

Sampling Theory and Methods

M. N. MURTHY

*Professor and Head of Design Division
National Sample Survey
Indian Statistical Institute*

NOT TO BE ISSUED

Statistical Publishing Society

P. V. 83

© Copyright 1967

STATISTICAL PUBLISHING SOCIETY

204/1, Barrackpore Trunk Road

Calcutta 35, India

UNIVERSITY OF KALYANI LIBRARY
ACC. NO 106964 Maths.
CALL NO B.S.W 17

First Published in 1967

PV 83

Ex. by

Ex. by

Printed in India
At the Eka Press, Calcutta-35

To my parents
Kamala Bai and
Lakshmanachar

Preface

THIS book attempts to give an up-to-date account of the theoretical frame-work and the methodology of sampling from finite populations with particular attention to methods currently practised and to those having potentialities of wide application. It is intended to serve a two-fold purpose : primarily as a text book for students at the universities as well as in institutions imparting professional training in statistics, and as a work of reference for instructors, research scholars and professional workers in statistical surveys. This book is the outcome of my intimate association with the sample designing and planning of the large-scale nation-wide Indian National Sample Survey, and of my experience of teaching the subject over the years to the trainees of the various courses conducted by the Indian Statistical Institute. In fact, the very favourable reception accorded, both in India and abroad, to my earlier book *Introduction to Sampling Theory* (published a few years ago in the *Lecture Notes* series of the Research and Training School of the Institute) has induced me to prepare this revised and enlarged version of those *Notes*.

The theory and practice of sampling and estimation have been discussed in this book with several illustrative examples drawn from various fields of application such as agriculture, industry, economics and demography. The treatment of the first few chapters has been deliberately made somewhat elementary and didactic, many numerical examples having been put in to enable the beginner to acquire a thorough grasp of the concepts and the methods discussed. This grounding, it is expected, should help him considerably to follow the later chapters where the treatment is more advanced and concise. Numerous theoretical and practical problems covering different aspects have been provided to enable the reader to familiarize himself with and gain proficiency in the various techniques and methods.

An elementary knowledge of algebra, probability and statistics would be adequate for a study of this book.

In the treatment of any particular topic, the commonly used aspects are given in detail while the more advanced ones are only briefly mentioned. Alternative methods and derivations of results requiring relatively difficult proofs have been printed in smaller type and the reader may skip them at the first reading without any loss of continuity of the subject matter. The conditional probability approach has been used in the case of sampling designs involving two or more stages of randomization thereby achieving considerable simplification in the derivations.

Since a general discussion of the sampling methodology does not bring out clearly the problems of sample designing in planning the conduct of large scale surveys, also studies of nation wide statistical surveys covering social economic and agricultural characteristics have been given to provide an insight into the types of survey designs actually used in practice and I presume that these would be of special interest to professional workers engaged in or associated with planning and carrying out statistical surveys. Further, an extensive bibliography on sampling theory and methods with the papers classified by field of application and method of sampling has been included, and I believe this would be very useful to teachers and research workers in this field.

This book consists of sixteen chapters. In Chapter 1, a general discussion on the need for sample surveys is given. Chapter 2 is devoted to concepts, definitions and notations used in the later chapters. The basic sampling schemes (*simple random sampling, systematic sampling and varying probability sampling*) are covered in Chapters 3 to 6. Other sampling schemes (*stratified sampling, cluster sampling, multi-stage and multi phase sampling*) are discussed in Chapters 7 to 9. The estimation of ratios and differences and the use of auxiliary information at the estimation stage are considered in Chapters 10 and 11. Chapter 12 gives a description of the methods of making a sample design self weighting. The question of measurement

and control of non-sampling errors is considered in Chapter 13. The various steps involved in planning a statistical survey are discussed in Chapter 14. Chapters 15 and 16 present the details of the sample designs of the Indian National Sample Survey, carried out on an all-India basis as an integrated multi-subject survey, and of the Family Living Surveys, conducted in a number of middle class and working class (factory, mining and plantation) centres in India. Random numbers are given in Appendix 1 for use in practising sample selection techniques. In Appendix 2, a consolidated list of formulae for estimators and variance estimators is given for ready reference. Appendix 3 provides solutions to problems included at the end of each chapter.

I am grateful to the Department of Statistics, Government of India, for their permission to include the material given in Chapter 15, which is based on a paper of mine originally published by them, and to Mrs. B. N. Chinnappa and the Editors of *Sankhyā* for allowing me to include the matter given in Chapter 16 based on a paper published in that journal.

I wish to express my gratitude to Professor D. B. Lahiri for the numerous interesting and instructive discussions I have had with him and for his very useful suggestions during the preparation of this book. I am grateful to Dr. S. K. Mitra, Mrs. B. N. Chinnappa, Mr. A. S. Roy and Dr. A. Matthai for seeing through the manuscript and giving valuable comments, to Mr. K. B. Goswami for his generous assistance in getting the manuscript ready for the press, and to Miss Nilima Das and Miss P. K. Vasanthi for their able help in typing the manuscript. I greatly appreciate the efficient manner in which the staff of the Eka Press have printed the book in a short time. My thanks are particularly due to Dr. C. R. Rao, F.R.S., Director of the Research and Training School, Indian Statistical Institute, for the keen interest he has taken in this book right from the initial stages. My wife, Vijaya, has immensely helped me during the various stages of the preparation of this book by sharing the work of computations and proof correction.

Needless to say, I have amply drawn upon the rich and varied contributions the Indian Statistical Institute has made to the theoretical and the methodological aspects of sample surveys under the inspiring guidance of Professor P. C. Mahalanobis, F.R.S., during the last three decades.

Calcutta-35, India

September, 1967

M. N. MURTHY

Contents

CHAPTER	PAGE
1. NEED FOR SAMPLE SURVEY	1-22
1.1 Need for Statistical Information ✓	1
1.2 Types of Data ...	3
1.3 Complete Enumeration Survey ..✓	6
1.4 Need for Sampling ✓	10
1.5 Sampling and Non-sampling Errors ✓	11
1.6 Cost Aspect ...	13
1.7 Sampling and Complete Enumeration ✓	16
References ...	20
Complements and Problems ...	21
2. CONCEPTS, DEFINITIONS AND NOTATIONS	23-54
2.1 Unit and Population ...	23
2.2 Population Parameters ...	25
2.2a General Parameters ...	26
2.2b Normal Distribution ...	27
2.2c Gamma Distribution ...	29
2.2d Bivariate Population ...	31
2.3 Sampling Unit ...	31
2.4 Sampling Frame ...	32
2.5 Random Sample ...	36
2.6 Unbiased Estimator ...	38
2.7 Measures of Error ...	40
2.8 Stages of Randomization ...	41
2.9 Non-sampling Errors ...	43
2.10 Efficiency ...	44
2.11 Confidence Interval ...	45
2.12 Interpenetrating Sub-samples ...	47
2.13 Variance and Cost Functions ...	48

CHAPTER		PAGE
2 14 Notations	...	50
References	...	52
Complements and Problems	...	53
3. SIMPLE RANDOM SAMPLING	...	55-94
3 1 Sampling of One Unit	...	55
3 2 Sampling of Two Units	...	58
3 3 Sampling n units with Replacement	...	60
3 3a Expected Value of Sample Mean	...	60
3 3b Variance of Sample Mean	...	61
3 3c An unbiased Estimator of $V(\bar{y})$...	62
3 3d Interpenetrating Sub samples	...	64
3 4 An Alternative Estimator for \bar{Y}	...	65
3 5 Sampling Two Units without Replacement	...	67
3 6 Sampling n Units without Replacement	...	69
3 6a Expectation of Sample Mean	...	70
3 6b Variance of Sample Mean	...	70
3 6c Behaviour of Sampling Error	...	73
3 6d Unbiased Estimator of $V(\bar{y})$...	74
3 7 Estimation of Totals	..	76
3 7a Estimation of Population Total	...	76
3 7b Total and Mean of a Sub population	...	77
3 8 Estimation of Proportion of Units	...	78
3 8a SRS with Replacement	...	79
3 8b SRS without Replacement	...	80
3 8c Number of Units in a Class	...	81
3 9 Confidence Interval	...	81
3 9a Normal Distribution σ Known	...	81
3 9b Normal Distribution σ Unknown	...	82
3 9c Non normal Distribution	...	83
3 9d Confidence Interval for Proportion	...	84
3 10 Pooling of Estimates	...	86
3 10a SRS with Replacement	...	86
3 10b SRS without Replacement	...	87
3 10c Estimation of a Proportion	...	88
References	..	89
Complements and Problems	...	89

CHAPTER		PAGE
4.	SAMPLE SELECTION AND SAMPLE SIZE	95–132
4.1	Illustrative Populations ...	95
4.2	Stability of Coefficient of Variation ...	96
4.3	Behaviour of Sample Size ...	99
4.4	Sampling Error and Efficiency ...	101
4.5	Procedures of Selection ...	103
4.5a	Random Number Tables ...	103
4.5b	Association of One Number ...	104
4.5c	Association of Several Numbers ...	107
4.6	Changes in Sampling Frame ...	110
4.7	Haphazard Selection ...	111
4.7a	Generation of 'Random' Numbers ...	111
4.7b	Use of Colour Chart ...	112
4.7c	Biases in Selection ...	113
4.8	Determination of Sample Size ...	113
4.8a	Fixed Relative Standard Error ...	114
4.8b	Fixed Confidence Interval ...	117
4.8c	Sample Size When $E(L)$ is specified ...	118
4.8d	Prob. $\{L \leqslant 2d\}$ is specified ...	119
4.8e	Fixed $L : \sigma$ Unknown ...	120
4.8f	Cost Aspect ...	121
4.8g	Estimation of a Proportion ...	122
References	...	123
Complements and Problems	...	124
Annexure 4.1 : Village-wise Complete Enumeration Data	...	127
Annexure 4.2 : Strip-wise Data on Volume of Timber	...	131
5.	SYSTEMATIC SAMPLING ...	133–182
5.1	Sampling Procedure ...	133
5.1a	Sampling of Two Units ...	134
5.1b	A Modified Sampling Scheme ...	136
5.1c	Sampling of n Units ...	136
5.2	Circular Systematic Sampling ...	139
5.2a	Sampling of Two Units ...	140
5.2b	Sampling of n Units ...	141
5.3	Use of Fractional Interval ...	141

CHAPTER	PAGE
5 4 Operational Convenience	142
5 5 Sampling Variance	144
5 6 Comparison with SRS	146
5 7 Behaviour of Sampling Variance	148
5 7a Sampling of Strips in a Forest	149
5 7b Systematic Sampling of Villages	150
5 7c Systematic Sampling in Census	151
5 8 Estimation of Sampling Variance	153
5 8a Linear Systematic Sampling	153
5 8b Circular Systematic Sampling	155
5 8c Random Arrangement of Units	156
5 8d Use of Interpenetrating Sub samples	158
5 8e An Illustrative Example	159
5 9 Linear Trend	160
5 9a A Hypothetical Population	162
5 9b End Corrections	163
5 9c Centrally Located Sample	164
5 9d Balanced Systematic Sampling	165
5 9e Sampling for Multiple Characteristics	167
5 9f Illustrative Examples	169
5 10 Periodic Variation	169
5 11 Suitable Arrangement	172
5 12 Distribution of Sample Proportion	172
5 13 Determination of Sample Size	173
5 14 Two dimensional Systematic Sampling	175
References	176
Complements and Problems	177
6 VARYING PROBABILITY SAMPLING	183 232
6 1 Measure of Size of a Unit	183
6 2 Selection of one Unit with PPS	184
6 3 PPS sampling with Replacement	185
6 4 Efficiency of PPS Sampling	186
6 4a Sub units Approach	187
6 4b Linear Relationship	188
6 5 Cost Aspect . . .	190
6 6 Empirical Studies	192

CHAPTER	PAGE
6.7 Gain due to PPS Sampling ...	196
6.8 Area Sampling (A Special Case) ...	197
6.9 An Alternative Estimator ...	199
6.10 Selection Procedures ...	200
6.10a Cumulative Total Method ...	200
6.10b Lahiri's Method ...	202
6.10c Choice of the Methods ...	206
6.10d Area Sampling ...	207
6.10e Selection in Stages ...	208
6.11 Sampling without Replacement ...	208
6.11a PPS Sampling without Replacement ...	209
6.11b Random Group Method ...	214
6.11c PPS Systematic Sampling ...	215
6.11d Probability Proportional to Total Size ...	218
6.11e Comparison of Various Estimators ...	219
6.12 Integration of Surveys ...	220
6.13 Extension of PPS Sampling ...	223
References ...	225
Complements and Problems ...	226
7. STRATIFIED SAMPLING ...	233-292
7.1 Need for Stratification ...	233
7.2 Principle of Stratification ...	235
7.3 Design and Stratification Variable ...	237
7.4 Allocation of Sample Size ...	239
7.4a Variance and Cost Functions ...	239
7.4b Optimum Allocation ...	241
7.4c Proportional Allocation ...	243
7.4d Allocation Proportional to $W_s \bar{Y}_s$...	244
7.4e Allocation Proportional to $W_s R_s$...	245
7.5 Stratified SRS with Replacement ...	246
7.5a Optimum Allocation ...	246
7.5b Proportional Allocation ...	248
7.5c Gain due to Stratification ...	248
7.6 Stratified SRS without Replacement ...	250
7.6a Allocation to Strata ...	251
7.6b Gain due to Stratification ...	252

CHAPTER		PAGE
7.7	Estimation of a Proportion ...	253
7.8	Stratified PPS Sampling ...	255
7.8a	Allocation of Sample Size ...	256
7.8b	Area Sampling for Crop Survey ...	257
7.9	Illustrative Examples ...	258
7.10	Demarcation of Strata ...	261
7.10a	Theoretical Solutions ...	262
7.10b	Effect of using Auxiliary Variable ...	264
7.10c	Normal, Gamma and Beta Distributions ...	266
7.10d	Approximations to OPS... ...	270
7.10e	An Empirical Study ...	272
7.11	Determination of Number of Strata ..	274
7.12	Interpenetrating Sub-samples ...	276
7.13	Multiple Stratification ...	278
7.14	Technique of Post stratification ...	280
7.15	Controlled Selection... ...	282
References	285
Complements and Problems	...	287
8.	CLUSTER SAMPLING ...	293-316
8.1	Need for Cluster Sampling ...	293
8.2	Sampling of Equal Clusters ...	294
8.2a	Sampling of One Cluster ...	295
8.2b	Sampling of n Clusters ...	297
8.2c	Estimation of Efficiency ...	298
8.3	Optimum Cluster Size ...	299
8.3a	Variance Function ...	299
8.3b	Cost Function ...	301
8.3c	Determination of Cluster Size ...	302
8.3d	Use of Other Sampling Schemes ...	304
8.4	Estimation of a Proportion ...	305
8.5	Varying Cluster Size ...	306
8.5a	Simple Random Sampling ...	306
8.5b	Varying Probability Sampling ...	309
8.6	Illustrative Examples ...	311
References	313
Complements and Problems	...	314

CHAPTER		PAGE
✓ 9. MULTI-STAGE SAMPLING 317-360
9.1 Sampling Procedure	...	317
9.2 Estimation and Sampling Variance	...	319
9.3 Two-stage Sampling with SRS	...	322
9.3a SRS WOR at Both the Stages	...	323
9.3b SRSWR and SRS WOR at the Two Stages	...	325
9.3c SRSWR at Both the Stages	...	325
9.3d Estimation of Population Mean	...	326
9.3e Uni-stage and Cluster Sampling	...	326
9.4 Sampling with PPSWR and SRS WOR	...	328
9.5 Variance Function and Its Behaviour	...	329
9.6 Cost Function	...	334
9.7 Optimum Values of n and m	...	335
9.7a Expected Number of Sample SSU's Fixed	...	335
9.7b Cost Fixed	...	336
9.7c Variance Fixed	...	337
9.7d Graphical Method	...	338
9.7e Optimum Size of FSU	...	340
9.8 Efficiency of Two-stage Sampling	...	340
9.8a Behaviour of Efficiency	...	340
9.8b SRS WOR at Both Stages	...	342
9.8c PPSWR and SRS WOR at the Two Stages	...	343
9.9 An Illustrative Example	...	344
9.10 Three-stage Sampling Design	...	345
9.11 Multi-subject Surveys	...	347
9.12 Multi-phase Sampling	...	349
9.13 Composite Sampling Designs	...	352
References	...	353
Complements and Problems	...	354
10. METHOD OF RATIO ESTIMATION	...	361-404
10.1 Need for Ratio Estimation	...	361
10.2 Bias of Ratio Estimator	...	363
10.3 Mean Square Error	...	367
10.4 Ratio Method of Estimation	...	369

CHAPTER	PAGE
10 5 Basic Sampling Schemes	371
10 5a Simple Random Sampling	371
10 5b Systematic Sampling	373
10 5c Varying Probability Sampling	375
10 6 Stratified Sampling	376
10 7 An Empirical Study	378
10 8 Product Method of Estimation	380
10 9 Almost Unbiased Ratio Estimators	381
10 10 Unbiased Ratio Type Estimators	383
10 11 Unbiased Ratio Estimators	386
10 11a SRS without Replacement	387
10 11b Systematic Sampling	387
10 11c Stratified Sampling	387
10 12 Different Types of Ratio Estimators	388
10 12a Stratified Ratio Estimator	389
10 12b Chain Ratio Estimator	390
10 12c Double Ratio Estimator	390
10 12d Multiple Auxiliary Variables	392
10 13 Two phase Sampling	394
10 14 An Empirical Study	396
References	397
Complements and Problems	398
11 DIFFERENCE AND REGRESSION ESTIMATORS	405-424
11 1 Estimating a Difference	405
11 2 Regression Method of Estimation	406
11 3 Bias and Variance	408
11 4 SRS and Stratified SRS	410
11 4a Simple Random Sampling	410
11 4b Stratified SRS WOR	411
11 5 Two phase Sampling	412
11 5a Variance of Estimator	413
11 5b Cost Aspect	414
11 6 Sampling on Successive Occasions	415
11 7 Estimation of Relative Change	417
11 7a Change in a Proportion	417
11 7b Proportion of Common Units	418

CHAPTER		PAGE
11.8	Multiple Regression Estimator	419
References	...	420
	Complements and Problems	421
12.	SELF-WEIGHTING DESIGN	425-448
12.1	Need for Equal Weights	425
12.2	Stratified Uni-stage Sampling	428
12.2a	SRS within Strata	428
12.2b	PPS sampling within Strata	430
12.3	Stratified Two-stage Sampling...	431
12.4	Some Examples	433
12.5	Self-weighting at Tabulation Stage	436
12.5a	Rounding Off to Some Common Weights	436
12.5b	Rounding Off to Average Weights	437
12.5c	Randomized Rounded-off Weights	437
12.5d	Sub-sampling with PPSWR	438
12.5e	Sub-sampling PPS Systematically	438
12.6	An Empirical Study	439
12.6a	Sampling Design of the Survey	439
12.6b	Self-weighting Procedures	440
12.6c	<i>Results of the Empirical Study</i>	441
12.7	Repetition of Sample Observations	442
References	...	444
	Complements and Problems	444
13.	NON-SAMPLING ERRORS	449-480
13.1	Study of Non-sampling Errors	449
13.2	Sources of Non-sampling Errors	450
13.3	Treatment of Non-sampling Errors	452
13.4	Non-sampling Bias	454
13.5	Non-sampling Variance	456
13.6	Simple Random Sampling	457
13.7	Estimation of a Proportion	461
13.8	Cost Function	463
13.9	Non-response Error	463
13.10	Measurement and Control of Errors	466

CHAPTER	PAGE
13 10a Consistency Checks	467
13 10b Sample Check	468
13 10c Post census and Post survey Checks	469
13 10d External Record Check	472
13 10e Statistical Quality Control Techniques	472
13 10f Study of Recall Error	473
13 10g Interpenetrating Sub samples	474
Preferences	476
Complements and Problems	477
14 PLANNING OF SAMPLE SURVEYS	481-510
14 1 Scope	481
14 2 Formulation of Data Requirements	482
14 3 Survey <i>Ad hoc</i> or Repetitive	484
14 4 Method of Data Collection	485
14 4a Observation and Measurement	486
14 4b Personal Interview	486
14 4c Mail Enquiry	487
14 4d Method of Registration	487
14 4e Transcription from Records	488
14 5 Questionnaire <i>versus</i> Schedule	488
14 6 Survey Reference and Reporting Periods	490
14 7 Problem of Sampling Frame	492
14 8 Choice of Sampling Design	493
14 9 Pilot Survey	496
14 10 Field Work	497
14 11 Processing of Survey Data	499
14 11a Summarization of Data	499
14 11b Subject Analysis	499
14 11c Statistical Analysis	499
14 11d Planning of Processing Work	500
14 11e Fractile Graphical Analysis	501
14 12 Preparation of Reports	502
14 13 Integrated Multi subject Survey	504
14 14 Permanent Survey Organization	507
References	509

CHAPTER		PAGE
15. NATIONAL SAMPLE SURVEY 511-562
15.1 Scope of Survey 511
15.1a Historical Note 511
15.1b Subjects of Enquiry 512
15.1c Methods of Enquiry 512
15.1d Responsibility 513
15.1e Multi-subject Survey 513
15.1f Reporting Period 514
15.1g Interpenetrating Sub-samples 515
15.1h Participation of States 515
15.1i Fourteenth Round 516
15.2 Rural Sector 518
15.2a Subject Coverage 518
15.2b Survey Period 521
15.2c Fixation of Work-load 522
15.2d Programme of Work 528
15.2e Sample Design 529
15.2f Allocation of Sample Villages 530
15.2g Stratification 531
15.2h Investigation Zones 535
15.2i Selection of Villages 536
15.2j Interpenetrating Sub-samples 538
15.2k Hamlet-group Selection 540
15.2l Division of Sample Village 541
15.2m Selection of Households 543
15.2n Selection of Plots 545
15.2o Estimation Procedure 547
15.3 Urban Sector 551
15.3a Subject Coverage 551
15.3b Survey Period 552
15.3c Fixation of Work-load 552
15.3d Interpenetrating Sub-sample 553
15.3e Sample Design 554
15.3f Stratification 554
15.3g Allocation of Sample Blocks 555
15.3h Selection of Urban Blocks 555

CHAPTER	PAGE
15 3i Self weighting Design	558
15 3j Selection of Households	560
15 3k Estimation Procedure	561
16 FAMILY LIVING SURVEYS	563-592
16 1 Introduction	563
16 1a Historical Note	563
16 1b Origin of the Surveys	564
16 1c Organizational Set up	565
16 1d Scope of the Surveys	565
16 2 Working Class Survey	566
16 2a Selection of Centres	567
16 2b Fixation of Sample Size	569
16 2c Preliminary Survey	572
16 2d Tenement Sampling Method	574
16 2e Payroll Sampling Method	580
16 3 Middle Class Survey	582
16 3a Selection of Centres	583
16 3b Fixation of Sample Size	585
16 3c Preliminary Survey	586
16 3d Sampling of Blocks	587
16 3e Sampling of Families	588
16 4 Estimation Procedures	588
16 4a Tenement Sampling Centres	588
16 4b Payroll Sampling Centres	590
BIBLIOGRAPHY	593-656
APPENDIX 1 Table of Random Numbers	657
APPENDIX 2 Formulae for Estimators of \bar{Y} and Variance Estimators	659
APPENDIX 3 Solutions to Problems	666
INDEX	677

LIST OF FIGURES

PAGE

1. Behaviour of Sampling Error with increase in Sample Size	12
2. Behaviour of the Sum of Cost of Survey and Loss due to Sampling and Non-sampling Errors with increase in Sample Size	15
3. Curves of Normal Distributions with Mean μ and with Standard Deviations σ_0 , $\sigma_1 (<\sigma_0)$ and $\sigma_2 (>\sigma_0)$	28
4. Specimen Curve of a Gamma Distribution	29
5. Specimen Map of a part of a Village showing Boundaries of Fields	35
6. A Sketch Map showing Boundaries of Blocks in an Urban Area	35
7. Behaviour of Relative Standard Error in Sampling with and without Replacement	73
8. Histograms of Distributions of Sample Mean based on 500 Samples of Sizes 2, 8 and 16	85
9. Diagrammatic Representation of Linear and Circular Systematic Sampling Schemes	140
10. Scatter Diagram of Cultivated Area with Villages in increasing order of Geographical Area	161
11. Scatter Diagram of 1961 Census Population with Villages in increasing order of 1951 Census Population	161
12. A Specimen Population with Regular Periodic Variation	170
13. Histograms of the Sampling Distribution of the Proportion of Cultivators for different Sample Sizes	174
14. A Representation of Two-dimensional Systematic Sampling showing the Selected Grids in the Cells by Dots	175
15. Relationship between Cultivated Area and Geographical Area for the Villages in a Tehsil	194
16. Relationship between 1951 and 1961 Census Population for the Villages in a Tehsil	194
17. A Diagrammatic Representation of Lahiri's Method of Sampling with PPS	204

	PAG
18 Selection of a Field with Probability Proportional to Area from a Map	20
10 A Sketch Map showing Village Boundaries	23
20 Behaviour of variance Function in Two stage Sampling when Cost is Fixed	33
21 Configurational Representation of the Regions of Preference for (i) \hat{R} , \hat{R}_1^* and \hat{R}_2^* and (ii) \hat{P} , \hat{P}_1^* and \hat{P}_2^*	39

Need for Sample Survey

1.1 NEED FOR STATISTICAL INFORMATION

Since the beginning of the twentieth century, the economic and social life of the people and the functional system of industry and business, transport and communications, educational and medical facilities and other activities of the community have undergone substantial changes due to spectacular developments in the fields of science and technology. The primitive community units or groups, producing goods and services out of their own resources to satisfy their own needs, have now given place to more complex ways of living, where the emphasis is on specialization in mass production and utilization of goods and services of a given type with a view to getting the maximum possible benefit per unit of cost. Considerable planning is required in such large-scale projects and any rational decision regarding efficient formulation and execution of suitable plans and projects or an objective assessment of their effectiveness, whether in the field of industry, business or governmental activities, has necessarily to be based on objective data regarding resources and needs. There is, therefore, the need for various types of statistical (quantified) information to be collected and analysed in an objective manner, and presented suitably so as to serve as a sound basis for taking policy decisions in different fields of human activity. In modern times, the primary users of statistical data are the State, industry, business, scientific institutions, public organizations and international agencies.

The concept of the role of the State has changed considerably during the last three or four decades. It is no longer interested only in the maintenance of law and order, but is also increasingly participating in the improvement of the economic and social life of its citizens. Though the State may not be owning many factories and farms, it regulates their output by taking suitable measures such as imposition of import and export restrictions, price controls, taxation and provision of subsidies as it is gradually taking upon itself the responsibility of sound development and maintenance of the economy. The State is also interested in ensuring a continuous progress of the educational system, health services and other welfare activities with a view to achieving better standards of living for its citizens.

To execute its various responsibilities, the State is in need of a variety of information regarding different sectors of the economy, sections of people and geographical regions in the country as well as information on the available resources such as manpower, cultivable land, forests, water, minerals and oil. If the resources were unlimited, the various requirements could easily be satisfied. Planning in such a case would be relatively simple, as it would consist in just providing each one with what he needs in terms of money, material, employment, education and entertainment. But such a situation is only hypothetical as in practice the resources, however large, are limited and the needs are usually not well defined and are elastic.

When the resources of a country fall short of its needs, the problem of efficient planning is not simple and it necessitates optimum utilization of the resources with a view to satisfying as many of the needs as possible in a balanced manner. It is to be noted that the State cannot afford to completely expend all the resources in meeting the present needs, as it must develop the resources for meeting the future needs of the country as well since the aim of a welfare State is the betterment of the social and economic life of its people not only of the present but also of the future generations. For the purpose of proper planning, therefore, fairly detailed data on the available resources and on the needs are to be collected.

For example, the country is in need of data on production and consumption of different types of products to enable it to take objective decisions regarding its import and export policies. Statistical information on the cost of living of different categories of people living in various parts of the country is of importance in shaping its policies in respect of wage and price levels. Similarly, reliable information on the rates of births, deaths and population growth and incidence of diseases, and on present nutritional standard, level of education and living conditions of the people is necessary for planning for the improvement of social and economic life of the people. This gives an indication of the mass of statistical data that would be immensely useful to the country in the maintenance and development of the economic and social welfare of its people.

In case of industry and business, it is important to have statistical information on cost, quantity and value of production, stock and supply position, etc. on the one hand and on consumer needs and preferences, profits and potentialities, etc. on the other hand for proper planning of production and sales campaigns. Data are also needed on the progress of work at different stages of production and sales activities to enable efficient management of the establishment. For instance, an establishment producing consumer goods such as soaps, hair oil, cigarettes, etc. would be in need of adequate information on consumer preferences and demands for such articles and on stock and supply position in their factories and sales offices.

1.2 TYPES OF DATA

The data on needs and resources, that would be required for proper planning and execution of projects and for assessing their effectiveness, may be classified into two groups : (1) *survey data* comprising of data already in existence which are collected and recorded by observation or enquiry, and (2) *experimental data* which can only be obtained through well-designed and controlled statistical experiments. Survey data can further be grouped into three classes : (1a) levels and relationships of characteristics at a point of time or during a period of time,

(1b) knowledge of trends and relationships among characteristics over time, and (1c) data which are fairly stable over time These types of data are briefly described below

(1a) This type consists of knowledge, which is more related to time, such as the levels and relationships of different characteristics Examples of this type of knowledge are data on population, area under different types of utilization, crop production, industrial production, number of medical men, teachers lawyers engineers labourers, etc including their relationships such as population density, crop yield per unit of area etc on the one hand which are the resources, and data on consumer demands and preferences, employment situations level of education, housing conditions, rate of population growth and of incidence of diseases, etc on the other hand, which are factors contributing to the needs to be satisfied As this type of information is likely to change with time this has to be collected periodically

(1b) The knowledge of trends and relationships of different characters among themselves over time constitutes this type These are to be obtained on the basis of the data to be collected periodically This type of knowledge helps in forecasting the demands and resources in the future as well as the effect of certain government policies on the economic and social life of the people For instance, the relationship between social and economic factors and living conditions of the people will be useful in forecasting the demands at some future date for various items such as cereals milk and milk products, fuel, cloth building materials educational institutions, medical facilities, etc

(1c) The information on the geography and geology of different parts of the country such as natural relief climatic conditions, soil type, mineral and oil deposits belongs to that type, where the facts are not likely to change much over time For this type of knowledge, intensive work by experts in these fields is necessary and this work may be time consuming and costly But once this information is obtained by intensive work it need not be taken up again for some time to come

(2) This type consists of the knowledge that could only be obtained on the basis of well-designed and controlled statistical experiments. Examples of this type are the rate at which the manure should be applied to maximize the yield of a particular crop, the insecticide that is most effective in destroying pests, and the production process in an industry yielding the maximum output per unit of cost. This type also includes data obtained through sustained experiments for establishing or verifying theories of physical sciences.

The methods of obtaining data relating to (1a) and to some extent to (1b) will only be discussed in the subsequent chapters. An indication of the items belonging to these types, the data on which are likely to be of interest to the Government, industry, business and other institutions, is given below :

Population, births, deaths, migration, income and expenditure, employment and unemployment, capital formation, livestock, prices, building construction, housing conditions, education and health statistics, consumer preferences and demands, land utilization, irrigation, agricultural production, cost of cultivation, small, medium and large-scale manufacturing, trade, transport and other enterprises, opinions and attitudes of the people towards certain present or future policies of the Government or towards certain products presently in the market or to be introduced shortly in the market.

The process of planning for economic and social advancement consists in rationally allocating the resources of the country first to the different sectors of the economy such as agriculture, manufacturing industry, trade, transport, housing, health, education and social services, and then within each of these sectors to the administrative divisions such as State, district, tehsil or county, village or parish, town or city. For this purpose, it is necessary to get statistical information regarding the different sectors of the economy for each of the administrative divisions or for groups of such divisions. An example of the situation, where data are obtained for each of the smallest administrative divisions, is provided by the decennial census of population carried out in a number of countries. In case of industry

and business also, region wise information on consumer demands and other relevant factors is required to enable proper planning of the production levels and sales campaigns

13 COMPLETE ENUMERATION SURVEY

One way of obtaining the required information at regional and country levels is to collect the data for each and every unit (person, household, field, factory, shop, etc as the case may be) belonging to the *population* or *universe*, which is the aggregate of all units of a given type under consideration, and this procedure of obtaining information is termed *complete enumeration survey*. The effort, money and time required for carrying out complete enumeration surveys to obtain the different types of data mentioned above will, generally, be extremely large. However, if the information is required for each and every unit in the domain of study, a complete enumeration survey is clearly necessary. But there are many situations, where only summary figures are required for the domain of study as a whole or for groups of units, and in such situations collection of data for every unit is only a means to an end and not the end itself. Hence, before proceeding with data collection by complete enumeration, one has to consider the following points

- (i) Is it the objective to get the data for each and every unit, or only summary information for all units taken together or for groups of units ? ,
- (ii) Is it necessary to obtain exact information, that is, without any error, for the purpose in view ?
- (iii) Does a complete enumeration survey always provide us with accurate information ?
- (iv) Under what circumstances a sample survey, namely, survey of a part of the population, is to be preferred to complete enumeration ?

