1.71	1.64	1.60	1.56
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45	2.35	2.28	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.44	1.42	1.38	1.35	1.32
6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41	2.32	2.24	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36	2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43	
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.85	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.65	2.56	2.47	2.40	2.33	2.23	2.15	2.03	1.94	1.85	1.75	1.68	1.59	1.54	1.46	1.40	1.37	
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
6.81	4.75	3.91	3.44	3.14	2.92	2.76	2.62	2.53	2.44	2.37	2.30	2.20	2.12	2.00	1.91	1.83	1.72	1.66	1.56	1.51	1.43	1.37	1.33	
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34	2.28	2.17	2.09	1.97	1.88	1.79	1.69	1.62	1.53	1.48	1.39	1.33	1.28	
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
6.70	4.66	3.83	3.36	3.06	2.85	2.68	2.55	2.46	2.37	2.29	2.23	2.12	2.04	1.92	1.84	1.74	1.64	1.57	1.47	1.42	1.32	1.24	1.19	
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20	2.09	2.01	1.98	1.81	1.71	1.61	1.54	1.44	1.38	1.28	1.19	1.11	
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00	

Source : E.S. Pearson and H.O. Hartley Biometrika tables for statisticians, Vol. 1 (Cambridge, England : Cambridge University Press 1966), pp. 171-173.

Appendix—6

Dunnett Table

Number of Groups Including Control Group

dfc	alpha	2	3	4	5	6	7	8	9	10
5	0.05	2.57	3.03	3.29	3.48	3.62	3.73	3.82	3.9	3.97
	0.01	4.03	4.63	4.98	5.22	5.41	5.56	5.69	5.8	5.89
6	0.05	2.45	2.86	3.1	3.26	3.39	3.49	3.57	3.64	3.71
	0.01	3.71	4.21	4.51	4.71	4.87	5	5.1	5.2	5.28
7	0.05	2.36	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53
	0.01	3.5	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89
8	0.05	2.31	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41
	0.01	3.36	3.77	4	4.17	4.29	4.4	4.48	4.56	4.62
9	0.05	2.26	2.61	2.81	2.95	3.05	3.14	3.2	3.26	3.32
	0.01	3.25	3.63	3.85	4.01	4.12	4.22	4.3	4.37	4.43
10	0.05	2.23	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24
	0.01	3.17	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28
11	0.05	2.2	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19
	0.01	3.11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16
12	0.05	2.18	2.5	2.68	2.81	2.9	2.98	3.04	3.09	3.14
	0.01	3.05	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07
13	0.05	2.16	2.48	2.65	2.78	2.87	2.94	3	3.06	3.1
	0.01	3.01	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99
14	0.05	2.14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07
	0.01	2.98	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93
15	0.05	2.13	2.44	2.61	2.73	2.82	2.89	2.95	3	3.04
	0.01	2.95	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88
16	0.05	2.12	2.42	2.59	2.71	2.8	2.87	2.92	2.97	3.02
	0.01	2.92	3.22	3.39	3.51	3.6	3.67	3.73	3.78	3.83
17	0.05	2.11	2.41	2.58	2.69	2.78	2.85	2.9	2.95	3
	0.01	2.9	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79
18	0.05	2.1	2.4	2.56	2.68	2.76	2.83	2.89	2.94	2.98
	0.01	2.88	3.17	3.33	3.44	3.53	3.6	3.66	3.71	3.75
19	0.05	2.09	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96
	0.01	2.86	3.15	3.31	3.42	3.5	3.57	3.63	3.68	3.72
20	0.05	2.09	2.38	2.54	2.65	2.73	2.8	2.86	2.9	2.95
	0.01	2.85	3.13	3.29	3.4	3.48	3.55	3.6	3.65	3.69
24	0.05	2.06	2.35	2.51	2.61	2.7	2.76	2.81	2.86	2.9
	0.01	2.8	3.07	3.22	3.32	3.4	3.47	3.52	3.57	3.61
30	0.05	2.04	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86
	0.01	2.75	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52
40	0.05	2.02	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81
	0.01	2.7	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44
60	0.05	2	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77
	0.01	2.66	2.9	3.03	3.12	3.19	3.25	3.29	3.33	3.37