The first question is of considerable importance as the requirement of data for each and every unit in the population precludes the

possibility of exploring methods of data collection and compilation, which may be more economical and operationally more convenient than the complete enumeration survey. As can easily be seen, data for every individual or unit are necessary in cases where the action is taken separately for each individual or unit. Examples of such situations are income-tax assessment where the income of each individual is assessed and taxed, preparation of voters' list for election purposes and recruitment of personnel in an establishment, or selection from among the applicants for studentship, where an attempt is made to choose the best person on the basis of certain criteria from among a group of persons necessitating study of each and every person in the group.] However, there are numerous situations, where the interest lies solely or mainly in the summary information for all the units taken together, or for groups of units. Since complete enumeration survey involving the collection of data for each and every unit in the population is so conventional and familiar, there may be a tendency to lay down the collection of data for every unit itself as the objective even in cases where the main interest lies only in summary figures. It is necessary to guard against this tendency and to lay down the objective clearly with the ultimate purpose in view.

For answering the second question, one has to examine how exactly the desired statistical information is usually used by the users in taking decisions regarding their activities. It is to be noted that exact planning for the future is not possible, since this would need accurate information on the resources that would be available and on the needs that would have to be satisfied in future. In general, past data are used to forecast the resources and the needs of the future and hence there is some element of uncertainty in planning. Because of this uncertainty, only broad (and not exact) allocations of the resources are usually attempted. Thus some margin of error may be permitted in the data needed for planning, provided this error is not large enough to affect the broad allocations. Moreover, a good deal of uncertainty arising from non-statistical sources involved in dealing with socio-economic problems makes it possible to arrive at only broad (and not exact) decisions and this in its turn makes it plausible to

allow some margin of error in statistical results. Since exact planning is in general impracticable, there is some degree of uncertainty about the outcome of planned projects and hence some margin of error is permissible in the statistical information needed even in cases of assessment of the present and past progress and effectiveness of the projects. The margin of error that is permissible in view of the uncertainties mentioned here may be termed the *permissible error*. This point may be illustrated by the following examples.

In comparing the populations of regions or of countries, we consider usually their populations rounded off to multiples of certain convenient numbers such as thousand or million and not their exact populations up to the last person. For instance, the population of India in 1951 may be taken as 36 crores or $36(10)^7$ and that of Japan in 1950 as 83 millions for practical purposes instead of the exact figures 356879394 and 83199637 respectively. This would mean that an error of ± 0.5 crore ($\pm 1.4\%$) in reporting the population of India and of ± 0.5 million ($\pm 0.6\%$) in case of the population of Japan are permissible for certain purposes.

Suppose in a particular country, the consumption of food grains in the current year is expected to be about 100 million tons and that it is found on the basis of past experience that there is no serious problem of shortage or surplus if the amount of food grains made available for consumption is between 98 and 102 million tons. In such a case it would be possible to decide as to whether any import or export of food grains is necessary and if it is found necessary, how much to import or export provided the production figure is estimated within a margin of error of ± 2 million tons. This shows that the allowable or permissible margin of error for obtaining the production figure in this case is ± 2 million tons.

Usually a certain amount of adjustment is possible in formulation and implementation of decisions relating to socio economic conditions and industrial and business matters. As mentioned above, moderate fluctuations in production of consumer goods from time to time may not appreciably affect the living conditions of the people and it may

not be necessary for the Government to make revisions of its policies in respect of production and distribution of such goods. The margin of adjustment may be taken as an indication of the permissible error for figures of production of these items. However, a proper specification of the allowable margin of error in a particular situation will depend much on the risk involved in taking decisions and on the cost and time required for data collection and supply of final results.

As regards the third question raised above, it may be noted that a complete enumeration survey need not necessarily provide us with accurate information as evidenced by the census experience of a number of countries. For instance, it is estimated that there was a net under-enumeration of 1.1% in the 1951 Indian Population Census (Registrar General, 1953) and of 1.4% in the 1950 Population Census of the United States of America (Eckler, 1953). In fact, these estimates of under-enumeration themselves are believed to be under-estimates. The extent of under- or over-enumeration is likely to be higher for regional and classificatory breakdowns. This shows that even in a population census, where the main aim is just to get a complete count of persons, comparatively an easy task from the view-point of data collection and compilation, the data are subject to some amount of error. From this, the extent of errors that would be present in large-scale complete enumeration surveys involving measurement or observation of certain characteristics can be easily visualized, especially when these characteristics require thorough grasp of difficult concepts and definitions on the part of both the investigators and the informants. The errors in a complete enumeration survey arise mainly from incomplete coverage, observational and tabulation errors due to the difficulties encountered in organizing a survey on such a large scale and in getting adequate trained personnel to carry out the survey. A rather extreme example of gross under-enumeration is provided by the census of industrial and commercial enterprises in the urban areas of France in 1946 which had resulted in about 24% under-enumeration (Chevry, 1949).

14 NEED FOR SAMPLING

Having realized that some margin of error is permissible in the data needed for practical purposes and that complete enumeration surveys need not necessarily provide accurate statistics, we find ourselves faced with the two questions (i) determining the permissible error and (ii) finding an efficient method of survey that would ensure this specified margin of permissible error at the minimum cost. The permissible error is to be determined considering the margin of adjustment possible in the decisions. Usually the permissible error for the larger sections of the universe under study (e.g., larger administrative divisions) will be relatively smaller than that for the smaller divisions. This is due to the fact that at lower levels local knowledge is likely to supplement the survey data to a considerable extent.

Before considering any alternative to a complete enumeration survey for the purpose in view, it may be noted that it is just not practicable in case of destructive surveys, and an example of such a situation is provided by the problem of obtaining the average life of electric bulbs in a batch. In such cases, one has to confine the observations, of necessity, to a part (or a *sample*) of the *population* or *universe*, and to infer about the population as a whole on the basis of the observations on the sample. Even in the other situations, an effective alternative to a complete enumeration survey can be a *sample survey*, where only some of the units selected in a suitable manner from the population are surveyed, and an inference is drawn about the population on the basis of observations made on the selected units. It can be easily seen that compared to a sample survey, a complete enumeration survey is time consuming, expensive, has less scope in the sense of restricted subject coverage and is subject to greater coverage, observational and tabulation errors. In fact, a complete enumeration survey is not at all feasible in situations, where the large amount of resources in terms of trained personnel and finance needed for such a survey is not available. However, since an inference is made about the *whole* from a *part* in a

sample survey, the results are likely to be different from the population values and the differences would depend on the selected part or sample. Thus, it is seen that the information provided by a sample is subject to a kind of error which may be termed *sampling error*. On the other hand, as only a part of the population is to be surveyed, there is greater scope for eliminating the ascertainment or observational errors by proper controls and by employing trained personnel than is possible in a complete enumeration survey.

It is of interest to note that if a sample survey is carried out according to certain specified statistical principles, it is possible not only to estimate the value of the characteristic for the population as a whole on the basis of the sample data, but also to get a valid estimate of the sampling error of the estimate. Here it may be mentioned that the concept of sampling is not of recent origin, since consciously or unconsciously sampling is resorted to in everyday life from time immemorial in making decisions. For instance, the trader examines samples of grains taken from sacks of grains to determine the quality of the whole stock and the housewife examines a spoonful of the dish she has prepared to determine whether it has been properly cooked. However, the development of the theory of sampling together with its proper application to practical problems is of fairly recent origin. A comprehensive historical review of the development of sampling theory is given by Seng (1951).

1.5 SAMPLING AND NON-SAMPLING ERRORS

To appreciate the need for sample surveys, it is necessary to understand clearly the role of sampling and other errors in complete enumeration and sample surveys. As mentioned before, the error arising due to drawing inferences about the population on the basis of observations on a part (sample) of it is termed *sampling error*. Clearly the sampling error in this sense is non-existent in a complete enumeration survey, since the whole population is surveyed. However, the errors mainly arising at the stages of ascertainment and processing of data, which are termed *non-sampling errors*, are common both to complete enumeration and sample surveys.

The sampling error usually decreases with increase in sample size (number of units selected in the sample) and in fact in many situations the decrease is inversely proportional to the square root of the sample size (cf Figure 1.1). From this it is seen that though the reduction in sampling error is substantial for initial increases in sample size it becomes marginal after a certain stage. In other words considerably greater effort is needed after a certain stage to decrease the sampling error than in the initial instances. Hence, after that stage sizable reduction in cost can be achieved by lowering even slightly the precision required. From this point of view, there is a strong case for resorting to a sample survey to provide estimates within permissible margins of error instead of a complete enumeration survey as in the latter the effort and hence the cost needed will be substantially higher due to the attempt to reduce the sampling error to zero.

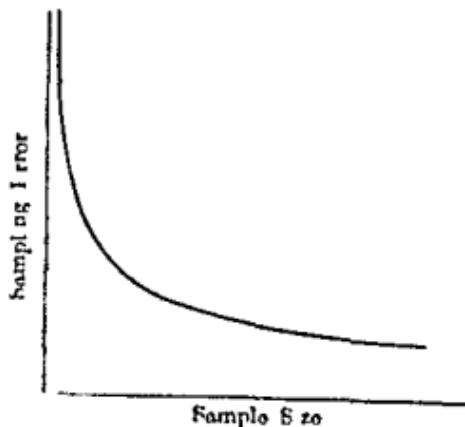


Figure 1.1 Behaviour of sampling error with increase in sample size

As regards the non sampling error, it can be easily seen that it is likely to be more in the case of a complete enumeration survey than in the case of a sample survey since it is possible to reduce the non sampling error to a greater extent by using better organization and suitably trained personnel at the field and tabulation stages in the latter than in the former. The behaviour of the non sampling error with increase in sample size is likely to be the opposite of that of

sampling error. That is, the non-sampling error is likely to increase with increase in sample size. In many situations, it is quite possible that the non-sampling error in a complete enumeration survey is greater than both the sampling and non-sampling errors taken together in a sample survey, and naturally in such situations, the latter is to be preferred to the former. A detailed discussion on the sources of non-sampling errors and on the techniques of assessing and controlling such errors in statistical surveys is given in Chapter 13.

1.6 COST ASPECT

Having noted that while only the results based on sample surveys are subject to sampling errors, non-sampling errors are usually present, though in varying degrees, in the results of both complete enumeration and sample surveys, we may proceed to examine the circumstances under which a sample survey is to be preferred to a complete enumeration survey from the points of view of cost and error. Since a complete enumeration survey may be considered as a particular case of a sample survey, where the sample size is equal to the total number of units in the population, a general formulation of the problem of obtaining statistical information can be taken as the evolution of a sampling method and the determination of the sample size (n) such that the sum of the cost of the survey $C(n)$ and the loss involved $L(n)$ in taking decisions on the survey results is the minimum. That is, the problem is to find the value of n which minimizes

$$T(n) = C(n) + L(n). \quad \dots \quad (1.1)$$

It may be noted that the loss $L(n)$ would depend on the margin of error consisting of both sampling and non-sampling errors. The theory of sample surveys deals with the procedures for selecting the units to be included in the sample and for estimating the value of the characteristic for the population as a whole on the basis of the data collected for the units in the sample, and with comparing the efficiencies of these procedures with reference to cost and error.

To fix the ideas, let us suppose the objective is to estimate the total production of cereals in a particular year for a finite population of N farms and that a complete enumeration survey provides results differing from the true value by $E_c\%$ a difference which arises due to the presence of non sampling errors Let $E_s\%$ be a measure of the divergence of the result provided by a suitably chosen sample of n farms from the true value and let this consist of $E_{s1}\%$ (sampling error) and of $E_{s2}\%$ (non sampling error) If the cost of collecting and tabulating data per farm in the census is C_c and that in the sample is C_s , and if the loss incurred is proportional to the error with L as the loss per one percent error, a sample survey is to be preferred to a complete enumeration survey whenever

$$NC_c + LE_c > nC_s + LE_s \quad (1.2)$$

Since it is desirable and feasible to use better trained personnel and to be more careful in data collection and compilation in a sample survey than is possible in a complete enumeration survey, C_s is expected to be greater than C_c , though nC_s is likely to be less than NC_c . Further, it is possible that in many practical situations the sampling and non sampling errors ($E_s\%$) in the sample survey for a suitably determined sample size (n) would be less than the non sampling errors ($E_c\%$) in a complete enumeration survey This shows that there may be many practical situations where a sample survey may be preferred to a complete enumeration survey on a joint consideration of cost and loss due to wrong decisions

Splitting the error $E_s\%$ involved in a sample survey into its components of samplng error ($E_{s1}\%$) and non samplng error ($E_{s2}\%$), the total cost for any given sample size may be written as

$$nC_s + LE_{s1} + LE_{s2} \quad (1.3)$$

It can be easily visualized that as n increases the components nC_s and LE_{s2} would tend to increase, whereas LE_{s1} is expected to decrease This would mean that the total cost including loss due to wrong decisions is likely to be large for a sample of size

one unit, since in that case the sampling error and hence the loss LE_{s1} would be large and that the total cost $T(n)$ would decrease with increase in n up to a certain stage, after which the cost of data collection and compilation and the loss due to non-sampling errors become large enough to offset the decrease in loss due to diminishing sampling error, finally resulting in an upward trend of the total cost (cf. Figure 1.2). The value of the sample size n , where the total cost reaches its minimum value, is considered as the *optimum sample size*. However, there may be situations, where the cost graph attains the minimum value only when n is almost equal to the total number of units in the population, indicating that in such cases complete enumeration surveys are to be preferred. Such situations, though not very common, may occur when the cost of data collection and compilation per unit is rather small and the possibility of non-sampling errors is negligible.

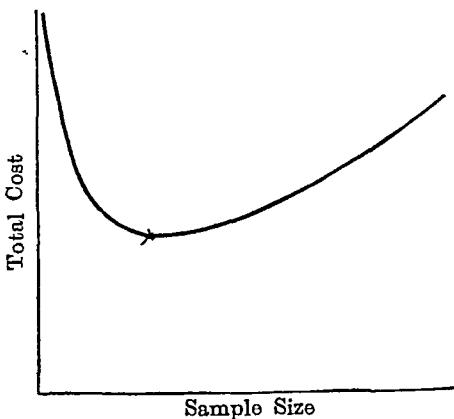


Figure 1.2. Behaviour of the sum of cost of survey and loss due to sampling and non-sampling errors with increase in sample size.

The above method of choosing between the two alternatives, complete enumeration survey and sample survey, or more generally of determining the optimum sample size is usually very difficult in practice, since it may not be easy to determine objectively the form of the loss function $L(n)$ and the values of the constants involved therein. Further, to be more realistic, we have to include in the

criterion denoting total cost other factors such as loss due to delay in supply of the survey results, which is likely to increase with the sample size, being the maximum for a complete enumeration survey. Here again, determination of the value of the loss for different periods of delay in supply of survey results is difficult in practice. However, with a clear conception of the possible actions that are expected to be taken on the basis of the possible survey results together with their financial implications and with extensive studies on cost and error aspects of similar surveys, it should be possible to arrive at the forms of cost and loss functions and the values of the constants involved in them. In practice, such situations are rather uncommon, and usually only an attempt is made to find that combination of selection and estimation procedures and sample size, which would provide estimates with specified margins of error at minimum cost, or alternatively, which would provide estimates with the minimum margins of error for a given cost.

At this stage it may be pointed out that in actual practice a decision on the sampling method and the sample size depends not only on cost and error considerations but also on other factors such as availability of qualified personnel ability on the part of informants to supply the required information, the existing organizational and administrative set up etc. In general, it is seen that cost is not the only constraint on the survey scheme and that feasibility, which depends on factors such as those mentioned above, is of paramount importance to the success of the survey scheme.

17 SAMPLING AND COMPLETE ENUMERATION

The relative merits and demerits of sample surveys have been discussed in detail, among others, by Mahalanobis (1950), Yates (1953), Zarkovich (1961) and Lahiri (1963) and a brief summary of the relevant points is given here.

Complete enumeration and sample surveys presuppose the existence of a certain minimum of facilities, such as funds, professional personnel for planning the survey methodology and supervision of field

operations, sufficiently qualified enumerators or investigators, *sampling frames* such as list of units and maps of area units, machine tabulation equipment, transport and communication facilities, etc. These facilities or combinations thereof do not always exist to the extent needed for a complete enumeration survey and hence in such cases it is impossible to have a complete enumeration survey. Recent experiences have shown that such cases are quite frequent and that sample surveys have been found to be particularly helpful in such situations, since the extent to which the above mentioned facilities are required in the latter case is usually much smaller than in the former.

The implications of the application of these two methods should be clearly understood. In a complete enumeration survey, data can be usefully tabulated for any administrative unit, irrespective of how small it may happen to be, as the results are free from sampling errors, provided of course the non-sampling errors are negligibly small. It will thus be seen that in countries with periodical complete censuses of agriculture, population, livestock, etc., totals for census items are preferred by as small units as villages or communes. In the case of a sample survey, it will not be possible to provide precise estimates for such small units. However, in many practical situations, the statistical information is needed mainly by large regional breakdowns such as provinces, groups of districts or counties and by broad breakdowns of the classificatory characters, and in such cases the sampling method is invariably more efficient.

At this stage, it may be noted that the need for data for very small administrative units may be the result of special needs and conditions in a country's economic and social development. Further, the desire for statistics for very small units in some countries might have arisen due to their using complete enumerations survey for a considerable period of time as the only form of data collection, which automatically provides data for smaller administrative units as a by-product. Experience has shown that if properly planned, a sample survey can give precise enough information on a country's agriculture, industry or any other sector of economic activity by moderately large

administrative or other geographical units and for fairly broad break downs of the classificatory characters. A critical enquiry into the ultimate objectives of the survey would often show that such a picture is satisfactory for most practical purposes. In that case, a sample survey with its many advantages becomes a rational choice.

It may be mentioned that substantial saving in cost of data collection and compilation can be achieved through the use of sample surveys by relaxing the need for statistics for very small administrative units and by not insisting on too precise estimates. The savings in cost so achieved can be utilized to collect data on other items of information. Thus the scope for having an enlarged subject coverage is considerably greater in a sample survey than in a complete enumeration survey.

A sample survey may also become a necessity in dealing with characteristics where serious biases or non sampling errors are expected, when special precautionary measures cannot be taken during collection and tabulation of data. The quality of data in a complete enumeration survey depends upon a large number of enumerators or investigators who cannot be given an intensive training because of cost and organizational difficulties involved. For that reason, more complex methods and costly measurements intended to reduce biases in data collection cannot be applied in complete enumeration surveys. Further, careful scrutiny and inspection at all stages of work will be more manageable and less expensive in a sample survey than in a complete enumeration survey. A convenient method of reducing costs and complications might be a combination of the complete enumeration survey and a sample survey, where the former is used to get information on a few items of basic importance and the latter on all other relevant items required for the purposes in view. If this combination is used the method of complete enumeration is reserved for a few simple items to be tabulated by small administrative units, while the sample survey is taken up for all other items for which estimates by larger regions are sufficient.

To sum up, the sample survey is less time-consuming, costs less, has greater operational flexibility and greater scope in subject coverage as compared to a complete enumeration survey. As regards error, sampling error is present in the results due to the fact that only a part of the whole is surveyed. The non-sampling error here is likely to be smaller than in a complete enumeration survey. The average sampling error in the data based on a sample survey can also be ascertained from the sample itself by adopting suitable sampling techniques.

It is usually possible to determine in advance, at least approximately, the likely error in an estimate from a properly designed sample for any given cost and vice-versa. Experience in India and other countries has shown that the cost of obtaining statistical data with specified permissible margins of error for making rational decisions is considerably less in a properly designed sample survey than in a complete enumeration survey. Even if complete enumeration has to be undertaken in specified situations, sampling techniques can be profitably used in getting advance information well ahead of processing of the complete enumeration material and in assessing the quality of the data provided by complete enumeration surveys.

Fisher (1950) sums up the advantages of the sampling method as compared with the traditional method of complete enumeration as follows :

"I have made four claims for the sampling procedure. About the first three, adaptability, speed and economy, I need say nothing further. Too many examples are already available to show how much the method has to give in these ways. But, why do I say that it is more scientific than the only procedure with which it may sometimes be in competition, the complete enumeration ? The answer, in my view, lies in the primary process of designing and planning an enquiry by sampling. Rooted as it is in the mathematical theory of the errors of random sampling, the idea of precision is from the first in the forefront. The director of the survey plans from the first for a predetermined and known level of precision; it is a consideration of which he never loses sight; and the precision actually attained, subject to well-understood precautions, is manifest from the results of the enquiry."

After considering in detail the relative advantages of complete enumeration and sample surveys, the United Nations Sub Commission on Statistical Sampling (1947) made the following recommendations in their first session itself .

- '(a) It is advisable to consider the desirability of carrying out a sample survey in conjunction with any attempted complete census (especially in the fields of agricultural and population enquiries) with a view (i) to assessing the margin of error, comparative speed, cost and convenience of organization, and (ii) to obtaining supplementary information The cost of such sample surveys will usually be relatively very small
- (b) It is desirable to make a sample survey instead of attempting a complete enumeration whenever adequate funds, physical facilities or personnel of sufficient ability are inadequate for a complete enumeration
- (c) It is sometimes desirable to use the same basic sampling structure for different types of statistics It is also sometimes possible to collect information required in enquiries simultaneously, although the extent to which this can be done is limited by the demands that can be made on the interviewers and respondents and by the fact that different types of surveys may require different interviewers of different technical qualifications
- (d) It is often advantageous to carry out a series of repeated sample surveys at short intervals along with or even instead of a complete census at long intervals In such a series of repeated surveys, it will usually be possible to make appreciable improvements in the sampling technique and thus reduce overall costs as well as to obtain more detailed information and information of better quality with the progress of time
- (e) All these recommendations are subject to this most important provision a sample survey should be carried out only under the technical guidance of professional statisticians not only with adequate knowledge of sampling theory but also with actual experience in sampling practice, and with the help of a properly trained field and computing staff "

REFERENCES

- CHEVREY, G (1949) Control of a general census by means of an area sampling method , *J Amer Stat Assn* , 44, 373-379
- ECKLER, A R (1953) Extent and character of errors in the 1950 census, *American Statistician*, 7, (5), 15-19, 21
- FISHER, R A (1950) The Sub Commission on Statistical Sampling of the United Nations, *Bull Inter Stat Inst* , 32 (2), 207-209,

- LAHIRI, D. B. (1963) : Some thoughts on multi-subject sample survey system; *Contributions to Statistics*, 175-220, Presented to Professor P. C. Mahalanobis on his 70th Birthday, Statistical Publishing Society, Calcutta.
- MAHALANOBIS, P. C. (1950) : Cost and accuracy of results in sampling and complete enumeration; *Bull. Inter. Stat. Inst.*, 32, (2), 210-213.
- REGISTRAR GENERAL (1953) : *Sample Verification of the 1951 Census Count*; Census of India, Paper No. 1, Government of India, New Delhi.
- SENG, Y. P. (1951) : Historical survey of the development of sampling theories and practice; *J. Roy. Stat. Soc.*, (A), 114, 214-231.
- UNITED NATIONS (1947) : *Report of the Sub-Commission on Statistical Sampling to the Statistical Commission*; New York, reprinted in *Sankhyā*, 8, (1948), 393-402.
- YATES, F. (1953) : *Sampling Methods for Censuses and Surveys*; Second Edition, Chapter 1, Charles Griffin & Co., London.
- ZARKOVICH, S. S. (1961) : *Sampling Methods and Censuses*; Volume I, Food and Agricultural Organization of the United Nations, Rome.

COMPLEMENTS AND PROBLEMS

1.1 Describe the advantages of carrying out a sample survey in preference to a complete enumeration survey. Under what circumstances can complete enumeration be recommended in preference to a sample survey ?

1.2 Discuss critically the following statement :

“Will it (the survey) include the entire community or merely a sample? If funds and enumerators are available, we may make a complete enumeration. Often we must be satisfied with a sample.”

1.3 Discuss the scope for using sampling techniques in a complete enumeration survey such as a periodical population census in a country.

1.4 Indicate clearly with suitable illustrations how the term permissible error can be explained to an agency interested in getting statistical data on a specified topic. What would be the implications of over- or under-specification of the permissible error in respect of the cost of the survey and the actions to be based on the survey results?

1.5 In which of the following situations will a sample survey be preferred to complete enumeration on *a priori* considerations? Give reasons

- (i) study of nutrient content of food consumed by persons residing in a city,
- (ii) detection for treatment of persons affected by trachoma, an eye disease, in a given region,
- (iii) determination of production of food grains in the country during the current agricultural season
- (iv) promotion of family planning among families having more than three children,
- (v) study of incidence of tuberculosis and lung cancer in the rural areas of a country,
- (vi) measuring the volume of timber available in a forest,
- (vii) study of trade margin among retail traders in the case of a commonly consumed commodity,
- (viii) study of the birth rate in a region,
- (ix) election for a political office with adult franchise,
- (x) determination of the average life of a batch of electric bulbs,
- (xi) survey of preferences of radio listeners in a region,
- (xii) study of direction and extent of internal migration in a country

Concepts, Definitions and Notations

In this chapter, the basic concepts and definitions of the theory of sampling and estimation are discussed, as a thorough grasp of these ideas is necessary to enable a clear understanding and proper appreciation of the sampling techniques considered in the subsequent chapters. The notations used in the subsequent chapters are also described here.

2.1 UNIT AND POPULATION

An *elementary unit*, or simply a *unit*, is an element or a group of elements, on which observations can be made or from which the required statistical information can be ascertained according to a well-defined procedure. To enable collection of data in a complete enumeration survey or in a sample survey, it is necessary that the units should be well defined and that it should be possible to identify them physically. Examples of unit are person, family, household, farm, factory, retail store, tree, bird, automobile, a period of time such as an hour, day, etc., and the type of unit to be considered would naturally depend on the purpose in view. For instance, a family or household may be considered as the unit in a family budget enquiry, whereas the unit may be a farm or plot (parcel of land) in a crop survey. Similarly, in an industrial survey a factory may be taken as the unit, whereas a shop may be the unit in a retail stores survey. A *reporting unit* is the unit which actually supplies the required statistical information or from which the information can be conveniently

ascertained and such a unit may be the elementary unit itself or a unit representing a group of elementary units. For instance, the head of a family may be the reporting unit supplying information on individual members of the family or it may be the parent office of a chain of establishments reporting about its constituents. Further, a *unit of analysis* is the unit used at the stage of tabulation and such a unit may also be the elementary unit itself or a group of elementary units. However, the *unit of analysis* and the *reporting unit* need not necessarily be identical. For instance, the reporting unit may be a household, or more specifically the head of a household, whereas the *unit of analysis* may be the household itself or an individual member of the household.

The collection of all units of a specified type in a given region at a particular point or period of time is termed a *population* or *universe*. Thus, we may consider a population of persons, families, farms, cattle, houses or automobiles in a region or a population of trees or birds in a forest or a population of fish in a tank, etc depending on the nature of data required. The term population can also be applied to the collection of observations at different points of time for a unit or group of units in a particular region. A population is said to be a *finite population* or an *infinite population* according as the number of units in it is finite or infinite. Populations consisting of a mass of matter which is not made up of easily identifiable and naturally formed elementary units or groups of such units, are termed *continuous populations*. Such populations will have to be artificially sub-divided into suitable elementary units for purposes of observation and experimentation. Examples of continuous populations are water in a tank and a sheet of metal. The *artificial units* in these two cases may be columns of water standing on areas of specified dimensions into which the surface of water can be sub divided and pieces of metal in area-divisions into which the sheet of metal can be sub-divided. Since in most of the survey situations commonly met with in practice we would be mainly concerned with finite populations, our attention would be confined only to finite populations in what follows.

Parts or segments of a population, for each of which the required statistical information is to be obtained separately for the specific purpose in view, are considered to constitute the *domains of study* or *sub-populations*. Since the results are to be obtained separately for each domain of study, any error in their specification may lead to serious limitations in the interpretation of the survey results, whether it be a complete enumeration survey or a sample survey. For instance, if in a country-wide survey, the interest lies in obtaining some specified statistical information separately for certain regions into which the country is divided, besides obtaining that information for the country as a whole, then the regions are said to form the domains of study. Here, it is necessary to specify the regions, which are the domains of study, clearly without any ambiguity with reference to the ultimate aim in view. Similarly, if it is desired to study the living standards of people having different means of livelihood, such as agriculture, industry, business, services, etc., then these classes of population constitute the domains of study and they should be well defined. The domain of study may also be termed the *target population*.

2.2 POPULATION PARAMETERS

Suppose a finite population consists of the N units U_1, U_2, \dots, U_N and let Y_i be the value of the variable y , the characteristic under study, for the i -th unit U_i , ($i = 1, 2, \dots, N$). For instance, the unit may be a person and the characteristic y may be the age last birthday, the unit may be a household and the variable y may be the total consumer expenditure during a particular month, the unit may be a shop and y may be the turnover of sales during a specified period, or the unit may be a farm and the characteristic may be the area under a particular crop, or the unit may be an hour and the characteristic may be the catch of fish at a fishing centre. Any function of the values of all the population units (or of all the observations constituting a population) is known as a *population parameter* or simply a *parameter*. Some of the important parameters usually required to be estimated in surveys are defined here.

22a GENERAL PARAMETERS

The total of the values $\{Y_i\}$, namely,

$$Y = \sum_{i=1}^N Y_i = Y_1 + Y_2 + \dots + Y_N, \quad (21)$$

and their arithmetic mean,

B2817 K7

$$\bar{Y} = \frac{Y}{N} = \frac{1}{N} \sum_{i=1}^N Y_i, \quad (22)$$

106964

are termed the *population total* and the *population mean* respectively. In case of dichotomization of the units in a population as those belonging to a particular class or possessing a specified characteristic, the value Y_i taken by the i th unit can be taken as 1 if it belongs to the particular class or has the specified characteristic and as 0 otherwise. Then the population total Y becomes the total number of units in the population belonging to the particular class or having the specified characteristic and \bar{Y} is the proportion of such units and it is denoted by P , which stands for *population proportion*.

Since the value of the variable varies from unit to unit, it is of interest to get an idea of the variability involved. A measure of the variability of the values of the units about the population mean is provided by the mean of the squared deviations of the values from the mean. This is known as the *population variance* and is given by

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \bar{Y}^2 \quad (23)$$

In other words, σ^2 is a measure of the inherent heterogeneity in the population. To be comprehensible and easily comparable, the measure of variability is to be considered in relation to the mean \bar{Y} . Since σ^2 and \bar{Y} are not in the same units of measure, the measure of variability is taken as the square root of the population variance and this

is termed the *standard deviation* (σ). The ratio of the standard deviation to the population mean is known as the *coefficient of variation* (C), that is,

$$C(y) = \frac{\sigma}{\bar{Y}}. \quad \dots \quad (2.4)$$

It may be noted that the population coefficient of variation $C(y)$ is a pure number free from units of any measure, and that conventionally it is expressed as a percentage. The square of this measure is called the *relative variance*.

The significance of the measure of variability σ defined above can be better understood by considering the *population frequency distribution*, which is a representation of the frequencies of units in the population according to the values of the variable under consideration. Here we shall consider two types of distributions commonly met with in practice.

2.2b NORMAL DISTRIBUTION

The *Normal Distribution*, which is characterized by the continuous frequency function

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2/\sigma^2}, \quad -\infty < y < +\infty, \quad \dots \quad (2.5).$$

is of considerable use in practice. In this case the variable y is said to be normally distributed with mean μ and variance σ^2 and this situation is symbolized by the expression $y \cap N(\mu, \sigma^2)$. The frequency curve of this distribution is given in Figure 2.1.

Though this distribution is continuous and refers to an infinite population, there are finite populations whose discrete distributions approximate the normal distribution. The population which conforms to this distribution is completely specified by the values of μ and σ , which are the only two distribution parameters in this case.

From Figure 2.1, it may be noted that the distribution would be peaked for small values of σ and flat for large values of σ . Further, the distribution is symmetric about μ . From the shape of the curve, it is clear that the percentage ($\alpha/2$) of units having values greater than μ by a certain quantity ($k\sigma$, say) is equal to the percentage of units having

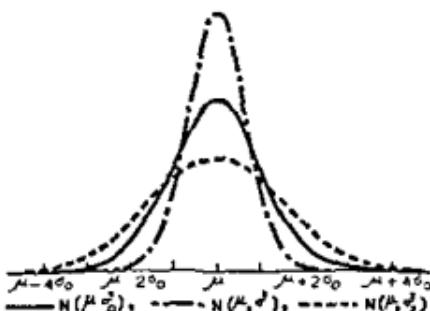


Figure 2.1 Curves of normal distributions with mean μ and with standard deviations σ_0 , σ_1 ($< \sigma_0$) and σ_2 ($> \sigma_0$)

values less than μ by $k\sigma$ and that it decreases with increase in k . The values of $(1-\alpha)/2$ for different values of k have been tabulated (C. R. Rao, Mitra and Matthai, 1966). The value of k corresponding to a specified value of $(\alpha/2)$ is termed the α -percent point of the distribution. The values of k for some values of $P (= 1-\alpha)$, the percentage of units having values between $\mu-k\sigma$ and $\mu+k\sigma$, are given in Table 2.1.

TABLE 2.1 VALUES OF k FOR SOME VALUES OF P FOR A NORMAL DISTRIBUTION.

$P(\%)$	50	80	90	95	99
k	0.674	1.282	1.645	1.960	2.576

2.2c GAMMA DISTRIBUTION

The other distribution, whose frequency function is given by

$$f(y) = \frac{a^p}{\Gamma(p)} e^{-ay} y^{p-1}, \quad 0 \leq y < \infty, \quad \dots \quad (2.6)$$

where $\Gamma(p) = \int_0^\infty e^{-x} x^{p-1} dx$, is known as the *Gamma Distribution*. A number of characteristics of importance in practice, such as income of individual or family, size of land holding and output of industrial establishment, are likely to follow this distribution. The only distribution parameters in this case are a and p , and the mean and the variance of this population are given by p/a and p/a^2 respectively. The frequency curve of this distribution is given in Figure 2.2. From this figure, it can be seen that the distribution is not symmetric and

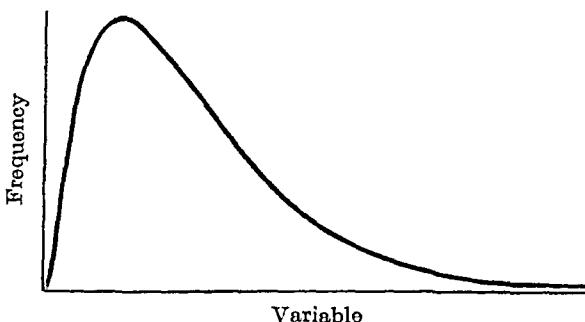


Figure 2.2. Specimen curve of a gamma distribution.

is skew unlike the normal distribution and that the percentages (P) of units having values between $\mu - k\sigma$ and $\mu + k\sigma$ for different values of k are different from those for the normal distribution. From the shape of the curve, it may be noted that populations, where a large number of units have low values for the characteristic under consideration and very few units have very large values, are likely to follow this distribution. A substantial part of the variability between

units in such populations is due to the few units having extremely large values. An example of this type of distribution is given in Table 2.2 and this relates to the distribution of land holdings by holding size. The values of the parameters a and p for this population are 0.0451 and 0.2451 respectively, since the population mean and variance turn out to be 5.4339 and 120.5420 respectively.

TABLE 2.2 DISTRIBUTION OF LAND HOLDINGS BY HOLDING SIZE IN THE RURAL SECTOR OF INDIA . 1954-55.

holding size class (acres)	number of operational holdings (in 1000)	average holding size (acres)
(1)	(2)	(3)
0.00 ¹	6768	—
0.01 — 0.99	19224	0.210
1.00 — 2.49	8692	1.690
2.50 — 4.99	9315	3.165
5.00 — 7.49	5415	6.123
7.50 — 9.99	3356	8.682
10.00 — 14.99	3443	12.176
15.00 — 19.99	1925	17.290
20.00 — 24.99	1032	22.213
25.00 — 29.99	709	27.402
30.00 — 49.99	1215	37.984
50.00 & above	686	83.534
all classes	61780	5.434

¹ holdings of size less than 0.005 acre are shown in this class.

1 acre = 0.4047 hectare

Source : National Sample Survey (1960) : *Report on Land Holdings, 1954-55*, (2), Report No 30, p 13, Cabinet Secretariat, Government of India.

2.2d BIVARIATE POPULATION

In the case of a *bivariate population*, where each unit in the population has two values (Y_i, X_i) corresponding to two characteristics y and x , the ratio of the population totals Y and X is termed the *population ratio* (R). In this case, a measure of the linear relationship between y and x is given by the *correlation coefficient* ρ defined as

$$\rho = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}, \quad \dots \quad (2.7)$$

where $\text{Cov}(x,y)$ stands for covariance between x and y , and is given by

$$\text{Cov}(x,y) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

and where σ_x and σ_y denote the standard deviations of the variables x and y respectively. The linear relationship is denoted by the form

$$y = \alpha + \beta x, \quad \dots \quad (2.8)$$

where $\alpha = \bar{Y} - \beta \bar{X}$ and $\beta (= \rho \sigma_y / \sigma_x)$ is termed the *regression coefficient*.

2.3 SAMPLING UNIT

Elementary units or groups of such units, which, besides being clearly defined, identifiable and observable, are convenient for purposes of sampling, are called *sampling units*. For instance, in a family budget enquiry, usually a family is considered as the sampling unit, since it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit. The sampling units, whether elementary units themselves or groups of such units, should also be well defined and

identifiable. However, the sampling units need not necessarily be non overlapping, though in a number of situations they are usually mutually exclusive. Examples of overlapping sampling units are provided by areas of given shape and size which are located in crop fields for carrying out crop yield surveys. This area, which may be a rectangle, circle or triangle of specified dimensions, is the sampling unit, since it can be considered to be groups of basic cells of unit area, which constitute the elementary units in this case. The sampling units should be so specified that each and every elementary unit in the population under consideration occurs at least in one sampling unit, as otherwise some elementary units will not have a chance of being selected at all in any sample. The collection of all sampling units of a specified type constitutes a population of sampling units.

2.4 SAMPLING FRAME

For using sampling methods in the collection of data, it is essential to have a *frame* of all the sampling units belonging to the population to be studied with their proper identification particulars and such a frame is termed the *sampling frame*. This may be a list of units with their identification particulars or a map showing the boundaries of the sampling units. As the sampling frame forms the basic material from which a sample is drawn, it should be ensured that the frame contains all the sampling units of the population under consideration but excludes units of any other population. The frame should be up to date, and free from errors of omission and duplication of sampling units. In some cases, the sampling frame may include, besides the identification particulars, some auxiliary information, which can be utilized for increasing the efficiency of selection of sampling units and of inference based on the sample observations. For instance, in the 1961 census list of villages available in India, which serves as the rural sampling frame, village wise data regarding area, population and distribution of workers by different means of livelihood classes, such as cultivators, workers at household industry, etc. are given besides names of the villages.

It may be pointed out that though the sampling frame may not contain in some cases the full identification particulars of all the sampling units, it should be possible to collect detailed identification particulars at least for the selected units, as otherwise it would not be possible to locate the selected units and ascertain the required information from them. An example of this situation is provided by the rural sampling frame in India mentioned above, which consists of a list of villages with their names and the exact particulars of location of the villages are available only in the regional administrative offices. Further, in some situations it is possible that only a frame of groups of sampling units is available. In such situations, the frame can be considered useful for drawing a sample of sampling units, only if it is practicable to prepare a frame of all sampling units at least in the groups of sampling units selected from the basic frame. For instance, if household is the sampling unit and only a frame of villages or segments (well-defined area units consisting of groups of households) is available, then it is possible to make use of this frame, provided a frame of all the households can be prepared in the selected villages or segments. It may also be noted that in the case of overlapping sampling units the frame may consist of only the population of the elementary units with a clear specification of the procedure of forming the sampling units. In the case of continuous populations, the frame may just consist of the location specifications of the population together with the procedure of forming the artificial units meant to be used as sampling units. Specimens of sampling frames are given in Table 2.3 and in Figures 2.3 and 2.4. The first example relates to a frame of villages with some auxiliary information provided by the 1961 Census of India. Figure 2.3 shows a sampling frame of fields (or plots) in a village, which is usually used for selection of fields for crop survey. In Figure 2.4 is given a frame for an urban area showing the boundaries of the blocks into which that area has been divided such that each block has about 150 households.

From what has been said above, it may be noted that sampling frames may broadly be divided into two groups : (i) list of sampling units and (ii) maps showing boundaries of area units. The former

TABLE 2.3 SPECIMEN SAMPLING FRAME FOR THE RURAL SECTOR IN INDIA SHOWING THE MAJOR ITEMS OF AVAILABLE INFORMATION.

State—Punjab		District—Hoshiarpur						Tehsil—Dasuya			
sl. no.	name of village	area in acres	number of houses		number of persons		literate		workers		
			house	holds	total	males females	males	females	males	females	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
51	Masut Palkot	663	232	234	1318	658	660	313	190	275	5
52	Mohan	263	54	54	302	158	144	37	9	82	2
53	Nagal Khungan	469	136	136	679	365	314	165	65	181	-
54	Malhapur Bodil	266	47	47	249	145	104	35	3	72	-
55	Salahpur	161	34	34	212	101	111	56	22	46	-
56	Gahot	146	52	54	251	116	135	72	50	54	5
57	Lodichak	381	144	138	643	308	335	163	94	136	4
58	Rajpur	155	107	107	602	317	285	182	110	131	23
59	Kotli	165	41	41	228	119	109	31	4	60	1
60	Jhappind	247	37	37	255	118	137	71	59	59	3
61	Chattowal	219	71	72	430	208	222	90	63	96	8
62	Sohian	246	93	93	506	249	257	117	59	134	3
63	Dargahert	233	25	26	123	68	55	31	13	34	7
64	Korala Kalan	875	215	215	1234	639	595	345	188	299	29
65	Kaala Kalan	327	—	—	uninhabited	—	—	—	—	—	—
66	Bagol Kalan	298	17	17	100	46	54	15	6	30	-
67	Bagol Khurd	333	32	32	139	87	52	13	54	54	2
68	Tilluwal	169	58	58	344	182	162	73	35	81	31
69	Khun Khun Kalan	640	287	287	1444	750	694	357	168	369	9

Source : Census of India (1961) Primary Census Abstracts of Hoshiarpur District in Punjab State

frame, which may be termed *list frame*, consists of a list of sampling units with their proper identification particulars and in many cases this frame may also contain some auxiliary information on the sampling units. In this category may be included all types of frames, where the sampling units may be selected and identified uniquely without necessarily referring to a map and where identification of area sampling units in the field is not involved. On the other hand, the latter type of frame, which may be termed *area or map frame*, shows geographical boundaries of sampling units or groups of sampling units which are generally area units necessitating reference to

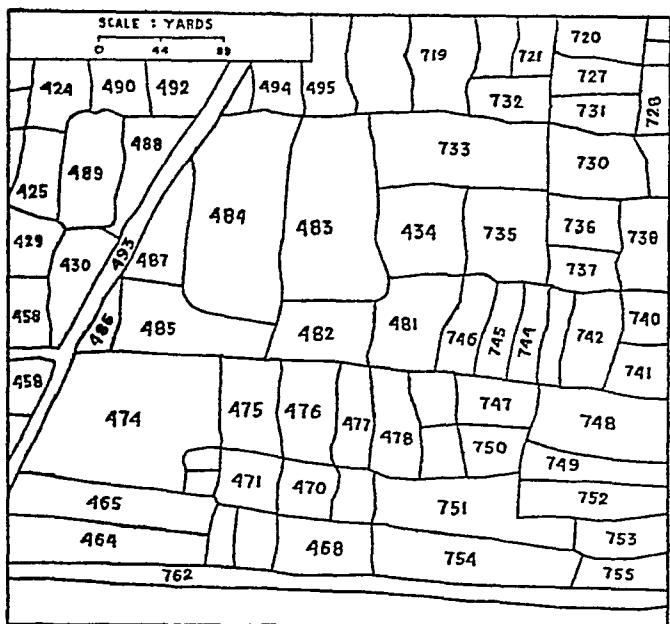


Figure 2.3. Specimen map of a part of a village showing boundaries of fields.

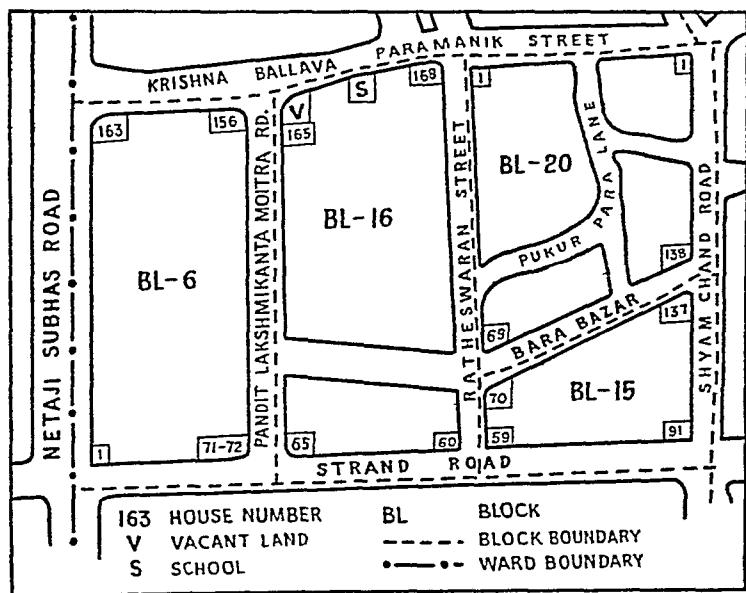


Figure 2.4. A sketch map showing boundaries of blocks in an urban area.

the map for proper identification of the boundaries of the selected sampling units Different aspects of sampling frame have been discussed, among others, by Mahalanobis (1944) Yates (1953), Seal (1962) and Hansen, Hurwitz and Jribine (1963)

25 RANDOM SAMPLE

One or more sampling units selected from a population according to some specified procedure are said to constitute a *sample*. The sample will be considered as *random* or *probability sample*, if its selection is governed by ascertainable laws of chance. In other words, a random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection. For example if a population consists of the N sampling units

$$U_1, U_2, \dots, U_n, \dots, U_N,$$

then we may select a sample of n units by selecting them unit by unit with equal probability for every unit at each draw with or without replacing the sampling units selected in the previous draws. Cochran, Mosteller and Tukey (1954) have defined a *probability sampling scheme* as one in which

- (i) each individual in the sampled population has a known probability of entering the sample,
- (ii) the sample is chosen by a process involving one or more steps of automatic randomization consistent with these probabilities and
- (iii) in the analysis of the sample, weights appropriate to the probabilities (i) are used”

It is to be noted that there are a number of samples of a specified type that can be formed by grouping units of a given population and that in probability or random sampling one of these samples is selected with specified probabilities. A clear specification of all possible samples of a given type with their corresponding probabilities

is said to constitute a *sample design*. In the example of selecting a sample of n units with equal probability with replacement mentioned earlier, the sample design consists of N^n possible samples (taking into account the orders of selection and repetitions of units in the sample) with $1/N^n$ as the probability of selection for each of them, since in each of the n draws any one of the N units may get selected. Similarly, in sampling n units with equal probability without replacement, the number of possible samples (ignoring orders of selection of units) is $\binom{N}{n}$ and the probability of selecting each of the samples is $1/\binom{N}{n}$.

The sampling units selected in the sample may be termed *sample units* and the values of the characteristics under study for the sample units are known as *sample observations*. The number of sampling units selected in a sample is termed *sample size (n)*, and the ratio of sample size to total number of population units is termed *sampling fraction (f)*.

A sample selected by a non-random process is termed *non-random sample*. A non-random sample, which is drawn using certain amount of judgement with a view to getting a representative sample, is termed *judgement* or *purposive sample*. In purposive sampling, units are selected by considering the available auxiliary information more or less subjectively with a view to ensuring a reflection of the population in the sample. This type of sampling is seldom used in large-scale surveys mainly because it is not generally possible to get strictly valid estimates of the parameters under consideration and of their sampling errors due to the risk of bias in subjective selection and the lack of information on the probabilities of selection of the units. Hence, the subsequent discussions are confined to random or probability sampling.

In the subsequent chapters we shall consider a number of random processes or probability schemes, ordinarily used in selecting a sample and make a comparative study of their relative merits and demerits in drawing inferences about the population on the basis of the sample observations.

The recent developments in sampling theory have been reviewed by Sukhatme (1959) Seth (1961) Dilenius (1962) and Murthy (1963a). Attempts have also been made to evolve a unified theory of sampling and the work in this field has been discussed by Godambe (1965) and Hanurav (1966). Koop (1963) and Murthy (1963b) have considered generalized estimators for specified classes of parameters based on any sampling scheme.

26 UNBIASED ESTIMATOR

Suppose a sample of n units is selected from a population of N units according to some probability scheme and let the sample observations be denoted by y_1, y_2, \dots, y_n . Any function of these values which is free from unknown population parameters is called a *statistic*. An *estimator* is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a *random variable* as its value differs from sample to sample and the samples are selected with specified probabilities. The particular value which the estimator takes for a given sample is known as an *estimate*. Let the probability of getting the i th sample be P_i and let t_i be the estimate that is the value of an estimator t of the population parameter θ based on this sample ($i = 1, 2, \dots, M_0$) M_0 being the total number of possible samples for the specified probability scheme. The *expected value* or the average of the estimator t is given by

$$E(t) = \sum_{i=1}^{M_0} t_i P_i \quad (29)$$

since the value t_i of the estimator based on the i th sample occurs with probability P_i . The estimator t is said to be an *unbiased estimator* of the population parameter θ if its expected value is equal to θ , that is

$$E(t) = \sum_{i=1}^{M_0} t_i P_i = \theta, \quad (210)$$

irrespective of the y -values. In case $E(t)$ is not equal to θ , the estimator t is said to be a *biased estimator* of θ and the *bias* of t is given by

$$B(t) = E(t) - \theta. \quad \dots \quad (2.11)$$

The estimator t is said to be positively or negatively biased for θ according as the value of the bias is positive or negative. It is assumed here that the sample observations are free from observational or ascertainment error. The situation, where the sample observations are subject to such errors, is considered briefly in Section 2.9 and in greater detail in Chapter 13.

The estimator t is said to be a *consistent estimator* of the parameter θ if its value *approaches* θ *statistically* as the sample size n is increased in the sense that the probability of the difference $(t - \theta)$ being less than any specified small quantity tends to unity as n is indefinitely increased. When sampling from a finite population of N units, a consistent estimator may be defined as an estimator, which, besides satisfying the above condition as n is increased to N , attains the value of the parameter when all units in the population are included in the sample. In cases where an unbiased estimator does not exist or where a biased estimator is preferred to an unbiased estimator on the basis of certain considerations, it is desirable to use a consistent estimator, as it would ensure getting estimates in the immediate neighbourhood of the population parameter for large samples.

A technique of generating unbiased estimators for a class of parameters based on any sampling scheme has been given by Murthy (1963b). Murthy (1962) has also suggested a procedure for estimating the bias in the case of biased estimators of non-linear parametric functions with a view to obtaining atleast almost unbiased estimators by correcting the former for their bias.

27 MEASURES OF ERROR

Since a probability scheme usually gives rise to different samples the estimates based on the sample observations will, in general, differ from sample to sample and also from the value of the parameter under consideration. The difference between the estimate t_i based on the i -th sample and the parameter, namely $(t_i - \theta)$, may be called the error of the estimate and this error varies from sample to sample. An average measure of the divergence of the different estimates from the true value is given by the expected value of the squared error, which is

$$M(t) = E(t - \theta)^2 = \sum_{i=1}^{M_0} (t_i - \theta)^2 P_i, \quad . \quad (2.12)$$

and this is known as *mean square error* (mse) of the estimator. The mse may be considered to be a measure of the *accuracy* with which the estimator t estimates the parameter. The square root of the mse is termed *root mean square error*.

The expected value of the squared deviation of the estimator from its expected value is termed *sampling variance*. It is a measure of the divergence of the estimator from its expected value and is given by

$$V(t) = \sigma^2(t) = E\{t - E(t)\}^2 = E(t)^2 - \{E(t)\}^2 \quad . \quad (2.13)$$

This measure of variability may be termed the *precision* of the estimator t . The relation between mse and sampling variance, or between accuracy and precision, can be obtained by writing the former as

$$\begin{aligned} M(t) &= E\{t - E(t) + E(t) - \theta\}^2 \\ &= E\{t - E(t)\}^2 + \{E(t) - \theta\}^2, \end{aligned}$$

where the cross product term vanishes, since $E\{t - E(t)\} = 0$. Hence,

$$M(t) = V(t) + \{E(t)\}^2, \quad . \quad (2.14)$$

which shows that the mean square error of t is the sum of the sampling variance and the square of the bias. However, if t is an unbiased estimator of θ , the mse and the sampling variance are the same. The square-root of the sampling variance, $\sigma(t)$, is termed the *standard error* of the estimator t . The ratio of the standard error of the estimator to the expected value of the estimator is known as *relative standard error* (rse) or *coefficient of variation* (C) of the estimator and this is conventionally expressed as a percentage. The square of this measure is termed *relative variance* of the estimator.

2.8 STAGES OF RANDOMIZATION

In some cases of random or probability sampling, the sampling scheme may involve two or more stages or phases of randomization. In other words, the sample may be selected in stages according to specified probability schemes at the different stages. One type of such a scheme is where the sampling units at a particular stage are sub-divisions of sampling units at the preceding stage and the sub-divisions are selected from among the sampling units selected in the previous stage and this type of sampling is termed *multi-stage sampling*. For instance, in a crop survey, the villages or some suitably defined areas (groups of fields) may be selected in the first stage using a random process, and in the second stage some fields may be sampled from each of the selected villages or other well-demarcated areas according to a probability scheme. Another type of such a scheme is where a large sample of units is selected in the first stage or phase according to some probability design for collecting data on a few simple items of information and then a sub-sample of this large sample is selected in the next stage or phase for collecting more detailed information. This type of sampling is known as *multi-phase sampling*.

If the sample is selected in two stages or phases, the total number of possible samples (M'_0 , say) will be given by

$$M'_0 = \sum_{i=1}^{M_0} M_i, \quad \dots \quad (2.15)$$

where M_0 is the number of possible samples arising at the first stage of randomization and M_i is the number of possible samples arising at the second stage of randomization given that the i th sample has been selected at the first stage of randomization. The *unconditional probability* P_{ij} of getting the ij th second stage sample is given by

$$P_{ij} = P_i P_{j|i}, \quad (2.16)$$

where P_i is the *unconditional probability* of getting the i th sample at the first stage of randomization and $P_{j|i}$ is the *conditional probability* of getting the ij th sample at the second stage of randomization given that the i th sample has been selected at the first stage of randomization.

Theorem 1 Let t be an estimator of the parameter θ based on a sample drawn with two stages of randomization. The expected value and the variance of the estimator t are given by

$$E(t) = E_1 E_2(t) \quad (2.17)$$

and

$$V(t) = V_1 E_2(t) + E_1 V_2(t) \quad (2.18)$$

where E_1 and V_1 stand for the expected value and variance over the first stage of randomization and E_2 and V_2 stand for the *conditional expected value* and *conditional variance* over the second stage of randomization given that a particular sample has been selected at the first stage of randomization.

Proof Let t_{ij} be the estimate based on the ij th sample. By definition $E(t)$ is given by

$$E(t) = \sum_{i=1}^{M_0} \sum_{j=1}^{M_i} t_{ij} P_{ij} = \sum_{i=1}^{M_0} \sum_{j=1}^{M_i} t_{ij} P_{j|i} P_i$$

since $P_{ij} = P_i P_{j|i}$. That is

$$E(t) = \sum_{i=1}^{M_0} E_2(t) P_i = E_1 E_2(t)$$

since by definition $E_2(t) = \sum_{j=1}^{M_i} t_{ij} P_{j|i}$. Similarly, by definition we have

$$\begin{aligned} V(t) &= E(t-T)^2, \text{ where } T = E(t), \\ &= E_1 E_2 [\{t-E_2(t)\} + \{E_2(t)-T\}]^2 \\ &= E_1 E_2 \{t-E_2(t)\}^2 + E_1 \{E_2(t)-T\}^2, \end{aligned}$$

since $\{E_2(t)-T\}$ is independent of the second stage of randomization and $E_2\{t-E_2(t)\} = 0$. This shows that

$$V(t) = E_1 V_2(t) + V_1 E_2(t).$$

The estimator t will be unbiased if $T = \theta$. If $T \neq \theta$, the estimator t is biased and its mean square error is

$$M(t) = V(t) + \{B(t)\}^2,$$

where $B(t)$ is the bias given by $B(t) = E_1 E_2(t) - \theta$ and $V(t)$ is the variance as given in (2.18).

The expressions for $E(t)$ and $V(t)$ given in (2.17) and (2.18) can easily be generalized to situations where there are more than two stages of randomization. For instance, if there are three stages of randomization, $E(t)$ and $V(t)$ are given by

$$E(t) = E_1 E_2 E_3(t) \quad \dots \quad (2.19)$$

and

$$V(t) = V_1 E_2 E_3(t) + E_1 V_2 E_3(t) + E_1 E_2 V_3(t), \quad \dots \quad (2.20)$$

where E_3 and V_3 are the conditional expected value and variance given that particular samples have been selected in the first two stages of randomization.

2.9 NON-SAMPLING ERRORS

So far we assumed that the sample observations were free from observational, ascertainment and other types of non-sampling errors. However, this is not usually the case in practice and we have to take account of the possibility of non-sampling errors in the observations, which may arise from various sources such as defective sampling frame, ambiguity in definitions and procedures of data collection, and tabulation errors. Taking the trivial case of each unit giving

rise to a constant observational error, it can be seen that even the result of a complete enumeration survey will be biased. In actual practice the extent of non sampling error will depend on a variety of factors and may change from investigator to investigator, from informant to informant from tabulator to tabulator, etc. In such a situation, the average of the non sampling errors over all possible observations on the units is termed the *non sampling bias* and the variance of the non sampling errors over all possible observations on units is regarded as the *non sampling variance*. From this, it may be noted that the mean square error of an estimator as defined earlier consists of not only bias and variance arising due to sampling of units but also includes non sampling bias and non sampling variance. A detailed discussion of this problem is given in Chapter 13.

2.10 EFFICIENCY

Given two estimators t_1 and t_2 of the population parameter θ , the estimator t_1 is said to be more *efficient* than t_2 if its *mse* is less than that of t_2 , that is, $M(t_1) < M(t_2)$. The estimator t_1 is said to be more *precise* than t_2 if its variance is less than that of t_2 , that is, $V(t_1) < V(t_2)$. The *information* supplied by t_1 is measured by the inverse of its *mse*. The *sampling efficiency* or simply *efficiency*, E_e , of t_1 compared to that of t_2 is defined as

$$E_e(t_1 | t_2) = M(t_2)/M(t_1), \quad (2.21)$$

which is the ratio of the amounts of information supplied by them. The *precision* of t_1 compared to that of t_2 is defined as $V(t_2)/V(t_1)$ and for unbiased estimators, the precision and the efficiency are the same. If $C(t_1)$ and $C(t_2)$ are the cost of survey operations leading to the estimators t_1 and t_2 respectively, then the *cost efficiency*, E_c , of t_1 compared to that of t_2 is defined as

$$E_c(t_1 | t_2) = M(t_2)C(t_2)/M(t_1)C(t_1), \quad (2.22)$$

which is the ratio of the amounts of information per unit of cost in the two cases.

2.11 CONFIDENCE INTERVAL

The frequency distribution of the samples according to the values of the estimator t based on them (that is, the sample estimates) is termed the *sampling distribution* of the estimator t . It is important to note that though the population distribution may not be normal, the sampling distribution of the estimator t is usually close to normal, provided the sample size is sufficiently large. If the estimator t is unbiased and is normally distributed, the interval $\{t - k\sigma(t), t + k\sigma(t)\}$ is expected to include the parameter θ in $P\%$ of the cases, where P is the proportion of the area between $-k$ and $+k$ of the distribution of the *normal deviate* $(t - \theta)/\sigma(t)$ having mean 0 and standard deviation 1. This means that if a large number of samples are drawn and the limits $t - k\sigma(t)$ and $t + k\sigma(t)$ are calculated, then the parameter θ will be between these limits in about $P\%$ of the cases. The interval considered here is said to be a *confidence interval* for the parameter θ with a *confidence coefficient* of $P\%$ with the *confidence limits* $t - k\sigma(t)$ and $t + k\sigma(t)$. The values of the confidence coefficient P commonly used together with the corresponding values of k are the same as those given earlier in Table 2.1 while considering the normal distribution.

It is of interest to note that in case the estimator t is biased, the confidence coefficient P associated with a specified confidence interval $t \pm k\sigma(t)$ decreases with increase in the absolute value of the bias in the estimator t , $|E(t) - \theta|$, provided the estimator t is normally distributed. Cochran (1963) has shown that the decrease in confidence coefficient P is only marginal for increases even up to 20% in the ratio of the bias $B(t)$ to the standard error $\sigma(t)$, and that for increases of this ratio beyond this point, the decrease in the confidence coefficient becomes marked.

In practice, the actual value of $\sigma(t)$, the standard error of t , is not generally known and hence it is usually estimated from the sample itself. If the estimator t is unbiased for θ and if its sampling distribution is close to normal, then the distribution of the statistic

$t' = (t - \theta)/s(t)$ where $s(t)$ is the estimated standard error of t , follows the *Student's t Distribution*, which has the frequency function

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma(\frac{1}{2}) \Gamma\left(\frac{v}{2}\right)} \frac{1}{\sqrt{v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty, \quad (2.23)$$

where v is the number of degrees of freedom on which the estimate of the standard error is based. The values of k for different values of P , the confidence coefficient, and of v the degrees of freedom, have been tabulated for the t distribution (C R Rao, Mitra and Matthai, 1966) and making use of this information confidence intervals can be set up as mentioned earlier. The values of k for some values of P and v are given in Table 2.4. If v is large, the distribution of t can be considered to be normal for all practical purposes.

TABLE 2.4 THE VALUES OF k FOR SOME VALUES OF P AND v

v	values of k when P (%) is				
	50	80	90	95	99
(1)	(2)	(3)	(4)	(5)	(6)
1	1 000	3 078	6 314	12 706	63 657
2	0 816	1 886	2 920	4 303	9 925
3	0 705	1 638	2 353	3 182	5 841
4	0 741	1 533	2 132	2 776	4 604
5	0 727	1 476	2 015	2 571	4 032
6	0 718	1 440	1 943	2 447	3 707
7	0 711	1 415	1 895	2 365	3 499
8	0 706	1 397	1 860	2 306	3 355
9	0 703	1 383	1 833	2 262	3 250
10	0 700	1 372	1 812	2 228	3 169
11	0 697	1 363	1 796	2 201	3 106
12	0 695	1 356	1 782	2 179	3 055
13	0 694	1 350	1 771	2 160	3 012
14	0 692	1 345	1 761	2 145	2 977
15	0 691	1 341	1 753	2 131	2 947
16	0 690	1 337	1 746	2 120	2 921
17	0 689	1 333	1 740	2 110	2 893
18	0 688	1 330	1 734	2 101	2 878
19	0 688	1 328	1 729	2 093	2 861
20	0 687	1 325	1 725	2 086	2 845
∞	0 674	1 282	1 645	1 960	2 578

2.12 INTERPENETRATING SUB-SAMPLES

If a sample is selected in the form of two or more sub-samples drawn according to the same sampling scheme such that each sub-sample provides a valid estimate of the parameter under consideration, the sub-samples so drawn are called *interpenetrating sub-samples*. The sub-samples may or may not be independently selected. One main advantage of having independent interpenetrating sub-samples is that it is possible to get easily an unbiased estimate of the variance of the estimator, even if the sample design is a complex one and the expressions for the variance of the estimator and for the usual estimator of this variance are complicated. Suppose t_1, t_2, \dots, t_k are unbiased estimates of the parameter θ based on k independent interpenetrating sub-samples. Then an unbiased estimator of θ based on all the sub-samples is given by the mean of the sub-sample estimates, namely,

$$\hat{\theta} = \bar{t} = \frac{1}{k} \sum_{i=1}^k t_i. \quad \dots \quad (2.24)$$

Theorem 2. An unbiased estimator of $V(\bar{t})$ is provided by

$$v(\bar{t}) = \frac{1}{k(k-1)} \sum_{i=1}^k (t_i - \bar{t})^2. \quad \dots \quad (2.25)$$

$$\begin{aligned} \text{Proof : } E\{v(\bar{t})\} &= \frac{1}{k(k-1)} E\left\{ \sum_{i=1}^k t_i^2 - k\bar{t}^2 \right\} \\ &= \frac{1}{k(k-1)} \left[\sum_{i=1}^k \{V(t_i) + \theta^2\} - k\{V(\bar{t}) + \theta^2\} \right], \end{aligned}$$

since $V(t_i) = E(t_i^2) - \theta^2$ and $V(\bar{t}) = E(\bar{t}^2) - \theta^2$. Simplifying, we get $E\{v(\bar{t})\} = V(\bar{t})$.

Further, it may be pointed out that if the estimator is unbiased and symmetrically distributed, then the probability of the parameter lying between the minimum and the maximum of the k estimates is $1 - (\frac{1}{2})^{k-1}$, thus readily providing a confidence interval.

The technique of interpenetrating sub samples was originally developed by Mahalanobis (1946) during the 1930's and since then this has been increasingly used in surveys for assessing sampling and non sampling errors. This technique can, in general, be used to study the differential effects of different methods and procedures by allotting two or more interpenetrating sub samples to different teams of workers at both the field and tabulation stages. The results based on interpenetrating sub samples, surveyed and analysed by different agencies or teams of workers following the same concepts, definitions and procedures, provide a valuable check on their reliability. Deming (1960) has extensively used this technique for obtaining quick estimates of sampling variance in the case of many sampling schemes.

2.13 VARIANCE AND COST FUNCTIONS

For designing a sample survey efficiently, it is essential to have some broad information on the variability in the population and on the cost of different steps in carrying out the survey. A measure of the error in the results of the survey is given by the mean square error of the estimator used which reduces to its variance in the case of an unbiased estimator. As mentioned earlier, the variance of the estimator usually decreases with increase in sample size, while the cost of the survey increases with increase in sample size. Further, variance and cost would also depend on the nature of the sampling unit. Hence, for a given sampling design it becomes necessary to take both these aspects into account in arriving at the optimum sampling unit and the optimum sample size n which would provide the maximum information per unit of cost. The different aspects of cost and variance functions have been considered in detail by Mahalanobis (1944) with special reference to crop surveys.

The variance of the estimator will, in general, be a function of the heterogeneity of the elementary units in the population (v , say), the nature of the sampling unit (characterized for simplicity by its

size x , which may be taken as the number of elementary units in it), the probability scheme used in selecting the sample (p , say) and the sample size (n), that is,

$$V(t) = f(v, x, p, n) \quad \dots \quad (2.26)$$

and such a function is called *variance function*. For instance, in case of equal probability sampling of units with replacement mentioned earlier, where the population mean is estimated by the sample mean t , the variance function of the estimator is of the form $V(t) = \sigma^2/n$, where σ^2 is the population variance. Similarly, it is possible in some other cases also to express the variance as an explicit function of σ , n , x and some other relevant factors such as the *intraclass correlation coefficient* (ρ_c) within the sampling units or groups of sampling units. In many situations, it may be necessary to find out the form of the variance function and the constants involved therein empirically through extensive studies.