Appendix—7

Critical values $q'(p, df; 0.05)$ for Duncan's multiple range test

df	$p \rightarrow$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.926	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.796	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814
6	3.460	3.586	3.649	3.680	3.694	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.625	3.625	3.625	3.625	3.625	3.625	3.625	3.625	3.625	3.625	3.625	3.625	3.625
8	3.261	3.398	3.475	3.521	3.549	3.566	3.575	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522	3.525	3.525	3.525	3.525	3.525	3.525	3.525	3.525	3.525	3.525
11	3.113	3.256	3.341	3.397	3.435	3.462	3.480	3.493	3.501	3.506	3.509	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510
12	3.081	3.225	3.312	3.370	3.410	3.439	3.459	3.474	3.484	3.491	3.495	3.498	3.498	3.498	3.498	3.498	3.498	3.498	3.498
13	3.055	3.200	3.288	3.348	3.389	3.419	3.441	3.458	3.470	3.478	3.484	3.488	3.490	3.490	3.490	3.490	3.490	3.490	3.490
14	3.033	3.178	3.268	3.328	3.371	3.403	3.426	3.444	3.457	3.467	3.474	3.479	3.482	3.484	3.484	3.484	3.484	3.484	3.484
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.432	3.446	3.457	3.465	3.471	3.476	3.478	3.480	3.480	3.480	3.480	3.480
16	2.998	3.144	3.235	3.297	3.343	3.376	3.402	3.422	3.437	3.449	3.458	3.465	3.470	3.473	3.476	3.477	3.477	3.477	3.477
17	2.984	3.130	3.222	3.285	3.331	3.365	3.392	3.412	3.429	3.441	3.451	3.459	3.465	3.469	3.472	3.474	3.475	3.475	3.475
18	2.971	3.117	3.210	3.274	3.320	3.356	3.383	3.404	3.421	3.435	3.445	3.454	3.460	3.463	3.466	3.469	3.472	3.473	3.474
19	2.960	3.106	3.199	3.264	3.311	3.347	3.375	3.397	3.415	3.429	3.440	3.449	3.456	3.462	3.466	3.469	3.472	3.473	3.473
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.390	3.409	3.423	3.435	3.445	3.452	3.456	3.461	3.463	3.467	3.470	3.472
21	2.941	3.088	3.181	3.247	3.295	3.332	3.361	3.385	3.403	3.418	3.431	3.441	3.449	3.456	3.461	3.465	3.469	3.471	3.471
22	2.933	3.080	3.173	3.239	3.288	3.326	3.355	3.379	3.398	3.414	3.427	3.437	3.446	3.453	3.459	3.464	3.467	3.470	3.470
23	2.926	3.072	3.166	3.233	3.282	3.320	3.350	3.374	3.394	3.410	3.423	3.434	3.443	3.451	3.457	3.462	3.466	3.469	3.469
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390	3.406	3.420	3.431	3.441	3.449	3.455	3.461	3.465	3.469	3.469
25	2.913	3.059	3.154	3.221	3.271	3.310	3.341	3.366	3.386	3.403	3.417	3.429	3.439	3.447	3.454	3.459	3.464	3.468	3.468