The cost of the survey can be considered as a function of the sample size n , size of the sampling unit x , the sampling scheme used in selecting the sample (p) and scope of the survey (s , say), that is,

$$C = g(n, x, p, s) \quad \dots \quad (2.27)$$

and such a function is called *cost function*. The cost usually increases with sample size and scope of the survey, but decreases with increase in the size of the sampling unit for the same total number of elementary units included in the sample. A simple form of a cost function is given by

$$C = C_0 + nC_1, \quad \dots \quad (2.28)$$

where C_0 is the overhead cost and C_1 is the average cost of surveying one unit. The forms of the cost functions for specified sample designs and survey organizations have to be evolved on the basis of suitably planned exploratory studies and study of similar surveys conducted earlier.

2 14 NOTATIONS

The capital letters are usually used to denote the population values and the small letters to denote the sample observations. For instance, Y_1, Y_2, \dots, Y_N denote the values of the N units in the population whereas y_1, y_2, \dots, y_n are used to denote the values of the units selected in the sample. It may be noted that while N is used to denote the total number of units in the population n is used to indicate the number of units in the sample that is the sample size. The population parameters are denoted by either capital letters of the English alphabet or by Greek letters. For example \bar{Y} and σ are respectively used to denote the population mean and the standard deviation. The estimators of parameters based on sample observations are indicated by small letters or with caps ($\hat{\cdot}$) on the corresponding symbols of population parameters. y and \hat{p} are examples of the notations used to denote the estimators of the population parameters. In what follows the word *unit* will stand for *sampling unit* unless otherwise stated. The notations used for population and sample characteristics in the subsequent chapters are given in Table 2 5.

TABLE 2 5 NOTATIONS FOR DENOTING POPULATION
AND SAMPLE CHARACTERISTICS

sr no	characteristic	notation used for		
		population	sample	
(1)	(2)	(3)	(4)	
1	number of units	N		n
2	sampling fraction	—		f
3	unit	U_1, U_2, \dots, U_N	u_1, u_2, \dots, u_n	
4	value of study variable y	Y_1, Y_2, \dots, Y_N	y_1, y_2, \dots, y_n	
5	value of auxiliary variable x	X_1, X_2, \dots, X_N	x_1, x_2, \dots, x_n	
6	initial probability of selection	P_1, P_2, \dots, P_N	p_1, p_2, \dots, p_n	
7	probability of inclusion in sample	$\Pi_1, \Pi_2, \dots, \Pi_N$	$\pi_1, \pi_2, \dots, \pi_n$	
8	mean value of y	\bar{Y}		\bar{y}
9	proportion	P		p
10	ratio	R		r

TABLE 2.5. (Contd.) NOTATIONS FOR DENOTING POPULATION AND SAMPLE CHARACTERISTICS.

sr. no.	characteristic	notation used for	
		population	sample
(1)	(2)	(3)	(4)
11.	parameter and estimator	θ	t or $\hat{\theta}$
12.	bias of t	$B(t)$	$b(t)$
13.	variance of t	$V(t)$ or $\sigma^2(t)$	$v(t)$ or $s^2(t)$
14.	covariance between t and t'	$\text{Cov}(t, t')$	$\text{cov}(t, t')$
15.	standard error of t	$\sigma(t)$	$s(t)$
16.	relative standard error of t (coefficient of variation of t)	$C(t)$	$c(t)$
17.	relative variance of t	$C^2(t)$	$c^2(t)$
18.	mean square error of t	$M(t)$	$m(t)$
19.	sampling efficiency of t compared to that of t'	$E_s(t t')$	$e_s(t t')$
20.	cost-efficiency of t compared to that of t'	$E_c(t t')$	$e_c(t t')$
21.	correlation coefficient	ρ	$\hat{\rho}$
22.	intraclass correlation coefficient	ρ_c	$\hat{\rho}_c$
23.	regression coefficient	β	$\hat{\beta}$
24.	summation over the units	$\sum_{i=1}^N$	$\sum_{i=1}^n$
25.	summation over cross-products	$\sum_{i=1}^N \sum_{i' \neq i}^N$	$\sum_{i=1}^n \sum_{i' \neq i}^n$

- (i) The population total and variance of the study variable y are denoted by Γ and σ^2 .
- (ii) The notations in column (4) for characteristics (11) to (23) indicate estimators of corresponding parameters shown in column (3).
- (iii) Subscripts s , i and j are generally used for *strata* (sub-divisions of population), units (or first stage/phase units) and second stage/phase units respectively.
- (iv) C is also used to denote cost.
- (v) Greek letters used are α — alpha; β — beta; Γ , γ — gamma; δ — delta, λ — lambda; θ — theta; μ — mu; ν — nu; Π , π — pye; ρ — rho.

REFERENCES

- COCHRAN, W G (1963) *Sampling Techniques*, Second Edition, Chapter 1, John Wiley & Sons, New York.
- COCHRAN, W G, MOSTELLER F and TUKEY, J W (1954) *Statistical Problems of the Kinsey Report* Chapter 3 and Appendices D and G, American Statistical Association, Washington, D C
- DALENTUS, T (1962) Recent advances in sample survey theory and methods, *Ann Math Stat*, 33, 325-349
- DEMING W E (1960) *Sample Design for Business Research*, Chapters 2 to 4, John Wiley & Sons, New York
- GODAMBE, V P (1965) A review of the contributions towards a unified theory of sampling from finite populations, *Rev Inter Stat Inst*, 33, (2), 242-258
- HANSEN, M H, HURWITZ, W N and JABINE, T B (1963) The use of imperfect lists for probability sampling at the U S Bureau of the Census, *Bull Inter Stat Inst* 40, (1), 497-517
- HANURAV, T V (1966) Some aspects of unified sampling theory, *Sankhya*, 28 (A), 175-203
- KENDALL, M G and BUCKLAND, W R (1957) *A Dictionary of Statistical Terms* Oliver and Boyd London
- KOOP, J C (1963) On the axioms of sample formation and their bearing on construction of linear estimators in sampling theory for finite universes, Parts I, II & III, *Metrika* 7, 81 114, 165 204
- MAHALANOBIS P C (1944) On large scale sample surveys, *Phil Trans Roy Soc*, London 231, (B), 329 451
- MAHALANOBIS P C (1946) Recent experiments in statistical sampling in the Indian Statistical Institute, *J Roy Stat Soc* (A) 109 325-378
- MURTHY, M N (1962) Almost unbiased estimators based on interpenetrating sub samples, *Sankhya*, 24 (A), 303-314
- MURTHY, M N (1963a) Some recent advances in sampling theory, *J Amer Stat Assn*, 58, 737-755
- MURTHY, M N (1963b) Generalized unbiased estimation in sampling from finite populations, *Sankhya*, 25, (B) 245-262
- RAO, C R, MITRA, S K and MATTHAI A (1966) *Formulae and Tables for Statistical Work* Tables 3 1 and 4 1, Statistical Publishing Society, Calcutta
- SEAL, K C (1962) Use of out dated frames in large scale sample surveys, *Bull Cal Stat Assn*, 11, 68-84
- SETH, G R (1961) On some aspects of sampling theory and practice, *Proceedings of Indian Science Congress*, 48th Session, Roorkee
- SUKHATME, P V (1959) Major developments in the theory and applications of sampling during the last twenty five years, *Estatistica*, 17, 652-679
- YATES, F (1953) *Sampling Methods for Censuses and Surveys*, Second Edition, Chapter 4, Charles Griffin & Co, London

COMPLEMENTS AND PROBLEMS

2.1 For conducting surveys on each of the following subjects, define the population and a suitable sampling unit. Indicate the other possible sampling units, if any, in each case and discuss their relative merits :

- (i) income distribution of families in a city ;
- (ii) attitude of students in a State towards a proposed change in the method of examination in schools ;
- (iii) prevalence of a specified contagious disease in a district ;
- (iv) labour force in the urban areas ;
- (v) opinions of radio listeners in a city on a newly introduced radio feature ;
- (vi) pre-harvest acreage under specified food crops in a region ;
- (vii) monthly catch of marine fish along a given stretch of sea coast ;
- (viii) housing conditions in a city ;
- (ix) annual yield of apple fruit in a hilly district ;
- (x) building construction in the urban sector ;
- (xi) the prevalent level of retail prices of some consumer goods in a region ;
- (xii) volume of goods carried through mechanized road transport in a State.

2.2 For surveys on the above subjects, describe the appropriate sampling frames that can be used for sample selection, their sources, possible defects in them and the steps needed to rectify or at least reduce the defects. How do the defects in the sampling frame affect the results of a survey for which it has been used ?

2.3 Explain what you understand by *random sampling* and *non-random sampling*. What are their relative advantages and disadvantages ?

2.4 For an estimator t of a population parameter θ , define the expected value $E(t)$, variance $V(t)$ and mean square error $M(t)$. When is an estimator said to be *biased*? Assuming t to be normally distributed, explain how its bias affects the confidence coefficient of a specified confidence interval for the parameter θ .

2.5 If x is a random variable taking the values 1 and 0 with probabilities p and $q (= 1-p)$, show that $E(x) = p$ and $V(x) = pq$. If y is another random variable taking the values 1 and 0 with probabilities p' and $q' (= 1-p')$, derive the expected values and variances of $(x+y)$, $(x-y)$ and xy .

2.6 Explain clearly what is meant by *cost-efficiency* of an estimator t_1 compared to that of another estimator t_2 . Suppose t_1 and t_2 are unbiased for θ , with $V(t_1)$ being v times $V(t_2)$ and with the cost of obtaining t_1 being c times that of getting t_2 . For what values of c will t_1 be more efficient than t_2 from a joint consideration of the variance and cost aspects ?

2.7 Let t_1 and t_2 be two estimators of θ such that

(i) t_1 is distributed normally with mean θ and standard error 0.01θ , and

(ii) t_2 is symmetrically distributed with $E(t_2) < \theta$ and $P(|t_2 - \theta| < 0.01\theta) > 0.95$

Assuming that the costs of obtaining the estimators t_1 and t_2 are the same, determine which of the two estimators is to be preferred when

(a) loss function $L(\hat{\theta})$ is proportional to the length of the confidence interval for θ , $\hat{\theta}$ being an estimator of θ , and

(b) $L(\hat{\theta}) = 0$ if $\hat{\theta} \geq \theta$ and $L(\hat{\theta}) = c$, if $\hat{\theta} < \theta$

2.8 If t_1, t_2, \dots, t_k are k independent unbiased estimates of a parameter θ with $V(t_i) = V_{ti}, i = 1, 2, \dots, k$, show that any weighted combined estimator

$$t = \sum_{i=1}^k w_i t_i, \quad \left(\sum_{i=1}^k w_i = 1 \right),$$

is unbiased for θ . Obtain an unbiased estimator of $V(t)$ and show that this reduces to the expression (2.25) when $w_i = (1/k)$ for all i . Determine also the values of $\{w_i\}$ that would minimize $V(t)$.

(Sankaran, K. S., *Sanakhyā*, 25, (B), (1963), 345-350).

2.9 Derive the results given in (2.19) and (2.20) relating to expected value and variance of an estimator in the case of three stages of randomization

2.10 If the parameters $\theta_1, \theta_2, \dots, \theta_k$ are unbiasedly estimated by the estimators t_1, t_2, \dots, t_k based on a probability sample, find the expected value and the variance of the statistic $\sum_{i=1}^k a_i t_i$, where a_i 's are constants

Simple Random Sampling

3.1 SAMPLING OF ONE UNIT

The simplest form of random sampling consists in selecting the sample, unit by unit, ensuring equal probability of selection for every unit at each draw and this technique of selection is termed *simple random sampling* (srs). Suppose the sampling frame consists of a list of all the sampling units with adequate identification particulars, but without any information regarding the magnitude or value of the characteristic under consideration for the population units. In such a situation there is no *a priori* reason for selecting one unit more often than any other unit. In other words, in selecting a unit from this population, there is no reason for giving a greater probability or chance of selection to a particular unit than to any other unit. Hence, the units to be included in the sample may be selected one by one with equal probability at each draw and the sample so drawn is termed *simple random sample*.

To fix the ideas regarding this type of sampling, let us consider a finite population of four units

$$U_1, U_2, U_3, U_4$$

with

$$Y_1, Y_2, Y_3, Y_4$$

as the corresponding values of the characteristic y under consideration. Let the objective be the estimation of the population mean

$$\bar{Y} = \frac{1}{4} (Y_1 + Y_2 + Y_3 + Y_4)$$

on the basis of the observation made on one unit selected with srs. The chance of any particular unit being selected is $1/4$ and this is same for all units. This can be achieved by marking the numbers 1, 2, 3 and 4 on four identical counters or slips of paper and choosing any one of them after thorough shuffling.

It can be seen that this procedure gives rise to four possible samples, as any one of the four units may be selected to form the sample, and the probability of getting any particular sample is one in four. Intuitively it is clear that the sample observation itself is to be taken as an estimate of \bar{Y} . The estimator of \bar{Y} is generally denoted by \hat{Y} . In this case \hat{Y} is a random variable taking the values Y_1 , Y_2 , Y_3 and Y_4 with the same probability $1/4$. Further, it is to be noted that the value of the estimator would differ from sample to sample depending on the unit selected in the sample and that these estimates, based as they are on only a part of the population would, in general, be different from \bar{Y} . The difference between \hat{Y} based on a particular sample and \bar{Y} is termed the *error*. The possible samples with their probabilities, estimates of \bar{Y} and errors are given in Table 3.1.

TABLE 3.1 SIMPLE RANDOM SAMPLES OF ONE UNIT FROM A POPULATION OF FOUR UNITS

sr no	sample composition	probability	estimate of \bar{Y}	error $\hat{Y} - \bar{Y}$
(1)	(2)	(3)	(4)	(5)
1	U_1	$1/4$	Y_1	$Y_1 - \bar{Y}$
2	U_2	$1/4$	Y_2	$Y_2 - \bar{Y}$
3	U_3	$1/4$	Y_3	$Y_3 - \bar{Y}$
4	U_4	$1/4$	Y_4	$Y_4 - \bar{Y}$

The expected value of the estimator \hat{Y} is just the mean of the estimates given in column (4) of Table 3.1 and is given by

$$E(\hat{Y}) = \frac{1}{4} \sum_{i=1}^4 Y_i = \bar{Y}.$$

Hence, \hat{Y} is unbiased for \bar{Y} . The mean square error (mse) of \hat{Y} is given by

$$M(\hat{Y}) = \frac{1}{4} \sum_{i=1}^4 (Y_i - \bar{Y})^2,$$

which, in this case, is the population variance σ^2 . Since \hat{Y} is unbiased for \bar{Y} , the sampling variance and the mse are the same, that is, $V(\hat{Y})$ is also equal to σ^2 .

An Example

The sampling variance of the estimator \hat{Y} in sampling one unit with srs from the hypothetical population of 4 units given in Table 3.2 is 1.25, which results in a standard error of 1.118 ($=\sqrt{1.25}$) and an rse of 44.72% ($=(1.118/2.5)100$).

TABLE 3.2. A HYPOTHETICAL POPULATION OF 4 UNITS.

unit	U_1	U_2	U_3	U_4
value of y	1	2	3	4

By a simple generalization of the above discussion, it can be easily seen that in sampling one unit from a population of N units, the estimator \hat{Y} , which is the sample observation itself, is a random variable taking the values $\{Y_i\}$, ($i = 1, 2, \dots, N$), with the same probability $1/N$. The estimator is obviously unbiased, since its expected value is

$$E(\hat{Y}) = \frac{1}{N} \sum_{i=1}^N Y_i = \bar{Y}. \quad \dots \quad (3.1)$$

The mse and the sampling variance of the estimator, which are the same in this case, are given by

$$M(\hat{Y}) = V(\hat{Y}) = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \sigma^2 \quad (3.2)$$

The standard error $\sigma(\hat{Y})$ of the estimator is equal to the population standard deviation σ and the relative standard error (rse) of the estimator is the same as the population coefficient of variation $C(=\sigma/\bar{Y})$. In this case the sampling distribution of the estimator is also the same as the population distribution. It may be mentioned that it is not possible to estimate σ^2 on the basis of a sample of one unit.

3.2 SAMPLING OF TWO UNITS

Suppose we replace the unit selected first and repeat the above operation of sampling to select another unit with srs from the same population of N units. Let y_1 and y_2 be the values of the characteristic y for the unit selected originally and for the unit selected subsequently. Since the second unit has been chosen using the same procedure as for the unit selected originally, y_2 is also unbiased for \bar{Y} and it has the variance σ^2 . Hence the mean of the two sample observations, $\bar{y} = (y_1 + y_2)/2$, will also be an unbiased estimator for \bar{Y} . Since the two units have been selected independently, the covariance between y_1 and y_2 would be zero and hence the sampling variance of \bar{y} is

$$V(\bar{y}) = \frac{1}{4} [V(y_1) + V(y_2)] = \frac{\sigma^2}{2}$$

Thus we see that the sampling variance decreases to half its value, when the sample size is increased from one unit to two units. From this, it may be expected that it is possible in practice to obtain estimates with any prespecified precision by selecting a sufficiently large number of units in the sample, that is, by sufficiently increasing the sample size. The above procedure of selecting a sample of more

than one unit is known as *simple random sampling with replacement* or *equal probability sampling with replacement*, as the units are drawn from the whole population one by one, after replacing at each draw the units selected in the previous draws.

Illustration

To illustrate the concepts involved, let us consider the question of sampling by this method 2 units from the population of 4 units given in Table 3.2. It can be seen that the total number of possible samples is 16, since each of the two units selected in the two draws may be any one of the four units. Also each of the 16 possible samples has the same chance $1/16$ of being selected. However, not all the 16 samples are different in composition, since some of them contain the same units in different order, e.g., samples with serial numbers 2 (12) and 5 (21) in Table 3.3.

TABLE 3.3. ALL SAMPLES OF 2 UNITS FROM 4 UNITS (TABLE 3.2)
IN SIMPLE RANDOM SAMPLING WITH REPLACEMENT.

sr. no. of sample	units in the sample	probabi- lity	sample observations		sample mean \bar{y}	sampling error $\bar{y} - \bar{Y}$	$(y_1 - y_2)^2 / 4$
			y_1	y_2			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1.	1,1	1/16	1	1	1.0	-1.5	0
2.	1,2	1/16	1	2	1.5	-1.0	0.25
3.	1,3	1/16	1	3	2.0	-0.5	1.00
4.	1,4	1/16	1	4	2.5	0	2.25
5.	2,1	1/16	2	1	1.5	-1.0	0.25
6.	2,2	1/16	2	2	2.0	-0.5	0
7.	2,3	1/16	2	3	2.5	0	0.25
8.	2,4	1/16	2	4	3.0	+0.5	1.00
9.	3,1	1/16	3	1	2.0	-0.5	1.00
10.	3,2	1/16	3	2	2.5	0	0.25
11.	3,3	1/16	3	3	3.0	+0.5	0
12.	3,4	1/16	3	4	3.5	+1.0	0.25
13.	4,1	1/16	4	1	2.5	0	2.25
14.	4,2	1/16	4	2	3.0	+0.5	1.00
15.	4,3	1/16	4	3	3.5	+1.0	0.25
16.	4,4	1/16	4	4	4.0	+1.5	0
average					2.5		0.625

We find from columns (6) and (7) of Table 3.3 that these estimates differ, in general, from \bar{Y} , which is 2.5 in this case, and that the error usually varies from sample to sample. The expected value of \bar{y} is given by the average value of column (6), which turns out to be the population mean 2.5, thus verifying that the estimator is unbiased. Further, the mse, which is the same as the sampling variance in this case, is obtained by averaging the squares of errors given in column (7) and this turns out to be 0.625, verifying that $V(\bar{y})$ is $\sigma^2/2$.

An estimator of $V(\bar{y})$ is given by

$$v(y) = (y_1 - y_2)^2 / 4 \quad (3.3)$$

the values of which are given in column (8) of Table 3.3 for different possible samples. The expected value of $v(y)$ which is the average of the values in column (8) is 0.6°5 showing that this estimator is unbiased for $V(y) = \sigma^2/2$

3.3 SAMPLING n UNITS WITH REPLACEMENT

In the general case, simple random sampling with replacement (srswr) consists in choosing at random n numbers from 1 to N replacing at each draw the numbers previously drawn and selecting the corresponding units in the sample, allowing repetitions of units, if any

3.3a EXPECTED VALUE OF SAMPLE MEAN

Suppose the population under consideration has the N units U_1, U_2, \dots, U_N , the i th unit U_i having the value Y_i for the characteristic y . Let a sample of n units be selected with srswr and let y_i ($i = 1, 2, \dots, n$) be the value of the characteristic y for the sample unit selected in the i th draw. Since the sample unit at each draw has been selected from the whole population of N units with srs, y_i is a random variable taking the population values $\{Y_i\}$, ($i=1,2, \dots, N$), with equal probability and hence it is an unbiased estimator of the population mean \bar{Y} and has a sampling variance σ_y^2 , as shown earlier. Thus the sample mean \bar{y} namely

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (3.4)$$

is the mean of n independent unbiased estimates of \bar{Y} and hence it itself is unbiased, that is,

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

It may be noted that in this procedure of selection the same unit may be selected more than once in the sample and that in the case of repetitions of a unit, its value is repeated as many times as it is selected in the sample in building up the estimator considered here.

3.3b VARIANCE OF SAMPLE MEAN

The sampling variance of \bar{y} is, by definition,

$$\begin{aligned} V(\bar{y}) &= E\{\bar{y} - E(\bar{y})\}^2 = E\left\{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})\right\}^2 \\ &= \frac{1}{n^2} \sum_{i=1}^n E(y_i - \bar{Y})^2 + \frac{1}{n^2} \sum_{i=1}^n \sum_{i' \neq i} E(y_i - \bar{Y})(y_{i'} - \bar{Y}) = \frac{\sigma^2}{n}, \quad \dots \quad (3.5) \end{aligned}$$

since $E(y_i - \bar{Y})^2 = V(y_i) = \sigma^2$ and $E(y_i - \bar{Y})(y_{i'} - \bar{Y})$, which is the covariance between y_i and $y_{i'}$, is 0 for $i \neq i'$, as the selections at the i -th and the i' -th draws are independent. The standard error of \bar{y} is given by

$$\sigma(\bar{y}) = \sigma/\sqrt{n} \quad \dots \quad (3.6)$$

and the rse of \bar{y} is

$$C(\bar{y}) = \frac{1}{\sqrt{n}} \frac{\sigma}{\bar{Y}} = \frac{C_y}{\sqrt{n}}, \quad \dots \quad (3.7)$$

where C_y is the population coefficient of variation σ/\bar{Y} . It is to be noted that the standard error and the rse of \bar{y} are both inversely proportional to the square-root of the sample size. The behaviour of the sampling error for varying n , presented in Figure 1.1 of Chapter 1, shows clearly that the decrease in the sampling error is substantial for initial increases in n and that after a certain stage the decrease in the sampling error becomes marginal and incommensurate with the increase in the effort required for having a larger sample.

Alternative Derivation

Alternatively, $E(\bar{y})$ and $V(\bar{y})$ can be obtained as follows by noting that \bar{y} can be written as

$$\bar{y} = \frac{1}{n} \sum_{i=1}^N r_i Y_i, \quad \dots \quad (3.8)$$

where r_i is the number of repetitions of the i -th unit and it is 0 for units not included in the sample.

The values of $E(\bar{y})$ and $V(\bar{y})$ would depend on the values of $E(r_i)$, $V(r_i)$ and $\text{Cov}(r_i, r_{i'})$, for

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^N E(r_i) Y_i \quad (3.9)$$

$$V(\bar{y}) = \frac{1}{n^2} \sum_{i=1}^N V(r_i) Y_i^2 + \frac{1}{n^2} \sum_{i=1}^N \sum_{i' \neq i}^N \text{Cov}(r_i, r_{i'}) Y_i Y_{i'} \quad (3.10)$$

Since r_i is a random variable taking the value $j (= 0, 1, 2, \dots, n)$ with probability

$$\frac{1}{N^n} \binom{n}{j} (N-1)^{n-j} = \binom{n}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{n-j}.$$

$$E(r_i) = \frac{n}{N}, \quad V(r_i) = \frac{n}{N} \left(1 - \frac{1}{N}\right) \text{ and } \text{Cov}(r_i, r_{i'}) = -\frac{n}{N^2} \quad \text{for } i \neq i'.$$

Substituting these in (3.9) and (3.10), we get

$$E(\bar{y}) = \bar{Y},$$

$$V(\bar{y}) = \frac{1}{nN} \left(1 - \frac{1}{N}\right) \sum_{i=1}^N Y_i^2 - \frac{1}{nN^2} \sum_{i=1}^N \sum_{i' \neq i}^N Y_i Y_{i'}$$

$$= \frac{1}{n} \left\{ \frac{1}{N} \sum_{i=1}^N Y_i^2 - \bar{Y}^2 \right\} = \frac{\sigma^2}{n}.$$

B2817

key

106964

3.3c AN UNBIASED ESTIMATOR OF $V(\bar{y})$

One technique of getting an unbiased estimator of a parameter is to consider the expected value of a statistic, which has the same functional form as the parameter. In this case, the statistic corresponding to σ^2 is of the form

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

Considering its expected value, we find that

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

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

$$= \frac{1}{n} \sum_{i=1}^n (\sigma^2 + \bar{Y}^2) - \left(\frac{\sigma^2}{n} + \bar{Y}^2 \right) = \frac{n-1}{n} \sigma^2,$$

since $E(y_i^2) = V(y_i) + \bar{Y}^2$ and $E(\bar{y}^2) = V(\bar{y}) + \bar{Y}^2$. This shows that the statistic considered is not unbiased for σ^2 and that σ^2 is unbiasedly estimated by

$$(\hat{\sigma}^2) = s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \quad \dots \quad (3.11)$$

Therefore, an unbiased estimator of $V(\bar{y})$, namely, σ^2/n , is given by

$$v(\bar{y}) = s^2/n. \quad \dots \quad (3.12)$$

From this, it follows that the estimators of the standard error of \bar{y} , $\sigma(\bar{y})$, and of the rse, $C(\bar{y})$, can be taken as

$$\hat{\sigma}(\bar{y}) = s(\bar{y}) = \frac{s}{\sqrt{n}}, \quad \dots \quad (3.13)$$

$$\hat{C}(\bar{y}) = c(\bar{y}) = \frac{1}{\sqrt{n}} \frac{s}{\bar{y}}. \quad \dots \quad (3.14)$$

It is to be noted that these estimators are not unbiased and that they should be used with caution because of the possibility of the bias being large. However, in the case of large samples, the biases in these estimators are likely to be negligible.

Alternative Derivations

Alternatively, an unbiased estimator of $V(\bar{y})$ can be derived by noting that

$$V(\bar{y}) = \frac{\sigma^2}{n} = \frac{1}{n} \left(\frac{1}{N} \sum_{i=1}^N Y_i^2 - \bar{Y}^2 \right) \quad \dots \quad (3.15)$$

and by getting unbiased estimators of $\frac{1}{N} \sum_{i=1}^N Y_i^2$ and \bar{Y}^2 . Since $\frac{1}{n} \sum_{i=1}^n y_i$ is an unbiased estimator of $\frac{1}{N} \sum_{i=1}^N Y_i$, it can be easily seen that an unbiased estimator of $\frac{1}{N} \sum_{i=1}^N Y_i^2$ is provided by $\frac{1}{n} \sum_{i=1}^n y_i^2$. Further, since $E(\bar{y})^2 = V(\bar{y}) + \bar{Y}^2$, an unbiased estimator of \bar{Y}^2 is provided by $\bar{y}^2 - v(\bar{y})$, where $v(\bar{y})$ stands for an unbiased estimator of $V(\bar{y})$, which is under consideration. Substituting the estimators of $\frac{1}{N} \sum_{i=1}^N Y_i^2$ and \bar{Y}^2 in (3.15), we get

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

This is a simple equation in $v(\bar{y})$ and collecting coefficients of $v(\bar{y})$, we get

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

That is,

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

A second alternative procedure of obtaining an unbiased estimator of $V(\bar{y})$ consists in estimating unbiasedly $\frac{1}{N} \sum_{i=1}^N Y_i^2$ by $\frac{1}{n} \sum_{i=1}^n y_i^2$ and \bar{Y}^2 by $\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i' \neq i} y_i y_{i'}$. Since any pair y_i and $y_{i'}$ are statistically independent, the product $y_i y_{i'}$ estimates \bar{Y}^2 unbiasedly and hence, the mean of all the possible pairs is also unbiased for \bar{Y}^2 . Substituting these estimates in (3.15), we get

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

Since

$$\sum_{i=1}^n \sum_{i' \neq i} y_i y_{i'} = \left(\sum_{i=1}^n y_i \right)^2 - \sum_{i=1}^n y_i^2 = n^2 \bar{y}^2 - \sum_{i=1}^n y_i^2,$$

we have

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

3.3d INTERPENETRATING SUB-SAMPLES

If the sample size is large, the calculation of s^2 , which is an unbiased estimator of σ^2 , becomes time-consuming and expensive. To avoid this difficulty, an alternative estimator of variance of \bar{y} , which is easy to calculate, can be obtained by using the technique of interpenetrating sub-samples, mentioned in Section 2.12 of Chapter 2. Suppose the sample of n units is drawn in the form of k independent interpenetrating sub-samples of m units each ($n = mk$), which in this case is equivalent to dividing the sample of n units into k groups of m units each at random, since each of the n units in the sample is independently selected. It may be noted that each group is a sample of m units selected with srswr and hence, the group mean is an unbiased estimator of \bar{Y} with a sampling variance of σ^2/m . Let \bar{y}_i ,

($i = 1, 2, \dots, k$), be the mean of the i -th group. Then an unbiased estimator of the variance of \bar{y} , the mean of the group means $\{\bar{y}_i\}$, is given by (cf. Expression 2.25 of Chapter 2),

$$v'(\bar{y}) = \frac{1}{k(k-1)} \sum_{i=1}^k (\bar{y}_i - \bar{y})^2. \quad \dots \quad (3.16)$$

In this case, an unbiased estimator of σ^2 is provided by

$$s'^2 = nv'(\bar{y}) = \frac{m}{k-1} \sum_{i=1}^k (\bar{y}_i - \bar{y})^2. \quad \dots \quad (3.17)$$

The variances of $v(\bar{y})$ and $v'(\bar{y})$ are given by

$$V\{v(\bar{y})\} = \left(\beta_2 - \frac{n-3}{n-1} \right) \frac{\sigma^4}{n^3}, \quad \dots \quad (3.18)$$

and

$$V\{v'(\bar{y})\} = \left\{ \frac{\beta_2 + 3(m-1)}{m} - \frac{k-3}{k-1} \right\} \frac{\sigma^4}{n^2 k}, \quad \dots \quad (3.19)$$

where $\beta_2 = \mu_4/\sigma^4$, μ_4 being the fourth population moment, namely, $\mu_4 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^4$. Subtracting (3.18) from (3.19), we get after simplification

$$V\{v'(\bar{y})\} - V\{v(\bar{y})\} = \frac{1}{n^2(n-1)(k-1)} \frac{2(n-k)\sigma^4}{n^2(n-1)(k-1)}, \quad \dots \quad (3.20)$$

which shows that $v(\bar{y})$ is more efficient than $v'(\bar{y})$.

3.4 AN ALTERNATIVE ESTIMATOR FOR \bar{Y}

Since repetition of the observation of a repeated unit in a sample selected with srsrsw does not provide additional information for estimating \bar{Y} , the mean of the values of the distinct units in a sample of n units may be considered as an alternative estimator. That is, if y'_1, y'_2, \dots, y'_d denote the values of the distinct units in a simple random sample of n units selected with replacement ($d \leq n$), then the suggested alternative estimator is

$$\bar{y}' = \frac{1}{d} \sum_{i=1}^d y'_i. \quad \dots \quad (3.21)$$

This estimator is unbiased for \bar{Y} and is more efficient than the sample mean y , namely

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^d r_i y'_i,$$

where r_i is the number of repetitions of the i th distinct unit and $\sum_{i=1}^d r_i = n$

The variance of \bar{y} can be obtained by noting that in this case two stages of randomization are involved (i) d is a random variable taking values 1 to n with certain probabilities, and (ii) selection of the d distinct units from N units with equal probability *without replacement* (cf. Section 3.6) and applying the formula (2.18) of Chapter 2 we get

$$V(\bar{y}) = \left\{ E\left(\frac{1}{d}\right) - \frac{1}{N} \right\} \frac{N}{N-1} \sigma^2, \quad (3.22)$$

where $E\left(\frac{1}{d}\right) = \frac{1^{n-1} + 2^{n-1} + \dots + N^{n-1}}{N^n}$ Neglecting terms of degree greater than $\left(\frac{1}{N}\right)^2$ in (3.22) we get

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{2N} + \frac{n-1}{12N^2} \right) \frac{N}{N-1} \sigma^2 \quad (3.23)$$

An unbiased estimator of $V(\bar{y})$ is given by

$$v(\bar{y}) = \left\{ \left(\frac{1}{d} - \frac{1}{N} \right) + \frac{N-1}{N^n - N} \right\} s_d^2, \quad (3.24)$$

where $s_d^2 = 0$ for $d = 1$ and $s_d^2 = \frac{1}{d-1} \sum_{i=1}^d (y'_i - \bar{y})^2$ for $d \geq 2$

The second term in the curly brackets in (3.24), namely $(N-1)/(N^n - N)$, is likely to be negligibly small compared to the first term and hence the variance estimator may be taken as

$$v(\bar{y}) = \left(\frac{1}{d} - \frac{1}{N} \right) s_d^2 \quad (3.25)$$

It may be noted that if N is considerably larger than n , then the chance of repetition of a unit in the sample will be small and hence the gain in using \bar{y} instead of y will be only marginal. The results mentioned in this section have been discussed in detail by Basu (1958), Des Raj and Khamis (1958) and Pathak (1962).