26	2.907	3.054	3.149	3.216	3.266	3.305	3.336	3.362	3.382	3.400	3.414	3.426	3.436	3.445	3.452	3.458	3.463	3.468	3.468
27	2.902	3.049	3.144	3.211	3.262	3.301	3.332	3.358	3.379	3.397	3.412	3.424	3.434	3.443	3.451	3.457	3.463	3.467	3.467
28	2.897	3.044	3.139	3.206	3.257	3.297	3.329	3.355	3.376	3.394	3.409	3.422	3.433	3.442	3.450	3.456	3.462	3.467	3.467
29	2.892	3.039	3.135	3.202	3.253	3.293	3.326	3.352	3.373	3.392	3.407	3.420	3.431	3.440	3.448	3.455	3.461	3.466	3.466
30	2.888	3.035	3.131	3.199	3.250	3.290	3.323	3.349	3.371	3.389	3.405	3.418	3.429	3.439	3.447	3.454	3.460	3.466	3.466
31	2.884	3.031	3.127	3.195	3.246	3.287	3.321	3.348	3.368	3.387	3.403	3.416	3.428	3.438	3.446	3.454	3.460	3.466	3.466
32	2.881	3.028	3.123	3.192	3.243	3.284	3.317	3.344	3.366	3.385	3.401	3.415	3.426	3.436	3.445	3.453	3.459	3.465	3.465
33	2.877	3.024	3.120	3.188	3.240	3.281	3.314	3.341	3.364	3.383	3.399	3.413	3.425	3.435	3.444	3.452	3.459	3.465	3.465
34	2.874	3.021	3.117	3.185	3.238	3.279	3.312	3.339	3.362	3.381	3.398	3.412	3.424	3.434	3.443	3.451	3.458	3.464	3.464
35	2.871	3.018	3.114	3.183	3.235	3.276	3.309	3.337	3.360	3.379	3.396	3.410	3.423	3.433	3.443	3.451	3.458	3.464	3.464
36	2.868	3.015	3.111	3.180	3.232	3.273	3.306	3.335	3.358	3.378	3.395	3.409	3.421	3.432	3.442	3.450	3.457	3.463	3.463
37	2.865	3.013	3.109	3.178	3.230	3.271	3.305	3.334	3.357	3.376	3.393	3.408	3.420	3.431	3.441	3.449	3.457	3.463	3.463
38	2.863	3.010	3.106	3.175	3.228	3.270	3.303	3.331	3.355	3.375	3.392	3.407	3.419	3.431	3.440	3.449	3.456	3.463	3.463
39	2.861	3.008	3.104	3.173	3.226	3.268	3.301	3.330	3.353	3.373	3.391	3.406	3.418	3.430	3.440	3.448	3.456	3.463	3.463
40	2.858	3.005	3.102	3.171	3.224	3.266	3.300	3.328	3.352	3.372	3.389	3.404	3.418	3.429	3.439	3.447	3.456	3.463	3.463
48	2.843	2.991	3.087	3.157	3.211	3.253	3.288	3.318	3.342	3.363	3.382	3.398	3.412	3.424	3.435	3.445	3.453	3.461	3.461
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333	3.355	3.374	3.391	3.406	3.419	3.431	3.441	3.451	3.458	3.458
80	2.814	2.961	3.059	3.130	3.185	3.229	3.266	3.297	3.323	3.346	3.366	3.384	3.400	3.414	3.427	3.438	3.449	3.458	3.458
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.286	3.313	3.337	3.358	3.377	3.394	3.409	3.423	3.435	3.446	3.457	3.457
240	2.786	2.933	3.031	3.103	3.159	3.205	3.243	3.276	3.304	3.329	3.350	3.370	3.388	3.404	3.418	3.432	3.444	3.455	3.455
Inf	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294	3.320	3.343	3.363	3.382	3.399	3.414	3.428	3.442	3.454	3.454