3.5 SAMPLING TWO UNITS WITHOUT REPLACEMENT

From the expression σ^2/n for $V(\bar{y})$ in the case of srs wr, it is clear that though it decreases with increase in n , it does not become zero, even when n is equal to N . This seemingly paradoxical situation, which is not desirable, arises due to the fact that in sampling with replacement the same unit may get selected in the sample more than once and hence, a sample of N units drawn with replacement does not necessarily include all the N population units. Such a situation can be easily avoided by selecting the sample without replacing at each draw the units already selected in the previous draws. This procedure is known as *simple random sampling without replacement* (srs wor). Obviously, in this type of sampling there is no possibility of the same unit occurring more than once. This procedure is explained by the following example.

Suppose the objective is to estimate the mean \bar{Y} of a population of 4 units U_1, U_2, U_3 and U_4 having Y_1, Y_2, Y_3 and Y_4 as the values of the characteristic y on the basis of 2 units drawn with srs wor. It can be seen that there are 6 possible samples in this case, namely, $U_1U_2, U_1U_3, U_1U_4, U_2U_3, U_2U_4$, and U_3U_4 . Let us consider the value of the sample mean as an estimator of \bar{Y} . All possible samples together with their probabilities of selection and the estimates of \bar{Y} based on them are given in Table 3.4.

TABLE 3.4. ALL SAMPLES OF 2 UNITS FROM 4 UNITS IN SAMPLING WITHOUT REPLACEMENT.

sample	U_1U_2	U_1U_3	U_1U_4	U_2U_3	U_2U_4	U_3U_4
probability	1/6	1/6	1/6	1/6	1/6	1/6
estimate	$\frac{1}{2}(Y_1+Y_2)$	$\frac{1}{2}(Y_1+Y_3)$	$\frac{1}{2}(Y_1+Y_4)$	$\frac{1}{2}(Y_2+Y_3)$	$\frac{1}{2}(Y_2+Y_4)$	$\frac{1}{2}(Y_3+Y_4)$

From this, it can be seen that the sample mean \bar{y} is an unbiased estimator of \bar{Y} , for

$$E(\bar{y}) = \sum_{i=1}^3 \sum_{i'>i}^4 \left(\frac{Y_i + Y_{i'}}{2} \right) \frac{1}{6} = \frac{1}{12} \sum_{i=1}^3 \sum_{i'>i}^4 (Y_i + Y_{i'})$$

and since each unit occurs in three samples, we get

$$E(\bar{y}) = \frac{1}{12} \sum_{i=1}^4 3Y_i = \bar{Y}.$$

The sampling variance of \bar{y} is given by

$$\begin{aligned} V(\bar{y}) &= E(\bar{y} - \bar{Y})^2 = E(\bar{y}^*) - \bar{Y}^2 \\ &= \frac{1}{6} \sum_{i=1}^3 \sum_{i' > i}^4 \frac{1}{4} (Y_i + Y_{i'})^2 - \bar{Y}^2 \\ &= \frac{1}{24} \sum_{i=1}^3 \sum_{i' > i}^4 (Y_i^2 + Y_{i'}^2 + 2Y_i Y_{i'}) - \bar{Y}^2. \end{aligned}$$

Noting that Y_i^2 , ($i = 1, 2, 3, 4$), occurs in three samples and that

$$2 \sum_{i=1}^3 \sum_{i' > i}^4 Y_i Y_{i'} = 16\bar{Y}^2 - \sum_{i=1}^4 Y_i^2,$$

we get after simplification

$$V(\bar{y}) = \frac{1}{12} \sum_{i=1}^4 Y_i^2 - \frac{1}{3} \bar{Y}^2 = \frac{\sigma^2}{3}.$$

Comparing this with the sampling variance $\sigma^2/2$ of the sample mean based on a sample of two units drawn with srswr, we find that srs wr is more efficient

In Example

The concept of srs wr can be illustrated by considering the hypothetical population of four units given in Table 3.2. In Table 3.5 the values of \bar{y} and their errors are given for the 6 possible samples. The expected value of \bar{y} , which is given by the average of column (6) of Table 3.5, turns out to be the population mean 2.5, verifying that \bar{y} is an unbiased estimator of \bar{Y} . The variance of \bar{y} is the average of the squares of the errors given in column (7) and this turns out to be 0.416. This is the value of $\sigma^2/3$, since $\sigma^2 = 1.25$ for this population, verifying the formula for sampling variance. Further, it is of interest to note from column (8) that for each sample an unbiased estimate of $V(\bar{y})$ is given by $v(\bar{y}) = (y_1 - y_2)^2/8$, since the average of column (8) is 0.416, which is the value of $V(\bar{y})$.

TABLE 3.5. ALL SAMPLES OF 2 UNITS FROM 4 UNITS (TABLE 3.2) IN SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT.

sr. no. of sample	units in sample	probabi- lity	sample observations		sample mean \bar{y}	error $\bar{y} - \bar{Y}$	$(y_1 - y_2)^2 / 8$
			y_1	y_2			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1.	1,2	1/6	1	2	1.5	-1.0	0.125
2.	1,3	1/6	1	3	2.0	-0.5	0.500
3.	1,4	1/6	1	4	2.5	0	1.125
4.	2,3	1/6	2	3	2.5	0	0.125
5.	2,4	1/6	2	4	3.0	+0.5	0.500
6.	3,4	1/6	3	4	3.5	+1.0	0.125
average					2.5		0.416

3.6 SAMPLING n UNITS WITHOUT REPLACEMENT

In the general case, where n units are selected with srs wr, the number of possible samples is

$$NC_n = \binom{N}{n} = \frac{N(N-1)\dots(N-n+1)}{n(n-1)\dots2\cdot1},$$

which is the number of combinations of n units taken from N units and the probability of selecting any one of these possible samples is $1/\binom{N}{n}$. This can also be seen by noting that a particular sample of n units would get selected if any of the n units in the sample gets selected in the first draw, the probability for which is n/N , any one of the remaining $(n-1)$ units gets selected from the remaining $(N-1)$ units in the population at the second draw, the probability for which is $(n-1)/(N-1)$ and so on. Hence, the probability of getting a particular sample s of n units is given by

$$P(s) = \frac{n}{N} \cdot \frac{n-1}{N-1} \cdots \frac{1}{N-n+1} = 1/\binom{N}{n}.$$

An Example

For instance, when $N=4$ and $n=2$, as in the example given above, the number of possible samples is $\binom{4}{2} = \frac{4 \cdot 3}{2 \cdot 1} = 6$ and the probability of selecting any of these 6 samples is $1/6$.

3.6a EXPECTATION OF SAMPLE MEAN

Let y_1, y_2, \dots, y_n be the values of n units selected with srs w.r.t. Then an unbiased estimator of \bar{Y} is given by the sample mean \bar{y} . For, $E(\bar{y})$ is by definition

$$E(\bar{y}) = \sum_{s=1}^{\binom{N}{n}} \left(\frac{1}{n} \sum_{i=1}^n y_i \right) \cdot \frac{1}{\binom{N}{n}} = \frac{1}{n} \cdot \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \left(\sum_{i=1}^n y_i \right),$$

where $\sum_{s=1}^{\binom{N}{n}}$ stands for summation over all the $\binom{N}{n}$ possible samples. To evaluate this sum we have to find out in how many samples a given population unit occurs. Since in sampling n distinct units, each unit of the population can occur with $(n-1)$ other units selected out of the remaining $(N-1)$ units in the population, each unit occurs in $\binom{N-1}{n-1}$ of the $\binom{N}{n}$ possible samples. Hence,

$$\sum_{s=1}^{\binom{N}{n}} \left(\sum_{i=1}^n y_i \right) = \binom{N-1}{n-1} \sum_{i=1}^N Y_i$$

and noting that $\binom{N-1}{n-1}/\binom{N}{n} = \frac{n}{N}$, we get $E(\bar{y}) = \bar{Y}$.

3.6b VARIANCE OF SAMPLE MEAN

The variance of \bar{y} , by definition, is

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \sum_{s=1}^{\binom{N}{n}} (\bar{y} - \bar{Y})^2 \cdot \frac{1}{\binom{N}{n}}.$$

Substituting $\frac{1}{n} \sum_{i=1}^n y_i$ for \bar{y} , we get

$$\begin{aligned} V(\bar{y}) &= \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y}) \right\}_s^2 \\ &= \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \frac{1}{n^2} \left\{ \sum_{i=1}^n (y_i - \bar{Y})^2 + \sum_{i=1}^n \sum_{i' \neq i}^n (y_i - \bar{Y})(y_{i'} - \bar{Y}) \right\}_s. \end{aligned}$$

We have noted that each unit in the population occurs in $\binom{N-1}{n-1}$ of the $\binom{N}{n}$ possible samples and hence

$$\sum_{s=1}^{\binom{N}{n}} \left\{ \sum_{i=1}^n (y_i - \bar{Y})^2 \right\}_s = \binom{N-1}{n-1} \sum_{i=1}^N (Y_i - \bar{Y})^2.$$

Further, each pair of units $\{(U_i, U_{i'}), \text{ say}\}$ occurs in $\binom{N-2}{n-2}$ of the $\binom{N}{n}$ samples, since the two given units can occur with $(n-2)$ other units out of the remaining $(N-2)$ population units. Hence

$$\sum_{s=1}^{\binom{N}{n}} \left\{ \sum_{i=1}^n \sum_{i' \neq i}^n (y_i - \bar{Y})(y_{i'} - \bar{Y}) \right\}_s = \binom{N-2}{n-2} \sum_{i=1}^N \sum_{i' \neq i}^N (Y_i - \bar{Y})(Y_{i'} - \bar{Y}).$$

Noting that

$$\binom{N-1}{n-1} = \frac{n}{N} \binom{N}{n} \text{ and } \binom{N-2}{n-2} = \frac{n(n-1)}{N(N-1)} \binom{N}{n}$$

and that

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 + \sum_{i=1}^N \sum_{i' \neq i}^N (Y_i - \bar{Y})(Y_{i'} - \bar{Y}) = \left\{ \sum_{i=1}^N (Y_i - \bar{Y}) \right\}^2 = 0,$$

we get

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n} \left\{ \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 \right\} - \frac{1}{n} \frac{n-1}{N-1} \left\{ \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 \right\} \\ &= \left(1 - \frac{n-1}{N-1} \right) \frac{\sigma^2}{n} = \frac{N-n}{N-1} \frac{\sigma^2}{n}. \quad \dots \quad (3.26) \end{aligned}$$

Thus we see that $V(\bar{y})$ in srs wor is less than that in the case of srswr with the reduction factor $\{1-(n-1)/(N-1)\}$, showing that sampling without replacement is more efficient than sampling with replacement. This factor is termed the *finite population correction (fpc)*, as this factor takes account of the finite nature of the population in sampling without replacement and reduces to zero when n is increased to equal N and becomes almost 1 for large populations. This clearly shows that the gain in using srs wor over srswr would be substantial whenever the population itself is small or the sampling fraction, $f = (n/N)$, is not very small. Substituting $N = 4$ and $n = 2$ in the expression for $V(\bar{y})$ given in (3.26), we get $\sigma^2/3$, verifying the result given in Section 3.5.

The expression for $V(\bar{y})$ given in (3.26) can be written in a slightly different form by expressing σ^2 in terms of σ'^2 , where

$$\sigma'^2 = \frac{N\sigma^2}{N-1} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad \dots \quad (3.27)$$

Substituting $(N-1)\sigma'^2/N$ for σ^2 in (3.26), we get

$$V(\bar{y}) = \frac{N-n}{Nn} \sigma'^2 = (1-f) \frac{\sigma'^2}{n} \quad \dots \quad (3.28)$$

The fact that srs wor is more efficient than srswr can also be shown by expressing $V(\bar{y})$ in the case of sampling with replacement in terms of σ'^2 , namely,

$$V(\bar{y}) = \left(1 - \frac{1}{N}\right) \frac{\sigma'^2}{n} \quad \dots \quad (3.29)$$

and comparing (3.29) with (3.28).

Alternative Derivation

The expected value and variance of y in srs wor can alternatively be derived by writing \bar{y} as

$$y = \frac{1}{n} \sum_{i=1}^N r_i Y_i,$$

where r_i is 1 or 0 according as the i th unit in the population is included in the sample or not ($\sum_{i=1}^N r_i = n$). Noting that

$$E(r_i) = \frac{n}{N}, \quad V(r_i) = \frac{n}{N} \left(1 - \frac{n}{N}\right) \text{ and } \text{Cov}(r_i, r_{i'}) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2,$$

we get

$$E(\bar{y}) = \frac{1}{n} \sum_{t=1}^N E(r_t) Y_t = \frac{1}{N} \sum_{t=1}^N Y_t = \bar{Y},$$

and

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n^2} \sum_{t=1}^N V(r_t) Y_t^2 + \frac{1}{n^2} \sum_{t=1}^N \sum_{t' \neq t}^N \text{Cov}(r_t, r_{t'}) Y_t Y_{t'}, \\ &= \frac{1}{nN} \left(1 - \frac{n}{N} \right) \sum_{t=1}^N Y_t^2 + \frac{1}{nN} \left(\frac{n-1}{N-1} - \frac{n}{N} \right) \sum_{t=1}^N \sum_{t' \neq t}^N Y_t Y_{t'}, \end{aligned}$$

which after simplification becomes $(1-f)\sigma'^2/n$.

3.6c BEHAVIOUR OF SAMPLING ERROR

The standard error $\sigma(\bar{y})$ and the rse $C(\bar{y})$ in srs wor are given by

$$\sigma(\bar{y}) = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}} \quad \dots \quad (3.30)$$

and

$$C(\bar{y}) = \frac{\sigma(\bar{y})}{\bar{Y}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{C}{\sqrt{n}}, \quad \dots \quad (3.31)$$

where C is the population coefficient of variation. The behaviour of the rse of \bar{y} in sampling with srs wor as compared to that in the case of srswr is shown in Figure 3.1 for different values of N and n ,

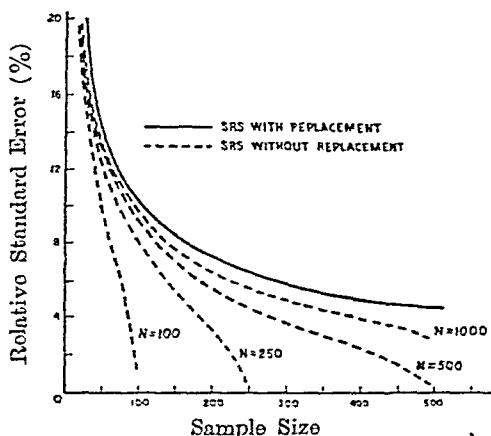


Figure 3.1. Behaviour of relative standard error in sampling with and without replacement.

taking C to be 100%. From this figure it is clear that as mentioned earlier the gain in using sampling without replacement in preference to sampling with replacement is substantial when the sample size is large compared to the number of units in the population and that it becomes marginal for relatively small sample sizes.

3.6d UNBIASED ESTIMATOR OF $V(\bar{y})$

An unbiased estimator of

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

can be obtained by getting unbiased estimators of $\frac{1}{N} \sum_{i=1}^N Y_i^2$ and \bar{Y}^2 .

The term $\frac{1}{N} \sum_{i=1}^N Y_i^2$ is estimated unbiasedly by $\frac{1}{n} \sum_{i=1}^n y_i^2$ just as $\frac{1}{N} \sum_{i=1}^N Y_i$ is estimated unbiasedly by $\frac{1}{n} \sum_{i=1}^n y_i$. Since $V(\bar{y}) = E(\bar{y}^2) - \bar{Y}^2$, that is $\bar{Y}^2 = E(\bar{y}^2) - V(\bar{y})$ an unbiased estimator of \bar{Y}^2 is given by $\bar{y}^2 - v(\bar{y})$ where $v(y)$ denotes an unbiased estimator of $V(\bar{y})$. Hence,

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

Solving for $v(\bar{y})$ we get

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

That is,

$$v(\bar{y}) = (1-f) \frac{s^2}{n}, \quad (3.32)$$

where $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$. From this, it is clear that an unbiased estimator of σ'^2 is given by s^2 . Thus we see that whereas s^2 estimates σ^2 unbiasedly in srs wr, it estimates σ^2 in srs wor. An unbiased estimator of σ^2 in the case of srs wor is provided by $(N-1)s^2/N$.

The estimators of the standard error $\sigma(\bar{y})$ and the rse $C(\bar{y})$ can be taken as

$$\hat{\sigma}(\bar{y}) = s(\bar{y}) = \sqrt{1-f} \frac{s}{\sqrt{n}} \quad \dots \quad (3.33)$$

and

$$\hat{C}(\bar{y}) = c(\bar{y}) = \sqrt{\frac{1-f}{n}} \frac{s}{\bar{y}}. \quad \dots \quad (3.34)$$

As in srswr, these estimators are not unbiased and should be used with caution. However, the bias is likely to be small in case of large samples.

Alternative Derivation

Alternatively, an unbiased estimator of $V(\bar{y})$ can be derived by expressing the sum of squares in σ^2 and σ'^2 , namely $\sum_{i=1}^N (Y_i - \bar{Y})^2$, as $\frac{1}{N} \sum_{i=1}^N \sum_{i' > i}^N (Y_i - Y_{i'})^2$

and estimating $\sum_{i=1}^N \sum_{i' > i}^N (Y_i - Y_{i'})^2$ unbiasedly by $\frac{N(N-1)}{n(n-1)} \sum_{i=1}^n \sum_{i' > i}^n (y_i - y_{i'})^2$

obtained from the sample. Substituting this in the expression for $V(\bar{y})$, we get

$$v(\bar{y}) = \frac{N-n}{Nn} \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i' > i}^n (y_i - y_{i'})^2.$$

Noting that

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

we get $v(\bar{y}) = (1-f)s^2/n$.

If n is large, the calculation of s^2 becomes time-consuming and expensive. This difficulty can, however, be overcome by dividing the sample of n units into k groups of m units each at random ($n = mk$) and obtaining an estimator of σ'^2 on the basis of the group means $\{\bar{y}_i\}$, ($i = 1, 2, \dots, k$). In that case, an unbiased estimator of σ'^2 is given by

$$s'^2 = \frac{m}{(k-1)} \sum_{i=1}^k (\bar{y}_i - \bar{y})^2.$$

For,

$$\begin{aligned} E(s'^2) &= \frac{m}{k-1} \left\{ \sum_{i=1}^k E(\bar{y}_i^2) - kE(\bar{y}^2) \right\} \\ &= \frac{m}{k-1} \left[k \left\{ \left(1 - \frac{m}{N} \right) \frac{\sigma'^2}{m} + \bar{Y}^2 \right\} - k \left\{ \left(1 - \frac{n}{N} \right) \frac{\sigma'^2}{n} + \bar{Y}^2 \right\} \right] \\ &= \frac{m}{k-1} \left\{ \left(\frac{1}{m} - \frac{1}{N} \right) - \left(\frac{1}{n} - \frac{1}{N} \right) \right\} \sigma'^2 = \sigma'^2 \end{aligned}$$

Hence, another unbiased estimator of $V(\bar{y})$, which is easier to compute, is given by

$$v'(\bar{y}) = (1-f)s'^2/n \quad . \quad (3.35)$$

As in the case of srswr, it can be shown that this variance estimator is less efficient than that given in (3.32).

3.7 ESTIMATION OF TOTALS

So far we have considered the question of estimation of \bar{Y} on the basis of a sample drawn with srs. In this section, the estimation of the population total and of the total of units belonging to a particular class is discussed.

3.7a ESTIMATION OF POPULATION TOTAL

Since the population total Y is simply $N\bar{Y}$, an unbiased estimator of Y can be obtained by multiplying an unbiased estimator of \bar{Y} by N . Further, the expressions for $V(\hat{Y})$ and $v(\hat{Y})$ can be got by multiplying the corresponding expressions for $\hat{\bar{Y}}$ by N^2 .

Hence an estimator of Y , its variance and the variance estimator in srswr and srs wr are respectively given by

$$\hat{Y} = N\bar{y}, \quad V(\hat{Y}) = N^2 \frac{\sigma^2}{n}, \quad v(\hat{Y}) = N^2 \frac{s^2}{n} \quad (3.36)$$

and

$$\hat{Y} = N\bar{y}, \quad V(\hat{Y}) = N^2 \frac{N-n}{N-1} \frac{\sigma^2}{n}, \quad v(\hat{Y}) = N^2(1-f) \frac{s^2}{n} \quad (3.37)$$

It may be noted that the rse of \hat{Y} is the same as that of \bar{y} .

3.7b TOTAL AND MEAN OF A SUB-POPULATION

If the interest lies in estimating the total of a part of the population, which may be one of the domains of study, the above procedure of estimating the population total may be used by taking 0 as the value for the units not belonging to the domain of study. For instance, in a consumer expenditure survey, if we are interested in estimating the total expenditure of households belonging to a specified income group or occupation class, the values for units not belonging to that group or class are to be taken as 0 in obtaining the estimate of the total for this part of the population. The total value Y' of the sub-population is given by

$$Y' = \sum_{i=1}^N Y'_i; \quad \dots \quad (3.38)$$

where $Y'_i = Y_i$ for units belonging to the sub-population and $Y'_i = 0$ for units not belonging to this sub-population. Thus we see that the sub-population total Y' has been written as a population total with a modification in the definition of the value of the unit. An unbiased estimator of Y' in srs is given by

$$\hat{Y}' = \frac{N}{n} \sum_{i=1}^n \hat{y}'_i; \quad \dots \quad (3.39)$$

where $\hat{y}'_i = y_i$ for sample units belonging to the specified group or class and $\hat{y}'_i = 0$ for the other sample units. The variance and the variance estimator of \hat{Y}' can be obtained by substituting σ'^2 and s'^2 for σ^2 and s^2 in (3.36) and (3.37), where

$$\sigma'^2 = \frac{1}{N} \sum_{i=1}^N (Y'_i - \bar{Y}')^2, \quad \bar{Y}' = \frac{1}{N} \sum_{i=1}^N Y'_i \quad \dots \quad (3.40)$$

and

$$s'^2 = \frac{1}{n-1} \sum_{i=1}^n (\hat{y}'_i - \bar{y}')^2, \quad \bar{y}' = \frac{1}{n} \sum_{i=1}^n \hat{y}'_i. \quad \dots \quad (3.41)$$

If the number of units (N' say) in the sub population under consideration is known, then an unbiased estimator of the mean of the sub population namely \hat{Y}' ($= \bar{Y}'/N'$), can be obtained by dividing \bar{Y}' by N' , that is,

$$\hat{Y}' = \frac{1}{N'} \frac{N}{n} \sum_{i=1}^n y'_i \quad (3.42)$$

The expression for variance and variance estimator of \hat{Y}' can be got by dividing the corresponding expressions for \bar{Y}' by N'^2 . That is,

$$V(\hat{Y}') = \frac{N^2}{N'^2} \frac{\sigma''^2}{n}, \quad i(\hat{Y}') = \frac{N^2}{N'^2} \frac{s''^2}{n} \quad (3.43)$$

in srsr, and in srs wor

$$V(\hat{Y}') = \frac{N^2(N-n)}{N'^2(N-1)} \frac{\sigma''^2}{n}, \quad i(\hat{Y}') = \frac{N^2(N-n)}{N'^2N} \frac{s''^2}{n} \quad (3.44)$$

If the value of N' is not known, then an estimator of \bar{Y}' is obtained as the ratio \hat{Y}'/\hat{N}' , where \hat{N}' is an estimator of N' . In this case, the estimator becomes a ratio estimator, which is generally biased, and such estimators are discussed in Chapter 10. The question of estimation of the number of units and proportion of units in a sub population is considered in the next section.

3.8 ESTIMATION OF PROPORTION OF UNITS

In this section, the question of estimating the proportion P of units of a population belonging to a specified class is considered. Examples of such situations are classification of villages factories, etc as big or small on the basis of some characteristics such as population, number of workers, etc and classification of persons having specified characteristics such as a particular educational qualification, disease, etc. Thus we may be interested in estimating the proportion of villages having medical facilities, or the proportion of graduates or unemployed persons in a region, or the proportion of persons having cancer or tuberculosis, or the proportion of factories having a fixed capital exceeding a specified amount, etc.

The estimation of P reduces to that of estimating a population mean, which has been discussed in detail in earlier sections, if we assign to each unit the value 1 or 0 according as that unit belongs or does not belong to that class or category. That is, Y_i , the value of the i -th unit in the population is 1 if that unit belongs to the specified class and 0 otherwise. With this definition of the value of Y_i , it can be easily seen that

$$Y = \sum_{i=1}^N Y_i = N', \quad \dots \quad (3.45)$$

where N' is the number of units belonging to that class and

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i = \frac{N'}{N} = P. \quad \dots \quad (3.46)$$

Further, we note that since Y_i takes only the values 1 or 0, Y_i^2 will also take only the values 1 or 0 and hence $\sum_{i=1}^N Y_i^2 = N'$. Thus we see that

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \bar{Y}^2 = P - P^2 = PQ, \quad Q = 1 - P. \quad \dots \quad (3.47)$$

3.8a SRS WITH REPLACEMENT

Suppose n units are drawn with srswr. In this case \bar{y} reduces to the sample proportion p , for

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{r}{n} = p, \quad \dots \quad (3.48)$$

where r is the number of units in the sample belonging to the specified category and hence reporting 1 for the value of the characteristic y . Since $V(\bar{y})$ has already been shown to be σ^2/n , we get

$$V(p) = PQ/n. \quad \dots \quad (3.49)$$

The population coefficient of variation in this case is

$$C = \sqrt{PQ}/P = \sqrt{Q/P} \quad \dots \quad (3.50)$$

and the rse of p is

$$C(p) = \frac{1}{\sqrt{n}} \sqrt{\frac{Q}{P}} \quad \dots \quad (3.51)$$

Further, since s^2 is an unbiased estimator of σ^2 , it follows that

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right) \\ &= \frac{1}{n-1} (np - np^2) = \frac{npq}{n-1}, \quad (q = 1-p), \quad \dots \quad (3.52) \end{aligned}$$

is unbiased for PQ . Hence, an unbiased estimator of $V(p)$ is given by

$$v(p) = pq/(n-1) \quad \dots \quad (3.53)$$

3.8b SRS WITHOUT REPLACEMENT

Proceeding similarly, we find that the sample proportion p is unbiased for P in the case of srs wor also $V(p)$ and $C(p)$ in this case are given by

$$V(p) = \frac{N-n}{N-1} \cdot \frac{PQ}{n} \quad \dots \quad (3.54)$$

and

$$C(p) = \sqrt{\frac{N-n}{N-1}} \cdot \frac{1}{\sqrt{n}} \sqrt{\frac{Q}{P}}. \quad \dots \quad (3.55)$$

Since we have found that in srs wor s^2 is an unbiased estimator of $\sigma'^2 = N\sigma^2/(N-1)$, an unbiased estimator of σ^2 is given by

$$(\hat{\sigma}^2) = \frac{N-1}{N} s^2 = \frac{N-1}{N} \cdot \frac{npq}{n-1}. \quad \dots \quad (3.56)$$

Hence, an unbiased estimator of $V(p)$ is given by

$$v(p) = (1-f) \frac{pq}{n-1}. \quad \dots \quad (3.57)$$

The behaviour of $V(p)$ with increase in n is similar to that of $V(\bar{y})$ considered earlier.

3.8c NUMBER OF UNITS IN A CLASS

Since the number of units N' in a class is NP , an unbiased estimator for N' is given by N times the sample proportion p . That is,

$$\hat{N}' = Np. \quad \dots \quad (3.58)$$

The expressions for $V(\hat{N}')$ and $v(\hat{N}')$ can be obtained by multiplying the corresponding expressions for p by N^2 . That is, in srs wr

$$V(\hat{N}') = N^2 \frac{PQ}{n}, \quad v(\hat{N}') = N^2 \frac{pq}{n-1}. \quad \dots \quad (3.59)$$

and in srs wor

$$V(\hat{N}') = N^2 \frac{N-n}{N-1} \frac{PQ}{n}, \quad v(\hat{N}') = N^2(1-f) \frac{pq}{n-1}. \quad \dots \quad (3.60)$$

3.9 CONFIDENCE INTERVAL

Since $V(\bar{y})$ in the case of simple random sampling decreases with increase in n , it follows that for a given population the \bar{y} 's would be more and more concentrated in the close neighbourhood of \bar{Y} with increasing n . The type and nature of the *sampling distribution* of \bar{y} would depend on the distribution of the population (termed *parent distribution*) and n . Here we shall study the sampling distribution of \bar{y} when the population follows either of the two types of distributions considered in Section 2.2 of Chapter 2, namely, normal and skew distributions, with a view to bringing out the significance of the sampling variance and the effect of increasing the sample size.

3.9a NORMAL DISTRIBUTION : σ KNOWN

In the case where the variable y is distributed as $N(\bar{Y}, \sigma^2)$, the sampling distribution of \bar{y} is also normal with mean \bar{Y} and variance $V(\bar{y})$. Since for a normal distribution about 95% of the values lie between the limits given by $\bar{Y} \pm 1.96\sigma$, we find that for about 95% of the samples the values of \bar{y} would be between the limits $\bar{Y} - 1.96\sigma(\bar{y})$

and $\bar{Y} + 1.96\sigma(\bar{y})$ This situation is denoted by the following probability statement

$$\text{Prob } \{\bar{Y} - 1.96\sigma(\bar{y}) \leq \bar{y} \leq \bar{Y} + 1.96\sigma(\bar{y})\} = 95\%,$$

which states that the probability of \bar{y} lying between the limits $\bar{Y} - 1.96\sigma(\bar{y})$ and $\bar{Y} + 1.96\sigma(\bar{y})$ is 95%. The above probability statement may also be written as

$$\text{Prob } \{\bar{y} - 1.96\sigma(\bar{y}) \leq \bar{Y} \leq \bar{y} + 1.96\sigma(\bar{y})\} = 95\%,$$

which shows that the probability of the interval $\{\bar{y} - 1.96\sigma(\bar{y}), \bar{y} + 1.96\sigma(\bar{y})\}$ containing \bar{Y} is 95%. This interval is known as *confidence interval* and the probability attached to it is termed the *confidence coefficient* (cf Section 2.11 of Chapter 2). The length of the confidence interval is $3.92\sigma(\bar{y})$ and since $\sigma(\bar{y})$ decreases as n increases, it is clear that by increasing n sufficiently, it would be possible to reduce the length of this confidence interval to any desired small range, thereby ensuring that \bar{Y} is in the immediate neighbourhood of \bar{y} with a confidence coefficient of 95%.

The confidence coefficient can be altered to any desired level $(1-\alpha)$ by using an appropriate value of k_α instead of 1.96 in working out the confidence interval. In this case we have

$$\text{Prob } \{\bar{y} - k_\alpha\sigma(\bar{y}) \leq \bar{Y} \leq \bar{y} + k_\alpha\sigma(\bar{y})\} = 1 - \alpha \quad \dots \quad (3.61)$$

This means that given the value of σ , we can find a range or interval based on any given sample, which is expected to contain the true value in $(1-\alpha)\%$ of the cases. The values of k_α for different values of the confidence coefficient usually used in practice are given in Table 2.1 of Chapter 2.

3.9b NORMAL DISTRIBUTION σ UNKNOWN

Since in practice the value of the population variance σ^2 is not known, usually the confidence interval has to be set up on the basis of the variance estimate $s^2(\bar{y})$ obtained from the sample itself. To build up a confidence interval using this variance estimate, we note that as mentioned in Section 2.11 of Chapter 2 the statistic $t = (\bar{y} - \bar{Y})/s(\bar{y})$

is distributed as Student's t with $(n-1)$ degrees of freedom, when $s(\bar{y})$ is computed using (3.13) or (3.33) as the case may be. Hence, the confidence interval for any specified confidence coefficient P ($=1-\alpha$) is given by $\{\bar{y}-t_\alpha s(\bar{y}), \bar{y}+t_\alpha s(\bar{y})\}$, where t_α is the value of t in the Student's distribution with $(n-1)$ degrees of freedom beyond which a proportion $(\alpha/2)$ of the values lie. That is,

$$\text{Prob.}\{\bar{y}-t_\alpha s(\bar{y}) \leq \bar{Y} \leq \bar{y}+t_\alpha s(\bar{y})\} = P. \quad \dots \quad (3.62)$$

The values of t_α for different values of P and v ($=n-1$) have been tabulated (cf. Table 2.4 of Chapter 2).

If σ^2 is estimated using the short-cut method based on the technique of independent interpenetrating sub-samples, that is, if s'^2 given in (3.17) is used as an estimator of σ^2 , then the statistic $t' = (\bar{y}-\bar{Y})\sqrt{n}/s'$ is distributed as Student's t with $(k-1)$ degrees of freedom, where k is the number of sub-samples on which s'^2 is based. However, if the number of degrees of freedom ($n-1$ or $k-1$ as the case may be) is fairly large (more than 30), the statistic t' is approximately normally distributed and the percentage points of the normal distribution can be used in setting up the confidence interval. It is of interest to note that if the number of sub-samples k is moderately large (about 10, say), the average length and the sampling distribution of the confidence interval based on the estimator s'^2 for any specified confidence coefficient are not much different from those of the confidence interval based on the estimator s^2 (Murthy, 1962).

3.9c NON-NORMAL DISTRIBUTION

When the population distribution is not normal, the sampling distribution of the estimator, which is the sample mean in this case, is usually approximately normal provided the sample size and the number of possible samples are large. Hence, the procedure of setting up confidence interval for \bar{Y} in the case of normal distribution discussed earlier can be applied to non-normal populations also, provided n is sufficiently large. It is difficult in general to determine how large

a sample should be for using the normal approximation in setting up confidence interval, as the rapidity with which the sampling distribution of an estimator approaches normality with increase in n depends to a large extent on the nature of the parent or population distribution and the sampling method used

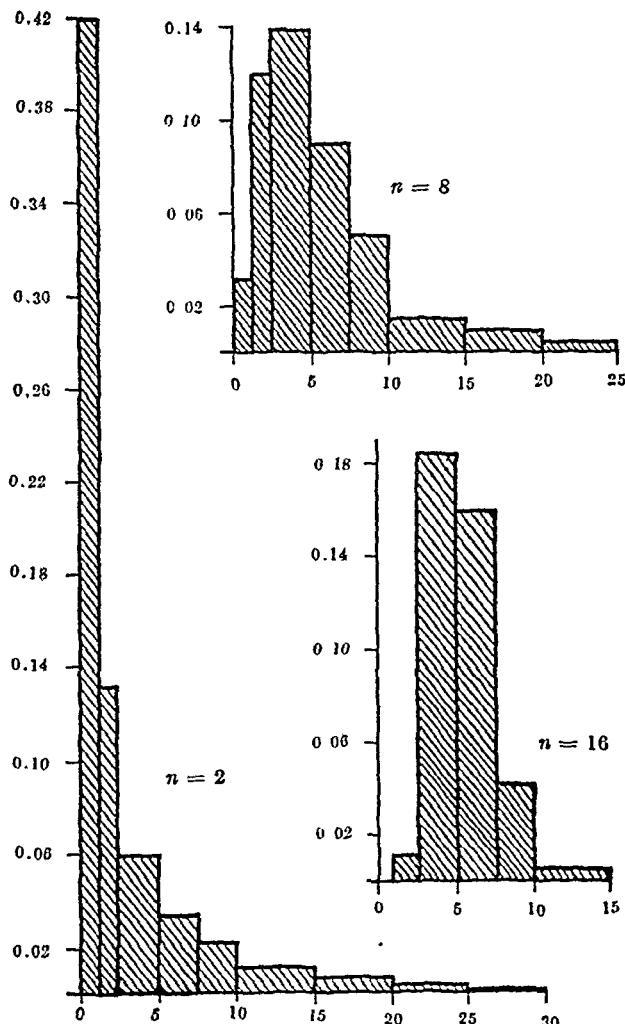
An illustration of this situation is given by the sampling distributions of the simple means based on simple random samples of different sizes drawn from the highly skew population given in Table 2.2 relating to the distribution of operational land holdings by holding size. For the sake of simplicity it is assumed that all the units belonging to a particular holding size class have the same value equal to the average holding size in that class. Figure 3.2 shows the histograms of the frequency distributions of 500 samples for different sample sizes, where the samples are drawn using the technique of srswr. From this set of figures it is clear that the skewness of the sampling distribution of the sample mean decreases as the sample size increases and that the distribution tends to cluster increasingly around the population mean approximating a normal distribution.

3.9d CONFIDENCE INTERVAL FOR PROPORTION

In estimating a proportion P on the basis of a simple random sample the sample proportion p can be considered to be normally distributed with mean P and standard deviation $\sigma(p)$ provided the sample size is sufficiently large. The decision as to how large n should be to make the application of normal distribution theory valid would depend on the value of P and the sampling method used. In large samples, the confidence interval for P is provided by

$$\text{Prob} \{p - k_\alpha s(p) \leq P \leq p + k_\alpha s(p)\} = 1 - \alpha, \quad (3.63)$$

where k_α is the $\alpha\%$ point in the normal distribution beyond which $(\alpha/2)\%$ of the values lie and $s(p) = \sqrt{pq/(n-1)}$ in the case of srswr and $s(p) = \sqrt{(1-f)pq/(n-1)}$ in the case of srs wor.



X-axis : land holding in acres ; Y-axis : proportion of samples

Figure 3.2. Histograms of distributions of sample mean based on 500 samples of sizes 2, 8 and 16.

3.10 POOLING OF ESTIMATES

In practice, there may be situations where the estimates based on several samples may have to be pooled to get a combined estimator, and we consider this problem in this section.

3.10a SRS WITH REPLACEMENT

Suppose $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k$ are sample means based on k samples of sizes n_1, n_2, \dots, n_k , each of which is selected with srswr from the same population. Then a combined estimator based on all the sample units is given by the weighted mean of the sample means

$$\bar{y} = \frac{1}{n} \sum_{i=1}^k n_i \bar{y}_i \quad \left(n = \sum_{i=1}^k n_i \right) \quad (3.64)$$

The variance of the combined estimator is

$$V(\bar{y}) = \frac{1}{n^2} \sum_{i=1}^k n_i^2 V(\bar{y}_i)$$

the $Cov(\bar{y}_i, \bar{y}_j)$ ($i \neq j$) terms being zero since the k samples have been selected independently. Noting that $V(\bar{y}_i) = \sigma^2/n_i$, we find $V(\bar{y}) = \sigma^2/n$, which is as it should be, since \bar{y} is nothing but the sample mean based on a combined sample of n units selected with srswr. It is to be noted that \bar{y} being the best linear estimator, is more efficient than the arithmetic mean of the k estimates.

$$\bar{y} = \frac{1}{k} \sum_{i=1}^k \bar{y}_i \quad (3.65)$$

Unbiased estimators of $V(y)$ and $V(\bar{y})$ are provided by

$$t(\bar{y}) = \frac{1}{n^2} \sum_{i=1}^k n_i s_i^2, \quad (3.66)$$

and

$$v(\bar{y}) = \frac{1}{k(k-1)} \sum_{i=1}^k (\bar{y}_i - \bar{y})^2, \quad . \quad (3.67)$$

where $s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$. However, more efficient estimators of $V(\bar{y})$ and $V(\bar{y}')$ can be obtained by estimating σ^2 by s^2 based on the combined sample of n units. That is,

$$v'(\bar{y}) = \frac{s^2}{n}, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2, \quad \dots \quad (3.68)$$

and

$$v'(\bar{y}') = \frac{s^2}{k^2} \sum_{i=1}^k \frac{1}{n_i}. \quad \dots \quad (3.69)$$

3.10b SRS WITHOUT REPLACEMENT

If the sample means $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k$ are based on k independent samples of sizes n_1, n_2, \dots, n_k , each of which is selected from the same population with srs wr, a combined estimator is provided by the weighted mean \bar{y} given in (3.64). Noting that $V(\bar{y}_i) = (1-f_i) \sigma'^2/n_i$, f_i being (n_i/N) , we get

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n^2} \sum_{i=1}^k n_i^2 V(\bar{y}_i) = \frac{\sigma'^2}{n^2} \sum_{i=1}^k \frac{(N-n_i)n_i}{(N-1)} \\ &= \frac{\sigma'^2}{n} \left\{ 1 - \frac{1}{Nn} \sum_{i=1}^k n_i^2 \right\}. \quad \dots \quad (3.70) \end{aligned}$$

An unbiased estimator of $V(\bar{y})$ is given by

$$v(\bar{y}) = \frac{1}{n^2} \sum_{i=1}^k n_i(1-f_i) s_i^2. \quad \dots \quad (3.71)$$

Since $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k$ are k independent estimates of \bar{Y} , an unbiased estimator of the variance of \bar{y}' ($= \sum_{i=1}^k \bar{y}_i/k$) is given by (3.67). It may be noted that as in the case of sampling with replacement the estimator \bar{y}' is less efficient than \bar{y} . However, \bar{y} and \bar{y}' become identical if the sample sizes n_1, n_2, \dots, n_k are the same.

3.10c ESTIMATION OF A PROPORTION

The results given above can readily be applied to the case of estimating proportions. In srs wr, the combined estimator based on k sample proportions p_1, p_2, \dots, p_k is given by

$$p = \frac{1}{n} \sum_{i=1}^k n_i p_i \quad (3.72)$$

and its variance is

$$V(p) = PQ/n, \quad (3.73)$$

for $V(p_i) = PQ/n_i$. An unbiased estimator of $V(p)$ is provided by

$$v(p) = \frac{1}{n^2} \sum_{i=1}^k n_i^2 \frac{p_i q_i}{n_i - 1}, \quad (q_i = 1 - p_i), \quad (3.74)$$

since $p_i q_i / (n_i - 1)$ is an unbiased estimator of PQ/n_i . An alternative variance estimator, which is more efficient, is $v'(p) = pq/(n-1)$, where $q = 1 - p$. It may be noted that p is nothing but the sample proportion in the combined sample of n units. Similarly, if p_1, p_2, \dots, p_k are sample proportions based on k independent samples of sizes n_1, n_2, \dots, n_k , each of which is selected from the population with srs wr, the combined estimator is given by (3.72) and its variance and unbiased estimator of variance are

$$V(p) = \frac{PQ}{n^2} \sum_{i=1}^k \frac{N-n_i}{N-1} n_i, \quad (3.75)$$

$$v(p) = \frac{1}{n^2} \sum_{i=1}^k \frac{(N-n_i)n_i^2}{N(n_i-1)} p_i q_i, \quad (3.76)$$

since $V(p_i) = \frac{N-n_i}{N-1} \frac{PQ}{n_i}$ and $\frac{n_i p_i q_i}{n_i - 1}$ is an unbiased estimator of $N P Q / (N-1)$ for a sample of size n_i selected with srs wr.

In case of both srswr and srs wor, the variance of the arithmetic mean of the sample proportions, namely $p' = \frac{1}{k} \sum_{i=1}^k p_i$, is unbiasedly estimated by

$$v(p') = \frac{1}{k(k-1)} \sum_{i=1}^k (p_i - p')^2. \quad \dots \quad (3.77)$$

REFERENCES

- BASU, D. (1958) : On sampling with and without replacement; *Sankhyā*, 20, 287-294.
- COCHRAN, W. G. (1963) : *Sampling Techniques*; Chapters 2 and 3, John Wiley & Sons, New York.
- CORNFIELD, J. (1944) : On samples from finite populations; *J. Amer. Stat. Assn.*, 39, 236-239.
- DES RAJ and KHAMIS, H. S. (1958) : Some remarks on sampling with replacement; *Ann. Math. Stat.*, 29, 550-557.
- MURTHY, M. N. (1962) : Variance and confidence interval estimation; *Sankhyā*, 24, (B), 1-12.
- PATHAK, P. K. (1962) : On simple random sampling with replacement; *Sankhyā*, 24, (A), 287-302.

COMPLEMENTS AND PROBLEMS

3.1 In selecting 3 units with srs wor from a population having 6 units with the values 0.1, 0.5, 0.8, 1.2, 1.5, 1.9, show that the sample mean is unbiased for the population mean by enumerating all possible samples. Calculate its sampling variance and verify that it agrees with the variance formula given in (3.26).

3.2 In a sample of 50 households drawn with srs wor from a village consisting of 250 households, only 8 households were found to possess a bicycle. These had 3, 5, 3, 4, 7, 4, 4 and 5 members respectively. Estimate unbiasedly the total number of households in the village possessing a bicycle (H_b) as well as the total number of persons in such households (P_b). Also estimate the rse's of these estimates by using the unbiased estimates of their variances.

3.3 The frequency distribution of 232 cities in a country by population size is given in Table 3.6. Calculate the rse of the estimator of the total population Γ (i) when a sample of 60 cities is selected with srs w/or and (ii) the two largest cities are definitely included in the survey and only 48 cities are drawn from the remaining 230 cities with srs w/or.

TABLE 3.6 FREQUENCY DISTRIBUTION OF 232 CITIES BY POPULATION SIZE (000)

population size class	no of cities	population size class	no of cities	population size class	no of cities
(1)	(2)	(1)	(2)	(1)	(2)
50 — 75	81	500 — 550	2	1800 — 1850	1
75 — 100	45	550 — 600	3	1850 — 1900	0
100 — 150	42	600 — 650	1	1900 — 2000	1
150 — 200	14	650 — 700	1	2000 — 2050	0
200 — 250	9	700 — 750	0	2050 — 2100	1
250 — 300	5	750 — 800	1	2100 — 2300	0
300 — 350	6	800 — 850	2	2300 — 2350	1
350 — 400	5	850 — 900	1	2350 — 2550	0
400 — 450	5	900 — 950	2	2550 — 2700	1
450 — 500	2	950 — 1500	0		

3.4 Obtain an estimate of the percentage (P) of unemployed persons in a large city using the data given in Table 3.7 and estimate its rse.

TABLE 3.7 ESTIMATES OF P BASED ON TWO SIMPLE RANDOM SAMPLES SELECTED WITH REPLACEMENT

sample number	sample size no of persons	percent unemployed
(1)	(2)	(3)
1	2345	19.678
2	1789	20.123

3.5 Using the information given in Table 3.8 find the rse that one would expect if a sample of one percent of the villages in a group of districts in a State is selected with srs w/or for estimating the total population Γ of the region.

TABLE 3.8. NUMBER OF VILLAGES, AVERAGE POPULATION AND STANDARD DEVIATION FOR A GROUP OF DISTRICTS IN A STATE.

district sr. no.	no. of villages	population per village	standard deviation
(1)	(2)	(3)	(4)
1.	1953	487	564
2.	1664	829	931
3.	1381	822	996
4.	1174	1083	1167
5.	531	1956	1940
6.	1391	664	625
7.	1996	456	779
8.	1951	372	556
9.	3369	339	591

3.6 Data relating to land holdings, viz., size, cultivated area and consumption of chemical fertilizers, collected in an agricultural survey, are given in Table 3.9 for a sample of 36 holdings selected with srswr from a population of 432 holdings in a village.

- (i) Estimate the proportions of holdings P_1, P_2, P_3 and P_4 in the four holding size classes 0–4.99, 5.00–9.99, 10.00–24.99 and 25 & above.
- (ii) Estimate for each holding size class the total cultivated area (A) and the total consumption of fertilizers (F).
- (iii) Estimate the rse's of the estimates obtained in (i) and (ii).

TABLE 3.9. DATA ON SIZE, CULTIVATED AREA AND CONSUMPTION OF FERTILIZERS FOR 36 SAMPLE HOLDINGS.

sr. no. of holding	holding size in acres	cultiva- ted area (acres)	ferti- lizer (lbs.)	sr. no. of holding	holding size in acres	cultiva- ted area (acres)	ferti- lizer (lbs.)
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	21.04	2.70	0	19	3.15	2.60	0
2	12.59	1.76	0	20	4.84	1.67	36
3	20.30	1.47	48	21	9.07	5.13	406
4	16.16	1.64	54	22	3.69	3.69	0
5	23.82	1.56	0	23	14.61	1.34	0
6	1.79	1.79	0	24	1.10	0.25	0
7	26.91	5.44	544	25	22.13	2.70	248
8	7.41	1.97	117	26	1.68	1.21	56
9	7.68	2.45	0	27	49.58	4.48	216
10	66.55	3.26	0	28	1.68	1.38	42
11	141.80	3.20	197	29	4.80	1.39	66
12	28.12	3.90	192	30	12.72	2.15	180
13	8.29	1.95	0	31	6.31	2.12	74
14	7.27	2.20	160	32	14.18	1.49	182
15	1.47	0.48	0	33	22.19	3.80	222
16	1.12	1.12	0	34	5.50	1.75	192
17	10.67	1.98	0	35	25.29	5.17	224
18	5.94	2.45	0	36	20.99	1.50	40

(1 acre = 0.4047 hectare; 1 lb. = 453.6 grammes).

3.7 To estimate the total employment in an industry comprising of 70 factories, a sample of 10 factories was selected in the following manner. Three factories with relatively larger number of employees were definitely selected for the survey and a sample of 7 factories was drawn with srs wr from the remaining 67 factories. Two estimators of the total employment in that industry are proposed.

$$(i) \hat{Y}_1 = \frac{70}{10} \sum_{i=1}^{10} w_i \text{ and } (ii) \hat{Y}_2 = \sum_{i=1}^3 w_i + \frac{67}{7} \sum_{i=4}^{10} w_i.$$

where w_i denotes the number of employees in the i th sample factory, the first three factories being those definitely included in the survey. Compare the mse of \hat{Y}_1 with that of \hat{Y}_2 using the following data

total number of employees for the 3 big factories	8050
total number of employees for the remaining 67 factories	60810
sum of squares of the number of employees for the remaining 67 factories	71740058

3.8 In a finite bivariate population of N units, the means and standard deviations of the variables x and y are \bar{X} , \bar{Y} and σ_x , σ_y respectively and the correlation coefficient between x and y is ρ . Derive the correlation coefficient $\rho(\bar{x}, \bar{y})$ between the sample means x and y based on the same sample of n units selected with srs wr. Also find $\rho(x+\bar{y}, \bar{x}-y)$.

3.9 Suppose \bar{x} , \bar{y} and \bar{z} are the average productions (in kilogrammes) per field growing paddy, wheat and maize respectively based on samples of n_1 , n_2 and n_3 fields drawn with srs wr from N_1 , N_2 and N_3 fields growing these crops in a village

(i) Suggest an unbiased estimator of the total production P of these food grains in the village and derive its sampling variance

(ii) If the sale prices of paddy, wheat and maize are respectively C_1 , C_2 and C_3 units of money per 1000 kilogrammes, what would be the estimator of the total sale value of the whole produce of the village? Obtain an unbiased estimator of its sampling variance

3.10 Suppose in a population of N units, NP units are known to have the value 0. Obtain the relative efficiency E of selecting n units from N units with srs wr as compared to selection of n units from the $N - N_1$ non zero units with srs wr in estimating the population total Y .

3.11 If the value of the population coefficient of variation C ($= \sigma/\bar{Y}$) is known at the estimation stage, is it possible to improve upon the estimator \bar{y} , the sample mean based on a sample of n units selected with srs wr? If so, give the improved estimator

and obtain its efficiency E by comparing its mse with $V(\bar{y})$. (Hint : Try the estimator $\lambda \bar{y}$, where λ is a constant to be determined.)

(Goodman, L. A., *Ann. Math. Stat.*, 24, (1953), 114-117;

Searls, D. T., *J. Amer. Stat. Assn.*, 59, (1964), 1225-1226).

3.12 Suppose from a sample of n units selected with srs w/o r a sub-sample of n' units is selected with srs w/o r, duplicated and added to the original sample. Derive the expected value and the approximate sampling variance of \bar{y}' , the sample mean based on the $n+n'$ units. For what value of the fraction n'/n does the efficiency of \bar{y}' compared to that of \bar{y} attain its minimum value?

(Hansen, M. H., Hurwitz, W. N. and Madow, W. G., *Sample Survey Methods and Theory*, (1953), Vol. II, pp. 139-141).

3.13 Derive the results (3.18) and (3.19) given in Sub-section 3.3d regarding the sampling variance of the variance estimators.

3.14 In srs w/o r find the variance of y_i , the i -th sample observation ($i = 1, 2, \dots, n$) and the covariance between y_i and $y_{i'}$, ($i' \neq i$). Using these results, derive the variance of the sample mean \bar{y} .

3.15 A sample of $2n$ units is drawn from a large population with srs wr and the population mean is estimated by the sample mean \bar{y} . Assuming that y is normally distributed, compare the variances of the following estimators of $V(\bar{y})$:

$$(i) v_1 = s^2/2n, \quad (ii) v_2 = (\bar{y}_1 - \bar{y}_2)^2/4, \quad (iii) v_3 = (s_1^2 + s_2^2)/4n,$$

where $s^2 = \frac{1}{2n-1} \sum_{i=1}^{2n} \sum_{j=1}^n (y_{ij} - \bar{y})^2$, \bar{y}_1 and \bar{y}_2 are the means based on two random groups of n units into which the sample of $2n$ units has been divided and

$$s_1^2 = \frac{1}{n-1} \sum_{j=1}^n (y_{1j} - \bar{y}_1)^2 \quad \text{and} \quad s_2^2 = \frac{1}{n-1} \sum_{j=1}^n (y_{2j} - \bar{y}_2)^2,$$

y_{ij} denoting the j -th observation in the i -th random group.

3.16 A sample of 3 units is drawn from a population of N units with srs w/o r. Derive the probabilities of the sample containing $d = 1, 2, 3$ distinct units. Show that the arithmetic mean of the values of the d distinct units in the sample, \bar{y}_d , is an unbiased estimator of the population mean \bar{Y} . Also derive $V(\bar{y})$ and obtain an unbiased estimator of this variance, at least approximately.

3.17 Prove that the estimator (3.21) is unbiased for \bar{Y} and derive the results (3.22), (3.23), and (3.24) given in Section 3.4.

(Des Raj and Khamis, H. S., *Ann. Stat. Math.*, 29, (1958), 550-557;

Pathak P. K., *Sankhyā*, 24. (A), (1962), 287-302).

3.18 Suppose there are an unknown number (N say) of similar objects in a box serially numbered from 1 to N . For estimating the value of N , a sample of n objects is picked up at random after thorough shuffling. Derive an unbiased estimator of

N using the serial numbers on the selected objects and obtain its variance and an unbiased variance estimator.

3.19 For estimating the proportion P of a rare item, usually a sample random sample is selected unit by unit till a specified number m of units possessing the rare attribute is selected. Obtain an unbiased estimator of P . Derive a simple but approximate expression for its sampling variance and estimate this variance approximately.

(Haldane, J. B. S., *Biometrika*, 33, (1945), 222-225)

3.20 A sample of r fish was taken by netting from a lake, marked and released. On a subsequent occasion it was found that it was necessary to catch n fish one by one to get exactly m of the marked fish. Assuming the fish to be randomly distributed and that there has been no change in the population of fish between the two fishing periods, estimate unbiasedly the total number of fish (N) in the lake and obtain an unbiased estimator of the variance of the estimator.

(Bailey, N. T. J., *Biometrika*, 38, (1951), 293-306)

3.21 Suppose there are two lists one having M units and the other having N units, and it is required to estimate the total number of units D common to both the lists. For this purpose samples of m and n units, selected with srs w/o r from the two lists are compared and it is found that d units are common between the two samples. Noting that the number of common units d can be assumed to be distributed as

$$P(d) = \binom{D}{d} f^d (1-f)^{D-d}, \quad \left(f = \frac{mn}{MN} \right).$$

in the limiting case of M and N becoming infinitely large with $D, m/M$ and n/N remaining fixed, suggest an unbiased estimator for D and derive its sampling variance. Also obtain an unbiased estimator for this variance.

(Deming, W. E. and Glasser, G. J., *J. Amer. Stat. Assn.*, 54 (1959), 405-415)

Sample Selection and Sample Size

4.1 ILLUSTRATIVE POPULATIONS

A few populations, which are illustrative of the types of populations obtaining in practice, are considered here with a view to bringing out the remarkable stability of population coefficient of variation over time and space, and over characteristics of similar nature and studying the relationship between sample size and rse of the estimator. The population relating to the distribution of *land holdings*¹ by size given in Table 2.2 is used in this chapter to study the behaviour of the rse of the estimator with increase in sample size in case of srswr and srs wor.

The complete enumeration data for 128 *villages*² comprising a *tehsil*³ in Madras State, compiled during the 1951 and the 1961 censuses in India, are presented in Annexure 4.1. The village-wise data consist of information on geographical area, cultivated area, number of persons in 1951 and in 1961, number of cultivators, workers at household industry and number of households. The data relating to the volume of timber and the length of each strip in the Blacks Mountain Experimental Forest in California, U.S.A., based on a complete enumeration and reported by Hasel (1942), have been

¹ A *land holding* is defined as all land that is directed or managed by one or more persons, alone or with the assistance of others, without regard to title, size or location.

² A *village* is a well-defined socio-economic area unit consisting of households and plots (parcels of land).

³ A *tehsil* is an administrative unit comprising of villages and towns.

given in Annexure 4.2. In this case, the forest area is divided into 10 blocks and each block is sub divided into strips of uniform width. The values of some basic parameters of these three populations—land holdings, village data and forest statistics—are given in Table 4.1.

It may be pointed out that in many situations actually met with in practice, the population under consideration would, in general be considerably larger than the populations considered here and hence these populations may not be quite adequate to bring out the advantages and the problems involved in sampling from large populations.

TABLE 4.1 VALUES OF SOME PARAMETERS FOR THREE POPULATIONS

sr no	characteristic	population		standard deviation σ	coefficient of variation $(\sigma/\bar{X}) \times 100$
		total Σ	mean \bar{X}		
(1)	(2)	(3)	(4)	(5)	(6)
I Population of Land Holdings					
1	holding size (in acres)	335711 (10^3)	5431	10.979	202
II Population of Villages					
2	geographical area (in acres)	408125	3579	2207	62
3	cultivated area (in acres)	248752	1943	1107	57
4	number of persons—1951	415149	3243	1953	60
5	number of persons—1961	443319	3463	2065	60
6	number of cultivators	10948	854	572	67
7	workers at household industry	8968	70	94	134
8	number of households—1951	93129	728	433	59
III Population of Strips in a Forest					
9	length of strip (in 10 chain units)	1231	6.994	2.901	42
10	volume of timber (in 1000 board units)	49740	289.6	155.3	53

4.2 STABILITY OF COEFFICIENT OF VARIATION

It is of interest to note that the population coefficient of variation (cv) is usually fairly stable over time and over characteristics of similar nature. For instance, the cv's of number of persons in 1951 and 1961 (items 4 and 5 in Table 4.1) happen to be the same, though there have been some changes in the population mean and the standard deviation. Further, there is not much difference in the cv's of related

characteristics such as geographical area and cultivated area (items 2 and 3), and population and number of households (items 5 and 8). This stability of cv over time and over related characteristics makes it possible to determine the sample size for estimating a parameter with a specified margin of error or forecast the percentage error in the estimate of a parameter for a given sample size, utilizing the information available for a previous time period for the characteristic under consideration or for some related characteristic.

In the case of the population of villages given in Annexure 4.1, the cv is substantially higher for the number of workers at household industry than those for other characteristics. An examination of column (7) of Annexure 4.1 shows that a substantial portion of this variation is due to the village with the serial number 100, which reports an extreme figure of 846 persons in this category. If this unit is excluded from the population, the cv reduces to 100% from 134%.

Another point needing special attention is that the cv remains fairly stable for moderate enlargement of the size of the population by addition of similar units, which would mean that the sample size required to provide estimates with a given precision would not depend much on the number of units in the population, N . This is an important point, since one is apt to presume on *a priori* grounds that the population cv depends much on the size of the population. The fact that this presumption is not generally true in practical situations is borne out by the examples given in Tables 4.2 and 4.3.

In Table 4.2 the values of the population cv are given for the number of persons based on village-wise counts obtained in the 1951 and the 1961 censuses separately for each of four tehsils, and cumulated over these tehsils of a *district*⁴. From columns (3) and (7), it can be seen that the increase in the cv is far from being proportional to N . For instance, from columns (6) and (7) we see that while N increases from 119 villages in a tehsil to 761 villages in a district, an increase

⁴ A *district* is an administrative unit consisting of *tehsils*.

TABLE 4.2 COEFFICIENT OF VARIATION FOR NUMBER OF PERSONS
AND SAMPLE SIZES REQUIRED TO ENSURE 5% SAMPLING ERROR

sr no of tehsil†	no of villages N	coefficient of variation 100 C	sample size n	sampling fraction 100 f	cumulated over tehsils			
					N	100 C	n	100 f
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1951 census data								
1	119	66	71	60	119	66	71	60
2	126	67	75	60	245	67	104	43
3	228	88	132	58	473	81	170	36
4	288	92	156	55	761	86	214	29
1961 census data								
1	123	67	74	60	123	67	74	60
2	133	69	79	59	256	68	108	42
3	241	90	139	58	497	83	178	36
4	309	96	169	55	806	89	228	28

† name of tehsil—1 Gohna 2 Rohtak, 3 Sonipat, 4 Jhajjar

Source District Census Handbook (1951 Census) and Primary Census Abstract (1961 Census) for Rohtak District of Punjab State in India

TABLE 4.3 COEFFICIENT OF VARIATION FOR NUMBER OF WORKERS
AND SAMPLE SIZES REQUIRED TO ENSURE 5% SAMPLING ERROR

sr no of factories**	no of factories** N	coefficient of variation 100 C	sample size n	sampling fraction 100 f	cumulated over States			
					N	100 C	n	100 f
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<i>industry ginning and pressing</i>								
1.	179	50	64	36	179	50	64	36
2	305	89	156	51	484	74	151	31
3	377	53	87	23	861	66	145	17
4	163	90	109	67	1021	70	165	16
<i>industry miscellaneous food products</i>								
1	537	70	144	27	537	70	144	27
2	424	75	147	35	961	80	202	21
3	129	74	81	63	1090	75	187	17
5	130	64	73	56	1220	74	186	15
<i>industry machinery except electricals</i>								
1	287	69	115	40	287	69	115	40
2	413	62	112	27	700	65	135	10
4	719	57	110	15	1419	61	135	10
5	231	64	96	42	1650	62	141	9

*States 1 Madras, 2 Maharashtra, 3 Mysore, 4 Punjab, 5 Uttar Pradesh

**employing 10 to 49 workers and using power and employing 20 to 99 workers and not using power

Source: List of Registered Factories in India, 1961

of about 5.4 times, that is, 540%, the cv has increased from 66% to 86%, an increase of only about one-third, that is, 30%. Further, it is of interest to note that here also the cv has remained quite stable even after a period of ten years.

In Table 4.3 the cv's for a different type of population, namely number of workers in factories, are given separately for each of four States and cumulated over States in respect of three industries. Besides noting the points already mentioned in connection with Table 4.2, we note from columns (3) and (7) of Table 4.3 that the cv need not necessarily be more for a larger population.

4.3 BEHAVIOUR OF SAMPLE SIZE

Since the rse of the sample mean, \bar{y} , based on a sample of n units selected with srs wor, is given by

$$C(\bar{y}) = \sqrt{\frac{N-n}{N-1}} \frac{C}{\sqrt{n}},$$

where C is the population coefficient of variation, the sample size required to ensure an rse of $e\%$ is given by

$$n = \frac{NC^2}{(N-1)e^2 + C^2}. \quad \dots \quad (4.1)$$

The sample sizes required to ensure 5% rse for the estimator of \bar{Y} in the case of srs wor separately for each tehsil and for cumulated groups of tehsils are shown in columns (4) and (8) of Table 4.2. Corresponding figures in the case of number of workers separately for each State and for cumulated groups of States are given in columns (4) and (8) of Table 4.3.

From columns (6), (8) and (9) of Tables 4.2 and 4.3, it is clear that n , required for getting estimates with a given precision, does not increase in proportion to N and that, in general, the sampling fraction f , required to ensure a given precision, decreases with increase in N . This shows that it is usually relatively more economical to obtain estimates of a given precision for a population as a whole

than to ensure the same precision separately for each part of it. For instance, if estimates with 5% precision are required for each tehsil separately, the total sample size needed is 434 villages, amounting to a sampling fraction of 57% as compared to a sample size of only 214 villages, amounting to a sampling fraction of 28%, required to ensure 5% precision for the district as a whole.

Another example to illustrate the relationship between n and the rse of the estimator is presented in Table 4.4. The data considered here relate to the 1951 Census of Livestock carried out in Yugoslavia. The number of agricultural holdings in the country was about 2.39 millions. The sample size in terms of number of holdings required for a desired rse has been worked out on the basis of srs (Zarkovich, 1961). From this table we see that the decrease in the rse becomes incommensurate with the increase in n . For instance, in estimating the number of horses, the increase in n is about 1300 for a reduction of 50% in the rse i.e., from 10% to 5%, whereas for a reduction from 2% to 1% the increase required in n is as much as 32500. It may also be observed that it is possible to get estimates with less than or about 2% rse for all the four characteristics with a sample size of about 12300 which is just about 0.5% of the total number of holdings.