Critical values $q'(p, df; 0.01)$ for Duncan's multiple range test

df	p →	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	90.024	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036
2	8.260	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321
3	6.511	6.677	6.740	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755	6.755
4	5.702	5.893	5.989	6.040	6.065	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074
5	5.243	5.439	5.549	5.614	5.655	5.680	5.694	5.701	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703
6	4.949	5.145	5.260	5.333	5.383	5.416	5.439	5.454	5.464	5.470	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472
7	4.745	4.939	5.056	5.134	5.189	5.227	5.256	5.276	5.291	5.302	5.309	5.313	5.316	5.317	5.317	5.317	5.317	5.317	5.317
8	4.596	4.787	4.906	4.986	5.043	5.086	5.117	5.142	5.160	5.174	5.185	5.193	5.199	5.202	5.205	5.206	5.206	5.206	5.206
9	4.482	4.671	4.789	4.871	4.931	4.975	5.010	5.036	5.058	5.074	5.087	5.098	5.106	5.112	5.117	5.120	5.122	5.123	5.123
10	4.392	4.579	4.697	4.780	4.841	4.887	4.923	4.952	4.975	4.994	5.009	5.021	5.031	5.039	5.045	5.050	5.054	5.057	5.057
11	4.320	4.504	4.622	4.705	4.767	4.815	4.852	4.882	4.907	4.927	4.944	4.957	4.969	4.978	4.986	4.993	4.998	5.002	5.002
12	4.260	4.442	4.560	4.643	4.706	4.754	4.793	4.824	4.850	4.871	4.889	4.904	4.917	4.927	4.936	4.944	4.950	4.955	4.955
13	4.210	4.391	4.508	4.591	4.654	4.703	4.743	4.775	4.802	4.824	4.843	4.859	4.872	4.884	4.894	4.902	4.909	4.916	4.916
14	4.167	4.346	4.463	4.547	4.610	4.660	4.700	4.733	4.760	4.783	4.803	4.820	4.834	4.846	4.857	4.866	4.874	4.881	4.881
15	4.131	4.308	4.425	4.508	4.572	4.622	4.662	4.696	4.724	4.748	4.768	4.785	4.800	4.813	4.825	4.835	4.843	4.851	4.851
16	4.099	4.275	4.391	4.474	4.538	4.589	4.630	4.664	4.692	4.717	4.737	4.755	4.771	4.785	4.797	4.807	4.816	4.824	4.824
17	4.071	4.246	4.361	4.445	4.509	4.559	4.601	4.635	4.664	4.689	4.710	4.729	4.745	4.759	4.771	4.782	4.792	4.801	4.801
18	4.046	4.220	4.335	4.418	4.483	4.533	4.575	4.610	4.639	4.664	4.686	4.705	4.722	4.736	4.749	4.760	4.771	4.780	4.780
19	4.024	4.197	4.312	4.395	4.459	4.510	4.552	4.587	4.617	4.642	4.664	4.684	4.701	4.716	4.729	4.741	4.751	4.761	4.761
20	4.004	4.177	4.291	4.374	4.438	4.489	4.531	4.567	4.597	4.622	4.645	4.664	4.682	4.697	4.711	4.723	4.734	4.743	4.743
21	3.986	4.158	4.272	4.355	4.419	4.470	4.513	4.548	4.578	4.604	4.627	4.647	4.664	4.680	4.694	4.706	4.718	4.728	4.728
22	3.970	4.141	4.254	4.337	4.402	4.453	4.496	4.531	4.562	4.588	4.611	4.631	4.649	4.665	4.679	4.692	4.703	4.713	4.713
23	3.955	4.126	4.239	4.322	4.386	4.437	4.480	4.516	4.546	4.573	4.596	4.616	4.634	4.651	4.665	4.678	4.690	4.700	4.700
24	3.942	4.112	4.224	4.307	4.371	4.423	4.466	4.502	4.532	4.559	4.582	4.603	4.621	4.638	4.652	4.665	4.677	4.688	4.688
25	3.930	4.099	4.211	4.294	4.358	4.410	4.452	4.489	4.520	4.546	4.570	4.591	4.609	4.626	4.640	4.654	4.666	4.677	4.677
26	3.918	4.087	4.199	4.282	4.346	4.397	4.440	4.477	4.508	4.535	4.558	4.579	4.598	4.615	4.630	4.643	4.655	4.667	4.667
27	3.908	4.076	4.188	4.270	4.334	4.386	4.429	4.465	4.497	4.524	4.548	4.569	4.587	4.604	4.619	4.633	4.646	4.657	4.657
28	3.898	4.065	4.177	4.260	4.324	4.376	4.419	4.455	4.486	4.514	4.538	4.559	4.578	4.595	4.610	4.624	4.637	4.648	4.648
29	3.889	4.056	4.168	4.250	4.314	4.366	4.409	4.445	4.477	4.504	4.528	4.550	4.569	4.586	4.601	4.615	4.628	4.640	4.640
30	3.881	4.047	4.159	4.241	4.305	4.357	4.400	4.436	4.468	4.495	4.519	4.541	4.560	4.577	4.593	4.607	4.620	4.632	4.632
31	3.873	4.039	4.150	4.232	4.296	4.348	4.391	4.428	4.459	4.487	4.511	4.533	4.552	4.570	4.588	4.600	4.613	4.625	4.625
32	3.865	4.031	4.142	4.224	4.288	4.340	4.383	4.420	4.452	4.479	4.504	4.525	4.545	4.562	4.578	4.592	4.606	4.618	4.618
33	3.859	4.024	4.135	4.217	4.281	4.333	4.376	4.413	4.444	4.472	4.496	4.518	4.538	4.555	4.571	4.586	4.599	4.611	4.611
34	3.853	4.017	4.128	4.210	4.273	4.325	4.369	4.406	4.437	4.465	4.490	4.511	4.531	4.549	4.565	4.579	4.593	4.605	4.605
35	3.846	4.011	4.121	4.203	4.267	4.319	4.362	4.399	4.431	4.459	4.483	4.505	4.525	4.543	4.559	4.573	4.587	4.599	4.599
36	3.840	4.005	4.115	4.197	4.260	4.312	4.356	4.393	4.425	4.452	4.477	4.499	4.519	4.537	4.553	4.568	4.581	4.594	4.594
37	3.835	3.999	4.109	4.191	4.254	4.306	4.350	4.387	4.419	4.447	4.471	4.493	4.513	4.531	4.548	4.562	4.576	4.589	4.589