TABLE 4.4 SAMPLE SIZES REQUIRED TO GET LIVESTOCK ESTIMATES WITH SPECIFIED SAMPLING ERRORS

sr no	characteristic	relative standard error (%)				
		10	5	3	2	1
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	horses	434	1735	4821	10846	43385
2	cattle	124	494	1372	3088	12351
3	sheep	493	1972	5479	12327	49308
4	pigs	257	1030	2662	6439	25755

Source Zarkovich S S (1961) *Sampling Methods and Censuses*, Part I, Table I FAO, Rome,

4.4 SAMPLING ERROR AND EFFICIENCY

We have already studied the behaviour of the sampling variance as n increases in case of srswr and srswor. The behaviour of the rse's and of the relative efficiencies of sampling with and without replacement with increase in n can be illustrated by using the three populations considered in Table 4.1. Let $C_1(\bar{y})$ be the rse of \bar{y} based on all the n units in the sample including repetitions, $C_1(\bar{y}')$ that of \bar{y}' based on the distinct units in the sample in the case of srswr, and $C_2(\bar{y})$ that of the sample mean \bar{y} in the case of srs wor. The expressions for these are given by

$$C_1(\bar{y}) = \frac{C}{\sqrt{n}}, \quad \dots \quad (4.2)$$

$$C_1(\bar{y}') = \sqrt{\frac{N}{N-1} \left\{ 1 - \frac{n}{2N} + \frac{n(n-1)}{12N^2} \right\}} \frac{C}{\sqrt{n}} \quad \dots \quad (4.3)$$

neglecting terms of degree greater than $(1/N)^2$, and

$$C_2(\bar{y}) = \sqrt{\frac{N-n}{N-1}} \frac{C}{\sqrt{n}}, \quad \dots \quad (4.4)$$

where C is the population cv. The values of $C_1(\bar{y})$, $C_1(\bar{y}')$ and $C_2(\bar{y})$ for various sample sizes are calculated in respect of the different populations and are shown in Table 4.5. The efficiencies E and E' of \bar{y} and \bar{y}' in the case of srswr compared to that of \bar{y} in srs wor are given by

$$E = \frac{V_2(\bar{y})}{V_1(\bar{y})} = \frac{C_2^2(\bar{y})}{C_1^2(\bar{y})} = \frac{N-n}{N-1} \quad \dots \quad (4.5)$$

and

$$E' = \frac{V_2(\bar{y}')}{V_1(\bar{y}')'} = \frac{C_2^2(\bar{y}')}{C_1^2(\bar{y}')'} = \frac{12N(N-n)}{6N(2N-n)+n(n-1)}. \quad \dots \quad (4.6)$$

This shows that the efficiencies E and E' are independent of the population coefficient of variation and that these are functions of only n and N . The values of the efficiencies for different sample sizes are also presented in Table 4.5.

TABLE 4.5 RELATIVE STANDARD ERRORS (%) OF SAMPLE MEAN FOR DIFFERENT SAMPLE SIZES

no.	characteristic	σ_x^2 (%) [*]	method ^{**}	sample size								
				5	10	15	20	25	50	75	100	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	
I Population of land holdings Number of holdings 61780 (10 ³)												
1 holding size	202	1	90.34	63.88	52.16	45.17	40.40	28.57	23.32	20.20	20.20	
		2	90.31	63.88	52.16	45.16	40.40	28.56	23.32	20.19	20.19	
		3	90.31	63.88	52.15	45.16	40.39	28.55	23.31	20.18	20.18	
		F	100.00	100.00	99.98	99.97	99.96	99.92	99.88	99.82	99.81	
		F'	100.00	100.00	99.99	99.99	99.99	99.98	99.96	99.94	99.92	
II Population of villages Number of villages 128												
2 geographical area	62	1	27.73	19.61	16.01	13.86	12.40	8.77	7.18	6.20	6.20	
		2	27.66	19.30	16.61	13.38	11.85	7.96	6.16	5.06	5.06	
		3	27.29	18.90	15.10	12.79	11.17	6.87	4.62	2.91	2.91	
3 number of persons (1961)	60	1	26.83	18.97	15.49	13.42	12.00	8.49	6.93	6.00	6.00	
		2	26.67	18.68	15.10	12.95	11.46	7.70	5.98	4.89	4.89	
		3	26.40	18.29	14.01	12.37	10.81	6.65	4.48	2.82	2.82	
4 number of culti- vators	67	1	29.96	21.10	17.30	14.08	13.40	9.48	7.74	6.70	6.70	
		2	29.79	20.86	16.86	14.46	12.80	8.60	6.66	5.48	5.48	
		3	29.49	20.42	16.32	13.82	12.07	7.43	5.00	3.15	3.15	
5 workers at house- hold industry	134	1	59.93	42.37	34.60	29.96	28.80	18.95	15.47	13.40	13.40	
		2	59.57	41.71	33.72	28.91	25.60	17.20	13.32	10.93	10.93	
		3	58.97	40.81	32.61	27.63	24.14	14.85	10.00	6.29	6.29	
		F	96.85	92.91	88.98	85.01	81.10	61.42	41.73	22.05		
		F'	93.00	95.89	93.67	91.33	88.87	74.57	56.32	39.16		
III Population of strips in a forest Number of strips 176												
6 length of strip	42	1	18.78	13.28	10.84	9.30	8.40	5.94	4.85	4.20	4.20	
		2	18.70	13.13	10.64	9.15	8.13	5.54	4.36	3.63	3.63	
		3	18.57	12.93	10.40	8.87	7.80	5.01	3.68	2.77	2.77	
7 volume of timber	55	1	24.60	17.39	14.20	12.30	11.00	7.78	6.35	5.50	5.50	
		2	24.49	17.19	13.94	11.98	10.64	7.25	5.70	4.75	4.75	
		3	24.31	16.04	13.62	11.61	10.22	6.60	4.82	3.62	3.62	
		F	97.71	91.86	92.00	89.14	86.29	72.00	57.71	43.43		
		F'	98.55	97.06	95.40	93.87	92.20	82.81	71.56	53.16		

* population coefficient of variation

** 1— $s_{\bar{x}w}\sigma_w$, the estimator being the sample mean2— $s_{\bar{x}w}\sigma_w$, the estimator being the sample mean based on distinct units,3— $s_{\bar{x}w}$, the estimator being the sample mean

$$E = V_2(\bar{y})/V_1(\bar{y}), \quad F' = V_2(\bar{y})/V'_1(\bar{y}')$$

From this table it is clear that the decrease in rse, though substantial for initial increases in n , is not commensurate with increases in n beyond a certain stage, and that the loss in efficiency in using srswr instead of srs wr increases as n increases, but becomes marginal for very large populations. It may be noted that when srs is used, the sample size has to be increased to about k^2 times its original value in order to reduce the relative standard error to $(1/k)$ th of its original value.

4.5 PROCEDURES OF SELECTION

One procedure of selecting a unit from a population of N units with equal probability consists in associating the units with N identical counters marked with numbers from 1 to N such that each numbered counter is associated with one and only one unit and then selecting one counter after thoroughly shuffling the counters. The unit corresponding to the number marked on the selected counter is considered to be sampled. It can be intuitively felt that with this selection process the proportion of times a particular unit gets selected in a large number of draws will be very near $(1/N)$ and this is its probability of selection. Since this probability is the same for all the units in the population in this case, this selection procedure will result in srs.

4.5a RANDOM NUMBER TABLES

The procedure of numbering each counter and selecting a counter after proper shuffling becomes tedious and cumbersome when the number of units in the population is large. Further, it is rather difficult to achieve thorough shuffling in practice. To overcome this difficulty tables of random numbers have been prepared. The random numbers are usually generated by some mechanism which, when repeated a large number of times, ensures approximately equal frequencies for the numbers from 0 to 9 and also proper frequencies for various combinations of numbers (such as 00, 01, ..., 99; 000, 001, ..., 999; etc.) that could be expected in a random sequence of the digits 0 to 9.

Several standard tables of random numbers are available, among which the following may be specially mentioned, as they have been tested extensively for randomness :

- (i) Tippett's (1927) random number tables consisting of 41600 random digits grouped into 10400 sets of four-digited random numbers;
- (ii) Fisher and Yates (1938) table of random numbers with 15000 random digits arranged into 1500 sets of ten-digited random numbers;
- (iii) Kendall and B. B. Smith (1939) table of random numbers having 100000 random digits grouped into 25000 sets of four-digited random numbers;
- (iv) Rand Corporation (1955) table of random numbers consisting of 1000000 random digits grouped into 200000 sets of five-digited random numbers; and
- (v) C. R. Rao, Mitra and Matthai (1966) table of random numbers with 20000 random digits arranged into 5000 sets of four-digited random numbers.

Two pages from (v) are reproduced in Appendix 1 to help in illustrating the use of random numbers for selecting a simple random sample.

4.5b ASSOCIATION OF ONE NUMBER

Suppose one village is to be drawn with equal probability from the population of 128 villages given in Annexure 4.1 of Chapter 4. The 128 villages are given running serial numbers from 1 to 128 so that each village is associated with one and only one of the numbers 1 to 128 and can, therefore, be uniquely identified by its serial number. Then the problem of selecting one village with equal probability reduces to that of selecting one of the numbers 1 to 128 at *random*. A three-digited column of random numbers is consulted, since the highest serial number that can be selected is 128, a three-digited number, and the village corresponding to the first number, which is less than or

equal to 128, is chosen. For this purpose, the numbers 001 to 128 will be considered for selection, a number like 008 being interpreted as 8 and the number 000 being ignored. Starting from the first column of four-digited random numbers given in Appendix 1 and confining our attention to only the first three digits from the left, we find that the first number, which is less than or equal to 128 and which is different from 000 is 112. Hence, the village with serial number 112 is considered as a sample of one unit selected with equal probability.

Suppose a sample of 10 villages is to be drawn with srs wr from the population of 128 villages for estimating the population mean or total of some characteristic. By referring to the random numbers given in Appendix 1 (columns 1 and 2), we find that the first ten three-digited numbers, which are less than or equal to 128 and different from 000, are 112, 059, 112, 116, 124, 090, 037, 078, 092, 062. The values of the characteristics under consideration for the sample villages having these numbers as serial numbers are given in Table 4.6. The various estimates and estimates of their standard errors and rse's are calculated and shown in Table 4.6 itself.

It may be noted that in the sample selected above, the village with serial number 112 has occurred twice. If instead we want a sample of 10 villages selected with srs wor, we should check up at each draw whether the unit selected in that draw has been selected in some previous draw and, in case of repetition, we have to select a fresh unit. In the example considered above, the number 112 got repeated at the 3rd draw. In this case, the number 112 is rejected and the numbers 116, 124, 090, 037, 078, 092, 062 and 077 are selected at the 3rd to 10th draws respectively. The estimates of population means of various characteristics and the estimates of their standard errors and rse's obtained from this sample are also shown in Table 4.6.

In this connection, it may be mentioned that if there has been a gap in the serial numbers already given, it is not necessary to re-number the units before selection with equal probability, provided there is no duplication of the same number. That is, suppose in the serial numbers originally given to the 128 units, the numbers

26 and 48 are missed and hence the last serial number is 130, then one unit can be selected with equal probability by selecting a number at random from 1 to 130 and accepting the unit corresponding to the selected number except in case the number is 26 or 48, when the process of selection is repeated.

TABLE 4.6 SIMPLE RANDOM SAMPLES OF 10 VILLAGES SELECTED WITH AND WITHOUT REPLACEMENT

sr no	order of selection		vill age sr.no	1951 census			1961 census			no of house-holds
	srs wr	srs wr		area (sq mile)	culti vated area*	number of persons	number of persons	culti vators	workers in industry	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	1	1	112	1 77	428	695	925	181	1	193
2	2	2	59	3 57	1314	1643	1808	394	36	381
3	3	—	112	1 77	428	695	925	181	1	193
4	4	3	116	3 56	1328	2639	2740	917	6	549
5	5	4	124	2 75	772	1577	1757	402	100	373
6	6	5	90	0 94	509	1146	1154	217	24	203
7	7	6	37	5 28	2622	2654	2768	1039	107	599
8	8	7	78	3 31	950	1010	1021	170	5	196
9	9	8	92	5 56	1881	953	1223	1	0	261
10	10	9	62	5 01	1053	4020	3652	955	51	715
11	—	10	77	4 39	1414	2144	2373	373	62	528

simple random sampling with replacement

\bar{y}	3 352	1132	1703	1797	446	33 10	371
$s(y) = s/\sqrt{n}$	0 501	221	342	301	120	12 92	60
$c(\bar{y}) = 100s(y)/\bar{y}$	14 95	19 52	20 08	16 75	26 91	39 03	16 17

simple random sampling without replacement

\bar{y}	3 614	1230	1878	1942	465	39 20	405
$s(\bar{y}) = \sqrt{1-f}(s/\sqrt{n})$	0 458	199	317	277	113	12 17	56
$c(\bar{y}) = 100s(\bar{y})/\bar{y}$	12 67	16 18	16 88	14 26	24 30	31 05	13 83

* in acres, 1 square mile = 640 acres, 1 acre = 0.4047 hectare

4.5c ASSOCIATION OF SEVERAL NUMBERS

It may be noted that in the above procedure of selection many random numbers get rejected, since all three-digited numbers greater than 128 and the number 000 are not considered at the time of selection. If a large sample is to be selected, this procedure may be time-consuming. In fact the expected proportion of random numbers which would be rejected in this case is 0.872, which is quite high. This difficulty can be overcome to a large extent by associating with each unit (village, in this case) an equal number of numbers from 1 to 896, which is the highest three-digited multiple of 128, that is, [1000/128] 128, such that no two units have the same numbers associated with them and by selecting a particular unit when any one of the numbers associated with it gets selected in choosing a number at random from 1 to 896. The number of numbers to be associated with each unit in this case is 7, which is the quotient in dividing 1000 by 128, that is, [1000/128]. It is clear that this procedure leads to equal probability of selection, for the probability of selection of any particular unit is 7/896, that is, 1/128, in the present example, since out of 896 possible numbers 7 are favourable for the selection of that unit. In this case, the expected proportion of random numbers which would get rejected is clearly only 0.104. It may be pointed out that there are different ways of associating 7 numbers from 1 to 896 with each of the 128 units and it is desirable to select that system which is operationally most convenient.

In general, to take a random number from 1 to N , where N is an r -digited number, a number is chosen at random from 1 to N' , the highest r -digited multiple of N . If N' is kN , then k numbers from 1 to N' are associated with each unit in the population and the unit corresponding to the random number selected from 1 to N' is selected. The proportion of numbers rejected in the general case is given by $(10^r - N')/10^r$. It may be noted that N' need not necessarily be an r -digited number and that it would be advantageous to make it an r' -digited multiple of N ($r' > r$) provided the difference between N' and the highest r' -digited number is very small.

Method 1 (Remainder Approach)

One way of associating 7 numbers from 1 to 896 with each unit is to assign to the s th unit ($s = 1, 2, \dots, 128$) the 7 numbers

$$\{s+128j\}, \quad j = 0, 1, 2, \dots, 6,$$

that is

$$s, s+128, s+256, \dots, s+768$$

A number R is selected at random from 001 to 896 and the serial number of the unit to be selected is given by the remainder obtained on dividing R by 128. In case the remainder is 0 the unit to be selected is 128. In general, if N is an r digit number, a number R is chosen at random from 1 to N , the highest r digit multiple of N , and the unit having the serial number equal to the remainder obtained on dividing R by N is considered as selected.

Method 2 (Quotient Approach)

Another way of associating 7 numbers with each of the 128 units is to associate with the s th unit U_s , ($s = 1, 2, \dots, N$), the 7 numbers

$$\{7(s-1)+j\} \quad j = 0, 1, 2, \dots, 6$$

In this case, a number R is chosen at random from 000 to 895 and the s th unit is selected if the quotient obtained on dividing R by 7 is $(s-1)$. For instance, if the random number selected is 492, the quotient on division by 7 is 70 and hence the 71st unit is to be selected. Comparing the two methods of associating the random numbers with the units in the population, it can be seen that *method 2* may be preferred as it involves only division by 7, which is simpler than division by 128 required in *method 1*. In general, if N is an r digit number, a random number R is chosen from 0 to $N-1$, N being the highest r digit multiple of N , and the unit U_s is selected if the quotient $(s-1)$ is obtained on dividing R by L , where $L = N/V$, treating the number N to be 0.

Method 3 (Modified Quotient Approach)

A third system of associating 7 numbers from 1 to 896 with each unit, which is a modification of *method 2* and which is likely to be more convenient than the two procedures given above, consists in associating the numbers

$$s, s+100, s+200, \dots, s+600$$

with the s th unit for $s = 1, 2, \dots, 100$ and the numbers

$$7(s-1), 7(s-1)+1, \dots, 7(s-1)+6,$$

for $s = 101, 102, \dots, 128$ with the exception that the number 896 is associated with the 101th unit instead of $7(s-1) = 700$. If the random number selected is less than or equal to 700, treating 000 as 1000, the serial number of the unit to be selected is given by the last two digits 00 standing for 100, without need for any division as in the first two procedures. However, if the selected random number is greater than

700 and less than or equal to 896, the serial number of the unit to be selected is to be taken as the integer next to the quotient, obtained on dividing the random number by 7, treating the number 896 as giving rise to selection of the 10th unit.

The effectiveness of associating severral numbers with each unit can be illustrated by sampling 10 units with equal probability from the population of 128 villages using *method 3* suggested here. We have seen earlier that we have to use up the 50 random numbers of column (1) and 11 numbers of column (2) to select 10 units using the procedure of rejecting all random numbers greater than 128. When *method 3* is used, the units that get selected in the sample are 43, 13, 80, 29, 72, 110, 49, 128, 12 and 33 showing that it has been possible to select 10 units by referring to only the first 11 random numbers instead of the 61 random numbers referred to in the previous case.

Method 4 (Modified Remainder Approach)

A modification of *method 1* is given here. Suppose the number of units in a population, N , is an r -dugited number. Let N' be the *nearest* higher r -dugited *rounded number*, which is not less than N and which can be conveniently used as the divisor in dividing any number. Let N'' be the highest r -dugited multiple of N' . Then if R is an r -dugited number selected at random using a table of r -dugited random numbers, the serial number of the unit to be selected will be the number got as remainder on dividing the selected random number R by N' provided R is not greater than N'' and the remainder is neither 0 nor greater than N . In all other cases, the selected number is to be rejected and the process is to be repeated. In case N' is equal to N , this method reduces to *method 1* and in that case, if the remainder is 0, the serial number of the unit to be selected is to be taken as N .

In the example where N is 128, N' may be taken as 200 and N'' as 1000, as it will be convenient to divide any random number selected from 1 to 1000 by 200. This amounts to associating with the i -th unit the five numbers i , $i+200$, $i+400$, $i+600$ and $i+800$ for $i = 1, 2, \dots, 128$. Since selection of any number in the ranges 129–200, 329–400, 529–600, 729–800, 929–1000 would result in the rejection of that random number, the proportion of numbers rejected during the process of selection will be 0.360. In the general case, the proportion of rejections is given by $\{(N' - N)/10^r\}(N''/N')$.

Method 5 (Independent Choice of Digits)

Another method, which is suggested by Matthai (1954), consists in making up the random number by combining two random numbers one referring to the first digit and the other relating to the other digits. For instance, if a number is to be selected at random from 1 to 498, two numbers are selected, one from 0 to 4 and the other from 00 to 99. Suppose the numbers are 3 and 86, the number chosen is 386. But if the number made up by combining the two random numbers is greater than 498 or is 000 it is rejected and the operation is repeated. In this example, a number from 0 to 4 can be selected by referring to a column or row of single-digit random numbers without any rejection by associating (0,5), (1,6), (2,7), (3,8), (4,9) with the numbers 0, 1, 2, 3, 4 respectively and hence the number of rejections would be very small,

4.6 CHANGES IN SAMPLING FRAME

Suppose a sampling frame of N units has undergone some changes in the sense that m units of the original frame have ceased to exist and m' new units have been added to the frame. Then the revised frame will consist of $N-m+m'$ (say, N') units. For selecting one unit with equal probability, the numbers 1 to N' may be associated with the N' units. This would mean renumbering of the units in the frame. This may be time consuming if N and N' are both large. The renumbering could be avoided by the following procedure. The old numbering is kept as such and the new units are given numbers from $(N+1)$ to $(N+m')$. A random number is selected from 1 to $N+m'$. The unit corresponding to this number is selected provided it is not one of the m units which have become non-existent. In case a non-existent unit is chosen, the draw is rejected and the procedure is repeated. It may be seen that this procedure gives equal probability to the $N-m+m'$ units in the revised frame.

Any particular unit, say U_i , may be selected in the first draw itself, the probability for which is $1/(N+m')$, or one of the m non-existent units is selected in the first draw resulting in its rejection and then the unit U_i may be selected in the second draw, the probability for which is $m/(N+m')$, and so on. Hence, the probability $P(U_i)$ of selecting the unit U_i in the first effective draw is given by

$$\begin{aligned}
 P(U_i) &= \frac{1}{N+m'} \left\{ 1 + \left(\frac{m}{N+m'} \right) + \left(\frac{m}{N+m'} \right)^2 + \dots \right\} \\
 &= 1/(N-m+m'), \quad . \quad (47)
 \end{aligned}$$

since $1+x+x^2+\dots = 1/(1-x)$, for $|x| < 1$

In this connection it may be mentioned that it is possible to reduce substantially the number of rejections of draws by giving all or some of the serial numbers of non-existent units to some or all of the new units. That is, if $m > m'$, then the serial numbers of m' of the m non-existent units may be assigned to the m' new units and if $m < m'$,

then the serial numbers of all the m non-existent units may be assigned to the first m of the m' new units, giving the fresh serial numbers $N+1$ to $N+m'-m$ to the remaining $m'-m$ new units.

These procedures (Sections 4.5 and 4.6) are given here just to illustrate the principle of equal probability selection and to indicate the possibility of effecting considerable reduction in the rejection of random numbers in the process of selection and of achieving substantial operational convenience by using a suitable system of associating the random numbers with the units. The essential point to be kept in view in equal probability sampling is that the same number of numbers is to be associated with each unit and that the numbers are to be selected at random.

4.7 HAPHAZARD SELECTION

At this stage it may be mentioned that there may be a feeling among the uninitiated in random sampling methods that equal probability selection could be achieved by selecting the units in a haphazard manner, that is, without following rigidly any specified routine procedure and hence there is no need to refer to random number tables. This feeling is apparently based on the belief that it is possible for a person to be quite unbiased in generating a series of numbers that would serve the purpose of random number tables. The results of some class room experiments conducted to illustrate that haphazard selection is *not* random selection are given here. These examples may be taken as indicative of the types of biases present in haphazard selection and as illustrative of the fact that in many situations haphazard selection is likely to be far from random selection.

4.7a GENERATION OF 'RANDOM' NUMBERS

In the first experiment six persons were instructed to write down 100 three-digited numbers just as they occurred to their minds. The frequency distributions of the digits 0, 1, 2, ..., 9 in the set of numbers obtained for each of the six persons, are shown in Table 4.7. This table brings out clearly the existence of a tendency on the part of the persons participating in the experiment to prefer some numbers more often than others. This type of bias may be termed *number bias*.

TABLE 4.7 FREQUENCY DISTRIBUTIONS OF THE NUMBERS 0, 1, 2, ..., 9 FOR THE SIX PERSONS

number	serial number of person						expected frequency
	1	2	3	4	5	6	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0	50	1	38	29	34	59	30
1	29	48	30	57	33	27	30
2	20	19	28	31	20	22	30
3	50	39	34	34	24	24	30
4	55	40	28	29	15	27	30
5	20	18	31	15	30	25	30
6	30	26	26	27	31	15	30
7	12	39	32	35	42	35	30
8	25	42	30	23	44	37	30
9	9	28	23	20	27	29	30
total	300	300	300	300	300	300	300

4.7b USE OF COLOUR CHART

In another experiment four persons were given a colour chart consisting of a square block of 100 cells, each cell being coloured blue, green, red, white or yellow in such a way that in each quadrant any colour occurs only once in each row and only once in each column. Each cell can be identified by its row number and column number. Thus (4 6) means the cell in the 4th row and the 6th column. Each of the four persons was instructed to select 100 cells with replacement haphazardly and to write down the row and the column numbers of these selected cells with a view to examining whether colour has some influence on their selection of numbers. The frequency distributions of colours, obtained for the four persons, are presented in Table 4.8. This table shows that there is some colour bias present in the persons, who experimented with the colour chart. A closer study of the results revealed a tendency to ignore the border cells of the colour chart at the time of selection.

TABLE 4.8 FREQUENCY DISTRIBUTIONS OF COLOURS OF SELECTED CELLS FOR THE FOUR PERSONS

colour	serial number of person				expected frequency
	1	2	3	4	
(1)	(2)	(3)	(4)	(5)	(6)
blue	14	26	20	12	20
green	28	21	15	21	20
red	15	12	20	22	20
white	25	23	20	19	20
yellow	18	18	25	26	20

4.7c BIASES IN SELECTION

From these illustrations, it is clear that in selecting numbers or objects haphazardly the personal biases, conscious or unconscious, of the experimenter, such as preference for certain numbers or colours, etc., invariably make it impossible to ensure equal probability to the numbers 0, 1, 2, ..., 9 or to units in the population. In fact, it is almost impossible to predetermine the proportions with which various combinations of the numbers occur in a sequence of numbers generated by a person or by some other non-random process affected by different types of biases. Such a procedure will yield only a non-random sample, which is not amenable to objective methods of sampling and estimation. Examples of biased sampling have been given, among others, by Yates (1935) and Cochran and Watson (1936).

4.8 DETERMINATION OF SAMPLE SIZE

One of the main advantages of a sample survey is that it is possible to ensure approximately a pre-specified margin of error in the sample results by suitably fixing the sample size. The margin of error that is permissible in the estimate, which is termed *permissible error*, may be taken as the maximum difference or percentage difference between the estimate and the parameter value that can be tolerated on considerations of loss or gain due to policy decisions based on the sample results.

For instance, if an error d on either side of the parameter value can be tolerated in the estimation of the population mean, the permissible error is specified as

$$|\bar{y} - \bar{Y}| = d.$$

Since the quantity $|\bar{y} - \bar{Y}|$ differs from sample to sample, the margin of error is specified by the probability statement,

$$\text{Prob. } \{|\bar{y} - \bar{Y}| \leq d\} = 1 - \alpha, \quad \dots \quad (4.8)$$

where $(1 - \alpha)$ may be taken as 95%, 99% or some other desired level of confidence. Since specification of the margin of error in terms of absolute difference between the estimator and the parameter value

requires at least a dimensional idea of the parameter value, the permissible error is usually specified as the percentage difference between the estimator and the parameter value. In the above case, the permissible error is specified by the probability statement

$$\text{Prob} \left\{ \left| \frac{\bar{y} - \bar{Y}}{\bar{Y}} \right| \leq P_d \right\} = 1 - \alpha, \quad \left(P_d = \frac{d}{\bar{Y}} \right), \quad (4.9)$$

which means that an error of $100 P_d \%$ on either side of the parameter value \bar{Y} can be tolerated. Before considering the translation of the above statement of the permissible error in terms of the standard error and the rse of the estimator, the question of determining n so as to ensure a specified value for the rse of an estimator is briefly discussed.

4.8a FIXED RELATIVE STANDARD ERROR

We have seen that the rse's of the sample mean in case of srswr and srs wor are given by

$$C_1(\bar{y}) = \frac{C}{\sqrt{n}} \quad \text{and} \quad C_2(\bar{y}) = \sqrt{\frac{N-n}{N-1}} \cdot \frac{C}{\sqrt{n}},$$

where C stands for the population cv. Thus we see that a knowledge of the value of C ($= \sigma/\bar{Y}$) would enable us to determine n so as to get estimates with any prespecified value of the rse. For, if the permissible margin of error, measured here by the rse of the estimator, is pre-fixed as e , then the sample size required is given by

$$n' = C^2/e^2 \quad \dots \quad (4.10)$$

in the case of srswr, and in the case of srs wor by

$$n'' = \frac{NC^2}{(N-1)e^2 + C^2} = \frac{Nn'}{N+n'-1}. \quad \dots \quad (4.11)$$

In this connection it may be pointed out that the remarkable stability of the population cv over time and space, and over characteristics of similar nature, noticed in Section 4.2, is of considerable

help in determining the sample size in a sample survey, as in practice the exact value of C for the specific characteristic under study would not be available and hence the cv of this characteristic for some past period or of some other characteristic related to the variable under consideration may have to be used for this purpose. It may be noted that (4.11) is approximately equal to (4.10), when N is large, which is as it should be, for the difference between the efficiencies of sampling with and without replacement becomes negligible when N is large.

The sample sizes required to estimate the population mean and total with specified relative standard errors in the case of srswr have been presented in Table 4.9 for different values of C . From this table, it is clear that the relationship between n and the rse of the estimator is similar to that indicated earlier, namely, the decrease in the rse becomes incommensurate with increase in n after a certain stage.

TABLE 4.9. SAMPLE SIZES REQUIRED TO ENSURE SPECIFIED RELATIVE STANDARD ERRORS IN SAMPLING WITH REPLACEMENT.

value of rse $C(\bar{y})\text{(%)}$	value of the population coefficient of variation (%)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
25	1	1	1	3	4	6	8	10	13	16	36	64
20	1	1	2	4	6	9	12	16	20	25	56	100
15	1	2	4	7	11	16	22	28	36	44	100	178
10	1	4	9	16	25	36	49	64	81	100	225	400
5	4	16	36	64	100	144	196	256	324	400	900	1600
4	6	25	56	100	156	225	306	400	506	625	1406	2500
3	11	44	100	178	278	400	544	711	900	1111	2500	4444
2	25	100	225	400	625	900	1225	1600	2025	2500	5625	10000
1	100	400	900	1600	2500	3600	4900	6400	8100	10000	22500	40000

Since sampling without replacement is more efficient than sampling with replacement, especially when the population is not very large, the sample sizes required to ensure specified values for the rse in the case of srs wor have been given in Table 4.10 for different values

of C to provide an idea of the gain in using sampling without replacement instead of with replacement. It is of interest to note that the statement made earlier in the case of srswr that the decrease in the margin of error becomes incommensurate with increase in n after a certain stage is not strictly valid in srs wr in case of small and medium populations with a fairly large C .

TABLE 4.10 SAMPLE SIZES REQUIRED TO ENSURE SPECIFIED RELATIVE STANDARD ERRORS IN SAMPLING WITHOUT REPLACEMENT.

value of rse $C(\bar{y})(\%)$	value of population coefficient of variation (%)											
	10	20	30	40	50	60	70	80	90	100	150	200
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)

$N = 100$

25	1	1	1	3	4	5	7	9	12	14	27	39
20	1	1	2	4	6	8	11	14	17	20	36	50
15	1	2	4	7	10	14	18	22	27	31	50	61
10	1	4	8	14	20	27	35	39	45	50	69	80
5	4	14	27	39	50	59	66	72	77	80	90	94
4	6	20	36	50	61	69	76	80	84	86	92	96
3	10	31	50	64	74	80	84	88	90	92	96	98
2	20	50	69	80	86	90	93	94	95	96	98	99
1	50	80	90	94	96	97	98	98	99	99	100	100

$N = 500$

25	1	1	1	5	4	6	8	10	13	16	34	57
20	1	1	2	4	6	9	12	16	19	24	51	83
15	1	2	4	7	11	16	21	27	34	41	83	154
10	1	4	9	16	24	34	45	57	70	83	155	222
5	4	16	34	57	83	112	141	170	197	222	322	381
4	6	24	51	83	141	151	190	222	252	278	369	417
3	11	41	83	154	179	222	261	294	322	368	417	450
2	24	83	155	222	278	322	355	381	401	417	459	476
1	83	222	322	381	417	439	454	464	471	476	489	494

$N = 1000$

25	1	1	1	3	4	6	8	10	13	16	35	60
20	1	1	2	4	6	9	12	16	20	24	53	91
15	1	2	4	7	11	16	21	28	35	43	91	151
10	1	4	9	16	24	35	47	60	75	91	183	286
5	4	16	35	60	91	126	164	204	245	286	474	616
4	6	24	53	91	135	183	235	286	336	385	585	714
3	11	43	91	151	218	286	353	416	474	527	714	816
2	24	91	183	286	385	474	551	616	670	714	849	909
1	91	286	474	616	714	783	831	864	890	909	957	976

4.8b FIXED CONFIDENCE INTERVAL

The value of e , the desired value of the rse $C(\bar{y})$, is usually fixed in such a way that the probability of the percentage difference between the estimate and the parameter being less than a prespecified value (P_d , say) is $(1-\alpha)$ ($= 95\%$ or 99% , say). That is, e is determined such that (4.9) is satisfied. As mentioned before, if the sample size and the number of possible samples are fairly large, the sample mean \bar{y} is likely to be normally distributed with mean \bar{Y} and standard deviation $\sigma(\bar{y})$, in which case

$$\text{Prob.} \left\{ \frac{|\bar{y} - \bar{Y}|}{\sigma(\bar{y})} \leq k_\alpha \right\} = 1 - \alpha, \quad \dots \quad (4.12)$$

where k_α is the $\alpha\%$ point of the normal distribution tabulated in Table (2.1) with $P = 1 - \alpha$. The expected proportions of sample with \bar{y} below $\bar{Y} - k_\alpha \sigma(\bar{y})$ and above $\bar{Y} + k_\alpha \sigma(\bar{y})$ would each be $\alpha/2$. Rewriting (4.12) after multiplying both sides of the inequality in the brackets by $C(\bar{y}) = \sigma(\bar{y})/\bar{Y}$, we get

$$\text{Prob.} \left\{ \left| \frac{\bar{y} - \bar{Y}}{\bar{Y}} \right| \leq k_\alpha C(\bar{y}) \right\} = 1 - \alpha. \quad \dots \quad (4.13)$$

Comparing (4.13) with (4.9), we see that to ensure (4.9) we should fix e , the value of $C(\bar{y})$ that should be achieved, as

$$e = P_d/k_\alpha. \quad \dots \quad (4.14)$$

Once P_d and $(1-\alpha)$ are specified, e can be calculated, which together with a knowledge of C will determine the sample size required, as given in (4.10) and (4.11). In this case, the length of the confidence interval L is a constant and is equal to $2d$.

In practice, it is not possible to know the exact value of C at the planning stage and hence this may have to be approximated by the true value or an estimate of the population coefficient of variation as obtained in a previous survey for the characteristic under consideration or for some associated characteristic. If no such information is available, it would be necessary to carry out an exploratory

study of the characteristic under consideration or a related characteristic to get an idea of C . Thus we see that because of the uncertainty involved in the value of C used in the determination of the sample size n , the probability statement given in (4.9) will be satisfied only approximately. Since we have seen in Section 4.2 that the population coefficient of variation is fairly stable over time and over similar types of characteristics the approximation achieved in determining n by using the past data for the same or a similar type of characteristic is likely to be quite adequate in practice.