Index

- Alias, 166
- Basic design, 72
- Best linear unbiased estimate, 17, 423
- Bias; definition, 339
 - due to errors in stratum weights, 505
 - in ratio estimator, 420
 - of regression estimator, 436
- Barlett's test, 305
- Census, 332
- Cluster sampling (single stage), 450
- Clusters of equal sizes, 451
- Circular, systematic sampling, 394
- Cochran's theorem, 19
- Combining inter-block and intra-block information, 221
- Confounded factorial experiment, 151
- Contrast, 12
- Covariance analysis, 236
- Cost function in determining sample size, 367, 433
- Defining contrast, 166
- Double sampling, 516
 - for stratification, 517
 - for ratio estimator, 519
 - for regression estimator, 523
- Duncan's multiple range test, 21
- Dunnett's test, 23
- Estimation of population variance, 347
- Estimation of variance component, 267
- Finite population correction, 345
- Frame, 331, 334
- Graeco Latin square design, 121
- Generalized interaction, 166
- Incomplete block design, 213
- Interaction, 135
- Inter-block analysis, 220
- Intra-block analysis, 218
- Judgement sampling, 330
- Latin square design, 90
- Linear regression estimator, 435
- Lottery method, 341
- Main effect, 124
- Missing values, 257
- Multiple comparison, 20
- Multiphase sampling, 516
- Neyman allocation, 367
- Nonresponse, 570
- Optimum allocation in stratified sampling, 367
- Orthogonal latin square design, 119
- Pilot survey, 333
- Principal block, 166
- Primary units, 472
- Proportional allocation, 358
- Proportion, 350
- Purposive selection, 330
- Quota sampling, 331
- Random numbers, 332, 339
- Random sampling, 331
- Random effect model, 6, 266
- Ratio estimator, 408
- Randomized block design, 101
- Replication, 9
- Regression estimator, 435

Sampling error, 336
Sampling fraction, 345
Sampling with replacement, 341
Simple random sampling, 341
Size of sample, 337
Steps in sample survey, 333
Strata, 357
Statified random sampling, 357
Systematic sampling, 393

Three-stage sampling, 503

Treatment combination, 136
Two-phase sampling, 516
Two-stage sampling, 472

Unbiased ratio-type estimator, 415

Variance component, 266

Weighted analysis, 300

Yates algorithm, 139