4.8c SAMPLE SIZE WHEN $E(L)$ IS SPECIFIED

Since the value of $C(\bar{y})$ is not known exactly, the confidence interval for \bar{Y} has to be based on an estimate of $\sigma(\bar{y})$ obtained from the sample. As has been pointed out earlier,

$$t = (\bar{y} - \bar{Y})/s(\bar{y}),$$

where $s(\bar{y})$ is a suitably chosen estimate of $\sigma(\bar{y})$, is generally distributed as Student's t with v degrees of freedom. If the estimator of the variance is taken as s^2 covering all the sample values, then v will be $(n-1)$ whereas if it is s^2 , derived from k group or sub sample means, v will be $(k-1)$.

In this case, the confidence interval for \bar{Y} with a confidence coefficient $1-\alpha$ is given by

$$\text{Prob}\{|\bar{y} - \bar{Y}| \leq L_\alpha s(\bar{y})\} = 1-\alpha, \quad (4.15)$$

where L_α is the $\alpha\%$ point of the t distribution with v degrees of freedom (cf Table 2.4 of Chapter 2 p. 46). The above probability statement can be obtained for any given sample size, which is large enough to validate the assumption of normality for the sampling distribution of \bar{y} , but the expected length of the confidence interval, $E(L)$, will depend on the sample size and the population standard deviation. For, the length of the confidence interval is given by

$$L = 2L_\alpha s(\bar{y}) \quad (4.16)$$

and in case of large samples

$$E(L) = 2k_a \sigma(\bar{y}). \quad \dots \quad (4.17)$$

The sample size can be determined so as to ensure a specified value for $E(L)$ or for $E(L)/\bar{Y}$. It may be noted that the permissible error P_d specified earlier is $\frac{1}{2}E(L)/\bar{Y}$. Hence, the rse of the estimator that would ensure a given value for the expected length of the confidence interval relative to the population parameter is given by

$$e = C(\bar{y}) = P_d/k_a, \quad \dots \quad (4.18)$$

where $P_d = \frac{1}{2}E(L)/\bar{Y} = k_a C(\bar{y})$, and this is the same as (4.14). To fix the sample size so as to ensure a given value e for $C(\bar{y})$, a knowledge of C is required. It may be noted that when the value of C is known only approximately this procedure ensures the specified confidence coefficient $1-\alpha$, though the expected length may be different from that specified, whereas in the procedure given in Sub-section 4.8b, even the specified confidence coefficient is only approximately achieved.

4.8d PROB. $\{L \leq 2d\}$ IS SPECIFIED

In the previous Section, a procedure of determining the rse e and hence n is given so as to ensure a specified value for $E(L)$. Here a procedure of determining the sample size proposed by Murthy (1963) is given which ensures

$$\left. \begin{aligned} \text{Prob. } \left\{ \left| \frac{\bar{y} - \bar{Y}}{\bar{Y}} \right| \leq P_d \right\} &= 1-\alpha \\ \text{Prob. } \{L \leq 2d\} &= 1-\beta. \end{aligned} \right\} \quad \dots \quad (4.19)$$

and

Since $L = 2k_a s(\bar{y})$, (cf. (4.16)), and $s(\bar{y}) = s/\sqrt{n}$ in srswr, the second statement in (4.19), may be written as

$$\text{Prob. } \left\{ k_a \frac{s}{\sqrt{n}} \leq d \right\} = 1-\beta,$$

that is,

$$\text{Prob. } \left\{ \frac{vs^2}{\sigma^2} \leq \frac{P_d^2 \bar{Y} n}{C^2 k_a^2} \right\} = 1-\beta, \quad \dots \quad (4.20)$$

since $d = P_d \bar{Y}$ and $C = \sigma/\bar{Y}$. Noting that vs^2/σ^2 follows a gamma distribution with the parameters $\alpha = \frac{1}{2}$ and $p = v/2$ and reducing (4.20) to an incomplete gamma function, which is already tabulated, we get

$$\text{Prob. } \{L \leq 2d\} = I(u, p), \quad \dots \quad (4.21)$$

where

$$I(u, p) = \frac{1}{\Gamma(p+1)} \int_0^{u\sqrt{p+1}} e^{-x} x^p dx, \quad p = \frac{v}{2} - 1 \text{ and } u = \frac{P_d^2 n \sqrt{v}}{C^2 k_a^2 \sqrt{2}}$$

The value of n required to ensure the two levels of probability given in (4.19) is given by

$$n = u \frac{C^2 k_a^2 \sqrt{2}}{P_d^2 \sqrt{v}}, \quad \dots \quad (4.22)$$

where u is obtained equating $I(u, p)$ to $1-\beta$. Thus we see that the sample size n can be calculated once the values of P_d , C and v are specified. It may be noted that if the sampling variance of \bar{y} is estimated on the basis of all the units, $v = n-1$ and that in that case only P_d and C need be specified for working out the sample size. It may be pointed out that at least a rough idea of C is necessary to determine the sample size in this procedure also. In case the value of C used in the determination of n is only approximate, then the second statement given in (4.19) will be satisfied only approximately, though the first statement will still be valid.

4.8e FIXED L σ UNKNOWN

Here another procedure of determining the sample size is given, which ensures a confidence interval of a desired length with a pre-specified value of confidence coefficient (Stein, 1945). The procedure consists in first selecting a sample of size n_1 determined on *a priori* considerations based on past experience and working out the sample size n required to get a confidence interval of a given length L_0 with a specified confidence coefficient $(1-\alpha)$ on the basis of

$$L_0 = 2k_a s_1 / \sqrt{n}$$

where s_1 is an estimate of σ based on the first sample of n_1 units and k_a is the $\alpha\%$ point of the t distribution with (n_1-1) degrees of freedom in the case of srswr. In the case srswor s_1/\sqrt{n} is to be multiplied by the factor $\sqrt{1-f}$. Thus the required sample size in the case of srswr is given by

$$n = 4 k_a^2 s_1^2 / L_0^2 \quad \dots \quad (4.23)$$

and in srswor we get

$$n = 4 \frac{k_a^2 s_1^2 N}{N L_0^2 + 4 k_a^2 s_1^2} \quad \dots \quad (4.24)$$

If $n \leq n_1$, the sample of size n_1 already taken is considered to be adequate for ensuring the length of the confidence interval to be less than L_0 . If $n > n_1$, an additional sample of $n_2 (= n - n_1)$ units is to be selected using the same procedure and the confidence interval is given by

$$\text{Prob.} \left\{ \bar{y} - k_\alpha \frac{s_1}{\sqrt{n}} < \bar{Y} < \bar{y} + k_\alpha \frac{s_1}{\sqrt{n}} \right\} = 1 - \alpha, \quad \dots \quad (4.25)$$

where \bar{y} is based on all the $n (= n_1 + n_2)$ units, whereas s_1 and k_α are based on the original sample of n_1 units. It can be easily seen that the length of the confidence interval is exactly L_0 because of (4.23). In srs wor, it is necessary to multiply s_1 by the factor $\sqrt{1-f}$ as mentioned earlier. If n_2 is fairly large compared to n_1 , it would be desirable to get a more efficient estimate of $\sigma(\bar{y})$ on the basis of the entire sample of n units than that based on a relatively smaller sample of n_1 units and to use k_α corresponding to $(n-1)$ degrees of freedom, as in that case the length of the confidence interval is likely to be smaller than L_0 . The effect of non-normality of the estimator on Stein's procedure has been studied by Bhattacharjee (1965).

4.8f COST ASPECT

In some situations the sample size may have to be determined on the basis of the resources available and in some other situations on the basis of the specified margin of error. Let the budget sanctioned for a survey be C' . Assuming the simplest type of cost function, namely,

$$C' = C_0 + nC_1, \quad \dots \quad (4.26)$$

where C_0 is the overhead cost and C_1 is the cost of surveying one unit, we get the sample size required, n , as

$$n = (C' - C_0)/C_1. \quad \dots \quad (4.27)$$

It may be noted that in this situation the rse of the estimator is not at our choice, but we can only find out the error, that is to be expected in the estimate with this sample size, if some prior information is available regarding the population coefficient of variation $C_y (= \sigma/\bar{Y})$.

Suppose the loss in terms of money is just proportional to the value of the rse of the estimator. Let l be the loss per 1% of rse of the estimator. Then the total of survey cost and loss involved in taking a decision based on a sample of size n selected with srswr is given by

$$L(n) = C_0 + nC_1 + \frac{C_y}{\sqrt{n}} l. \quad \dots \quad (4.28)$$

It may be seen that in the loss function $L(n)$ the cost of the survey increases with n while the loss due to the error in the estimate decreases with n . Hence if $L(n)$ is graphed against n , this graph will have a minimum value at some value of n . This value of n is termed the *optimum sample size*. This optimum value is obtained by differentiating $L(n)$ with respect to n and equating this partial derivative to zero, that is

$$C_1 - \frac{C_y}{2n^{3/2}} l = 0$$

Hence

$$n = \left(\frac{l}{2} \frac{C_y}{C_1} \right)^{2/3} \quad (4.29)$$

The optimum sample size in srs w/o r can also be similarly determined by considering the loss function $L(n)$ given by

$$L(n) = C_0 + nC_1 + \sqrt{\frac{N-n}{N-1}} \cdot \frac{C_y}{\sqrt{n}} l \quad (4.30)$$

It may be noted that in arriving at the optimum value of n it is necessary to have some fairly reliable information regarding the values of C_0 , C_1 , C_y and l based on some previous surveys or on exploratory studies taken up specifically for this purpose.

4.8g ESTIMATION OF A PROPORTION

Since the rse of the sample proportion p is a function of only the population proportion P , the determination of n , required to achieve a given precision for p , will depend only on P . The population cv's for different values of P are given in Table 4.11. Knowing the value of the cv for a given P , it is possible to find the sample sizes required to estimate P with specified rse's by referring to Table 4.9 or 4.10 as the case may be or by directly using the formula

$$n' = \frac{Q}{P} \frac{1}{e^2}, \quad Q = 1 - P, \quad (4.31)$$

where e is the desired rse of the estimate of P in srswr and

$$n'' = \frac{Nn'}{N+n'-1} \dots \quad (4.32)$$

in srs wr. The techniques of determination of sample size discussed in the previous sections may be applied to the case of estimating proportions by noting that in large samples p is normally distributed with mean P and standard deviation $\sigma(p)$.

TABLE 4.11. VALUES OF COEFFICIENT OF VARIATION FOR DIFFERENT VALUES OF PROPORTION.

P	$CV(\%)$	P	$CV(\%)$	P	$CV(\%)$
(1)	(2)	(1)	(2)	(1)	(2)
0.01	995	0.35	136	0.70	65
0.05	436	0.40	127	0.75	58
0.10	300	0.45	111	0.80	50
0.15	238	0.50	100	0.85	41
0.20	200	0.55	90	0.90	33
0.25	173	0.60	82	0.95	23
0.30	153	0.65	73	0.99	10

REFERENCES

- BHATTACHARJEE, G. P. (1965) : Effect of non-normality on Stein's two-sample test; *Ann. Math. Stat.*, 36, 651-663.
- COCHRAN, W. G. and WATSON, D. J. (1936) : An experiment in observer's bias in the selection of shoot heights; *Empire Journal of Experimental Agriculture*, 4, 69-76.
- FISHER, R. A. and YATES, F. (1938) : *Statistical Tables for Biological, Agricultural and Medical Research*; Table XXXIII, Oliver and Boyd, London.
- HASEL, A. A. (1942) : Estimation of volume of timber stands by strip sampling; *Ann. Math. Stat.*, 13, 179-206.
- KENDALL, M. G. and SMITH, B. B. (1939) : *Tables of Random Sampling Numbers*; Tracts for Computers, No. XXIC, Cambridge University Press, reprinted in 1946, 1951, 1954.
- MATTHAI, A. (1954) : On selecting random numbers for large-scale sampling; *Sankhyā*, 13, 157-160.

- MCHUGH, R. B (1961) Confidence interval inference and sample size determination
American Statistician, 15, (2), 14-17
- MURTHY, M. N (1963) A note on determination of sample size, *Sankhya*, 25,(A)
 381-382
- RAND CORPORATION (1955) *A Million Random Digits and 100000 Normal Deviates*
 The Free Press, Illinois
- RAO, C. R., MITRA, S. K. and MATHAI, A (1966) *Formulae and Tables for Statistical Work*
 Statistical Publishing Society, Calcutta
- STEIN, C (1945) A two sample test for a linear hypothesis whose power is independent of the variance, *Ann Math Stat*, 16, 243-258
- TIPPETT, L. H. C (1927) *Random Sampling Numbers*, Tracts for Computers No XV, Cambridge University Press, reprinted in 1952
- YATES F (1935) Some examples of biased sampling *Annals of Eugenics*, 6, 202-213
- ZAREKOVICH, S. S (1961) *Sampling Methods and Censuses*, Part I, Food and Agricultural Organization of the United Nations, Rome

COMPLEMENTS AND PROBLEMS

4.1 Suppose it is required to estimate the total output of a group of N factories in a region such that the sample estimate lies within 10% of the true value with a confidence coefficient of 95%

- (i) Calculate the sample sizes required in the case of srs wor when N is (a) 500
 (b) 1000 (c) 2500, (d) 5000 and (e) 10000

(ii) Find the changes in the sample sizes obtained in (i) if srs wr is used instead of srs wor

The population coefficient of variation is known to be 60% and the estimator can be assumed to be normally distributed

4.2 In a sample of 200 colleges selected from a population of 2000 colleges with srs wor, 140 colleges were in favour of a proposal to have a new form of examination

- (i) Estimate the 95% confidence limits for the number of colleges in the population in favour of the proposal

(ii) Do the above data furnish sufficient evidence at 95% confidence level to reject the hypothesis that only 50% of all the colleges are in favour of this proposal?

(iii) What sample size would reduce the expected length of the 95% confidence interval to half its length in (i)?

4.3 If the sample size required to estimate the proportion P of workers in a population with an rse of $\alpha\%$ is n in srs wor, determine the sample size n' , required to estimate the proportion of non workers with the same precision

4.4 A survey is to be conducted to study the prevalence of common diseases in a large population. For any disease that affects at least 1% of the persons in the population, it is desired to estimate the total number of cases with an rse not exceeding 20%.

(i) Determine the sample size needed in the case of srswr.

(ii) What will be the sample size requirement if estimates of the total number of cases is wanted separately for males and females with the same precision ? State the assumptions involved, if any.

4.5 A survey was conducted in a village consisting of 625 households by covering a sample of 50 households drawn with srs wr to estimate the average monthly household expenditure on toilet goods. The estimate turned out to be Rs. 4.20 with a standard error of Rs. 0.47. Using this information determine the sample size needed to estimate the same characteristic in a neighbouring village on the basis of a sample selected with srswr such that the length of the confidence interval at 95% confidence level is 20% of the true value. State the assumption involved in finding the sample size.

4.6 Suppose the loss and cost functions are of the form

$$L(e) = \lambda e^2 \quad \text{and} \quad C = C_0 + nC_1,$$

where λ , C_0 and C_1 are constants and e is the rse of the estimator. For srs wr find out the sample size n that minimizes the sum of the survey cost and the loss due to sampling error. Assume the knowledge of any parameters that need be known for this purpose.

4.7 Discuss critically the following statement :

"In a series of random sampling numbers, there may occur *patches* (runs of numbers or sequences of numbers), which are not suitable for use by themselves. But such patches should be present in a sufficiently long series, and may be used in conjunction with preceding and/or succeeding portions of the series for drawing a fairly large sample."

4.8 A pair of random numbers x and y is selected from a table of two-digited random numbers.

(i) Find the expected value and the variance of x , and also of y .

(ii) What is the expected value of the square of the difference between the two random numbers ?

4.9 State how you would select one out of two units with equal probability using a biased coin. Also extend this procedure for selecting one unit from a population of 8 units with srs.

4.10 The results of 100 throws of an unbiased coin are given below, where H and T stand for *head* and *tail* respectively.

T T H H T	T T T H T	T T H T H	T T H H H
H T T H H	H T T H H	H H T H T	H T T H T
T T H T T	T H T T T	T H H T T	T H H H T
H H T H T	H H H H H	T H T T T	T H T T H
H T T H T	T H H T H	H T T H T	H T T T T

Using these results draw a sample of 4 villages with srs wor from a list of 23 villages. Explain how the procedure adopted by you achieves srs wor

4.11 How would you select a point at random within a circle of radius 4 cm? State the limitations of the method, if any

4.12 In a village there are 352 fields. For estimating the yield of paddy, three random numbers 12, 273 and 105 are chosen and the fields with these survey numbers are to be surveyed. If a selected field does not grow paddy, the surveyor is asked to substitute it by the field with the next higher survey number, which grows paddy, but not already included in the original sample.

(i) State, giving reasons, whether this procedure leads to srs

(ii) Suggest a procedure for selecting a sample of 3 fields growing paddy with srs wor when a list of paddy growing fields is not available

(iii) What information will you require to estimate unbiasedly the total yield of paddy based on a sample selected according to your procedure in (ii)?

4.13 For selecting a sample of n words with srs wor from a dictionary, the following procedure is suggested. First select a page with equal probability from all the pages in the dictionary using a table of random numbers. In the selected page choose one of the two columns at random. Draw a number R at random from 1 to M , the maximum number of words in any column in any page or a number greater than the maximum. If R is less than or equal to the number of words in the selected column accept the R th word in the sample, otherwise repeat the operation starting from selection of the page till one word is selected. Then repeat the entire procedure till n different words are selected. Show that this procedure leads to srs wor.

4.14 (i) Derive the sample size required for estimating the number of units, D , common to two long lists of M and N units with an rse of $\alpha\%$ using samples of m and n ($= km$) units drawn from the two lists with srs wor, k being a given constant. Assume the binomial distribution for d in the limiting case of M and N becoming infinitely large and D , m/M and n/N remaining fixed (cf Problem 3.21 of Chapter 3).

(ii) When samples of 1000 and 500 persons, selected with srs wor from the lists of names of medical practitioners kept by two advertising agencies, were compared, it was found that there were 9 common names. Find the sample size needed to estimate the total number of names common to both the lists with an rse of 10%, assuming the same ratio of the sample sizes (2 : 1) used in the study.

ANNEXURE 4.1

TABLE 4.12. VILLAGE-WISE COMPLETE ENUMERATION DATA OBTAINED IN 1951 AND 1961 CENSUSES FOR A TEHSIL.

sr. no.	1951 Census			1961 Census			
	area in sq.miles	cultivated area (in acres)	number of persons	number of persons	no. of cultiv- ators	workers at household industry	no. of house- holds
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1.	6.97	2544	3295	3552	806	153	757
2.	1.55	428	378	420	193	0	78
3.	3.51	1177	2574	2819	819	116	548
4.	9.96	4567	4466	4892	1970	327	1064
5.	8.46	2618	3915	4040	1149	196	873
6.	9.71	4113	3249	3644	1510	151	763
7.	6.16	1869	3462	2303	611	97	541
8.	12.06	2713	4918	5082	888	39	984
9.	9.42	2237	2461	2554	1291	74	523
10.	1.21	600	511	506	257	0	97
11.	15.11	3420	6851	7409	1855	141	1455
12.	12.86	4012	4782	4873	1741	131	993
13.	7.73	1949	3753	3802	1345	63	758
14.	2.13	695	1299	977	343	22	198
15.	6.03	1569	1816	2252	586	5	445
16.	12.74	4562	4942	5542	1543	80	1066
17.	6.90	2221	2383	2538	719	179	541
18.	7.80	2423	2836	2988	797	52	631
19.	1.63	608	832	886	291	10	200
20.	3.03	1124	865	820	351	2	174
21.	1.29	527	588	605	126	38	111
22.	9.09	2767	6365	6325	1906	152	1197
23.	7.40	2770	3464	3665	1454	29	718
24.	2.56	719	941	1045	191	42	215
25.	2.95	607	1287	1513	336	99	276
26.	2.46	482	1058	1124	419	10	240
27.	4.03	1527	2111	2167	432	72	415
28.	3.74	1367	1337	1506	393	33	282
29.	1.95	767	827	772	192	5	169
30.	6.44	1648	2535	2772	1091	104	607
31.	11.33	2440	5820	6181	1830	393	1334
32.	9.28	2434	3378	3612	1219	126	716
33.	4.90	1638	1877	2342	543	35	472
34.	0.18	61	3402	3684	93	46	812
35.	13.25	4505	5769	5700	1675	133	1189

TABLE 4.12 (Contd.) VILLAGE WISE COMPLETE ENUMERATION DATA OBTAINED IN 1951 AND 1961 CENSUSES FOR A TEHSIL

sr no	1951 Census			1961 Census			
	area in sq miles	cultivated area (in acres)	number of persons	number of persons	no of cultivators	workers at household industry	no of households
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
36	5.37	1751	3148	3822	884	121	717
37	5.28	2622	2654	2768	1039	107	599
38	5.86	2848	4201	4585	941	51	958
39	6.91	3013	3523	3879	1348	84	878
40	4.28	1599	1714	1466	438	9	289
41	8.33	2949	3479	3850	1272	71	785
42	13.23	2641	7420	8086	2261	101	1668
43	4.38	1959	2681	2523	455	7	549
44	4.28	1371	2870	3139	916	5	657
45	7.70	3290	4435	4781	1380	28	1004
46	5.85	2526	3265	3592	904	75	756
47	4.97	2935	4096	4054	1438	35	793
48	2.71	1109	984	1185	445	22	237
49	14.30	2821	8200	7924	2299	269	1668
50	13.96	3678	8368	8493	2259	126	1667
51	6.40	811	3180	3312	446	65	704
52	9.13	1453	2683	2996	888	39	657
53	3.86	1665	1894	1891	277	37	332
54	8.07	2350	4021	4201	793	54	850
55	1.35	564	1241	1303	421	0	279
56	9.06	2487	3243	2915	1085	30	590
57	4.60	904	1306	1540	383	62	323
58	10.67	2040	6017	6572	1813	126	1414
59	3.57	1314	1643	1808	394	36	381
60	6.52	1506	7357	6758	1116	71	1512
61	6.17	1657	2988	3042	1145	110	629
62	5.01	1053	4020	3652	955	51	715
63	5.57	2071	6142	6394	1159	109	1400
64	2.53	872	1880	2105	826	34	409
65	5.64	1718	5133	5803	1565	77	1200
66	6.93	316	656	532	75	3	116
67	1.78	653	1764	1828	271	0	398
68	3.80	2357	3809	3968	821	69	858
69	6.35	3258	2780	2727	841	12	573
70	8.97	4051	5131	5059	1709	271	1036

TABLE 4.12. (Contd.) VILLAGE-WISE COMPLETE ENUMERATION DATA
OBTAINED IN 1951 AND 1961 CENSUSES FOR A TEHSIL.

sr. no.	1951 Census			1961 Census			
	area in sq.miles	cultivated area (in acres)	number of persons	number of persons	no. of cultivators	workers at household industry	no. of house- holds
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
71.	3.47	1209	3970	3984	1323	29	841
72.	2.55	1658	2039	2119	644	25	470
73.	6.27	2608	4281	4295	818	50	905
74.	3.61	1289	3664	3767	607	57	906
75.	4.59	599	3534	3363	925	44	749
76.	7.21	2573	4159	4444	1813	19	968
77.	4.39	1414	2444	2373	373	62	528
78.	3.31	980	1010	1021	170	5	196
79.	4.65	1543	2593	2559	633	32	546
80.	10.15	3060	4654	4961	1304	39	1033
81.	9.27	2600	3834	4699	1047	84	965
82.	2.32	1210	2188	2160	937	26	434
83.	6.07	2937	4049	4298	1137	52	914
84.	3.08	1867	2887	3114	559	35	689
85.	2.77	1337	1153	949	249	15	215
86.	4.66	1031	3732	3683	910	60	767
87.	4.58	1930	3796	3607	563	98	759
88.	2.24	1333	4283	4339	506	15	960
89.	2.58	1509	1470	1505	168	18	314
90.	0.94	509	1146	1154	217	24	253
91.	8.47	4424	5150	5709	1409	53	1207
92.	5.56	1881	953	1223	1	0	261
93.	10.87	4139	2366	2929	109	0	750
94.	7.35	4072	1292	3797	0	1	1229
95.	1.20	612	7153	8584	1167	129	1826
96.	16.36	5507	4865	6083	690	127	1401
97.	11.29	4634	9436	10472	2659	78	2073
98.	3.05	1667	381	550	67	0	107
99.	3.43	2013	1665	1801	591	13	379
100.	0.80	156	9113	10471	1506	846	2302
101.	3.67	1425	2540	3394	953	10	649
102.	5.17	2566	4816	5061	489	83	1091
103.	4.60	2394	2728	3458	451	29	689
104.	3.00	1356	4091	3596	399	2	536
105.	0.82	610	2069	4333	1074	125	1134

TABLE 412 (Contd.) VILLAGE WISE COMPLETE ENUMERATION DATA OBTAINED IN 1951 AND 1961 CENSUSES FOR A TEHSIL

sr no	1951 Census			1961 Census			
	area in sq.miles	cultivated area (in acres)	number of persons	number of persons	no of cultivators	workers at household industry	no of households
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
106	0 96	603	4340	4563	530	27	921
107	1 23	631	1563	1491	235	66	356
108	1 83	1074	1369	1499	269	31	314
109	3 03	1959	2323	2739	932	77	593
110	4 57	2366	1963	2195	367	67	467
111	4 32	2618	5596	6169	1367	77	1247
112	1 77	428	695	925	181	1	193
113	7 18	2075	8822	8916	230	65	1696
114	5 56	2296	4558	4650	1091	60	969
115	4 66	1870	3466	3461	553	104	682
116	3 56	1328	2639	2740	917	6	549
117	3 61	1612	2955	3151	1181	43	709
118	3 25	1653	1430	1483	473	0	310
119	1 91	933	1155	1284	440	22	265
120	8 15	2698	2895	2925	1069	112	629
121	1 44	730	1781	1953	561	6	397
122	5 72	2128	2976	3145	713	28	663
123	2 79	1753	4109	4029	98	67	812
124	2 75	772	1577	1707	402	100	373
125	4 03	2096	1328	2446	0	6	652
126	8 51	2862	4231	4194	1020	73	859
127	6 56	2377	4543	5158	783	118	1068
128	4 77	1318	1058	1116	222	8	236
total	715 82	248752	415149	443319	109248	8968	93129

1 square mile = 640 acres, 1 acre = 0.4047 hectare

Source District Census Handbook (1951 Census) and Primary Census Abstract (1961 Census) for Madurai District, Madras State

ANNEXURE 4.2

TABLE 4.13. STRIP-WISE COMPLETE ENUMERATION DATA ON LENGTH
(x) AND TIMBER VOLUME (y) FOR 10 BLOCKS OF THE BLACKS
MOUNTAIN EXPERIMENTAL FOREST.

sr. no.	block no.	strip no.	x	y	sr. no.	block no.	strip no.	x	y
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
1.		1	12	762	46.		14	6	192
2.		2	12	651	47.		15	6	224
3.		3	12	461	48.		16	6	165
4.		4	12	521					
5.		5	12	653	49.	4	1	7	174
6.		6	12	544	50.		2	7	159
7.		7	12	542	51.		3	7	157
8.		8	12	590	52.		4	7	140
9.		9	11	533	53.		5	7	198
10.		10	11	517	54.		6	7	169
11.		11	11	520	55.		7	7	127
12.		12	11	539	56.		8	7	181
13.		13	10	609	57.		9	7	156
14.		14	10	449	58.		10	7	207
15.		15	10	492	59.		11	7	181
16.		16	10	498	60.		12	7	121
					61.		13	7	125
17.	2	1	9	471	62.		14	7	86
18.		2	9	426	63.		15	3	115
19.		3	9	448	64.		16	3	129
20.		4	9	402					
21.		5	9	372	65.	5	1	5	121
22.		6	9	372	66.		2	5	170
23.		7	9	411	67.		3	6	316
24.		8	9	323	68.		4	7	207
25.		9	9	381	69.		5	7	319
26.		10	9	430	70.		6	7	266
27.		11	9	434	71.		7	9	445
28.		12	9	394	72.		8	9	406
29.		13	9	543	73.		9	9	427
30.		14	9	607	74.		10	9	449
31.		15	8	416	75.		11	9	382
32.		16	8	326	76.		12	9	243
					77.		13	9	360
33.	3	1	4	64	78.		14	9	344
34.		2	4	74					
35.		3	4	92	79.	6	1	9	331
36.		4	4	61	80.		2	9	378
37.		5	4	36	81.		3	9	295
38.		6	4	83	82.		4	9	238
39.		7	4	109	83.		5	9	305
40.		8	4	110	84.		6	9	284
41.		9	5	115	85.		7	9	363
42.		10	5	102	86.		8	9	352
43.		11	5	94	87.		9	9	345
44.		12	5	104	88.		10	9	354
45.		13	6	161	89.		11	9	401

TABLE 4 13 (Contd) STRIP WISE COMPLETE ENUMERATION DATA ON LENGTH (x) AND TIMBER VOLUME (y) FOR 10 BLOCKS OF THE BLACKS MOUNTAIN EXPERIMENTAL FOREST

sr no	block no	strip no	x	y	sr no	block no	strip no	x	y
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
90		12	9	381	133		10	15	519
91		13	8	408	134		11	15	571
92		14	8	376	135		12	9	403
93		15	8	374	136		13	9	354
94.		16	8	337	137		14	9	363
					138		15	9	562
95	7	1	7	248	139		16	6	224
96		2	7	286	140		17	6	246
97		3	7	340	141		18	5	256
98		4	7	360	142		19	5	245
99		5	8	336	143		20	5	245
100		6	8	333	144		21	4	248
101		7	8	330	145		22	4	250
102		8	8	404	146		23	4	200
103		9	8	379	147		24	2	98
104		10	6	280	148		25	2	152
105		11	4	206	149		26	1	41
106.		12	2	130	150		27	1	46
107.		13	1	45					
					151	10	1	1	21
108	8	1	3	144	152		2	1	23
109		2	3	159	153		3	2	40
110		3	4	209	154		4	3	43
111		4	4	210	155		5	4	50
112		5	4	227	156		6	4	96
113		6	5	247	157		7	5	145
114		7	5	277	158		8	5	156
115		8	5	372	159		9	5	127
116		9	6	304	160		10	5	161
117		10	6	208	161		11	5	214
118		11	7	305	162		12	5	143
119		12	7	353	163		13	5	201
120		13	8	369	164		14	5	198
121		14	8	361	165		15	5	173
122		15	8	333	166		16	5	220
123		16	8	290	167		17	5	297
					168		18	5	151
124	9	1	1	26	169		19	5	158
125		2	1	48	170		20	5	163
126		3	4	138	171		21	5	141
127		4	8	330	172		22	5	212
128		5	10	355	173		23	5	143
129		6	12	401	174		24	5	224
130		7	14	500	175		25	6	174
131		8	15	554	176		26	6	187
132		9	15	534					

x in 10 chain units y in 1000' board measurement units

Source Hasel A A (1942) Estimation of volume of timber stands by strip sampling *Annals of Mathematical Statistics* 13, 179-206

Systematic Sampling

5.1 SAMPLING PROCEDURE

In this chapter, we consider a sampling technique, known as *systematic sampling*, which is operationally more convenient than simple random sampling, and which at the same time ensures for each unit equal probability of inclusion in the sample. The technique of systematic sampling consists in selecting every k -th unit starting with the unit corresponding to a number r chosen at random from 1 to k , where k is taken as the integer nearest to N/n , the reciprocal of sampling fraction. The random number r chosen from 1 to k is known as the *random start* and the constant k is termed the *sampling interval*. A sample selected by this procedure is termed a *systematic sample with a random start*. It may be seen that the value of r determines the whole sample. In other words, this procedure amounts to selecting with equal probability one of the k possible groups of units (samples) into which the population can be divided in a systematic manner.

The main difference between the procedures of systematic sampling and simple random sampling is that enumeration of all possible samples and selection of one of them with equal probability are much simpler in the former than in the latter procedure. Apart from its operational convenience, which is of considerable importance in large-scale sampling work, this sampling procedure provides estimators more efficient than those provided by srs under certain conditions usually met with in practice.

Systematic sampling has been considered in detail by W G Madow and L H Madow (1941) L H Madow (1946) Cochran (1946) and Lahiri (1954). Reviews of the work done in this field have been given by Yates (1948) and Buckland (1951). The application of systematic sampling to forest surveys has been illustrated by Hasel (1942) Finney (1948) and Nur and Bhargava (1951). Use of systematic sampling in estimating catch of fish has been demonstrated by Sukhatme, Punse and Sistry (1958).

5 1a SAMPLING OF TWO UNITS

To fix up the ideas let us consider the estimation of the mean of a population of four units on the basis of a systematic sample of two units.

unit	U_1	U_2	U_3	U_4
value	Y_1	Y_2	Y_3	Y_4

In this case $N = 4$ and $n = 2$ and hence the sampling interval k is $2 (= 4/2)$. The two possible samples with their probabilities and sample means are given in Table 5 1.

TABLE 5 1 SYSTEMATIC SAMPLES OF 2 UNITS FROM 4 UNITS

random start	sample composition	probability	sample mean
(1)	(2)	(3)	(4)
1	$U_1 U_3$	1/2	$(Y_1 + Y_3)/2$
2	$U_2 U_4$	1/2	$(Y_2 + Y_4)/2$

The expected value of the sample mean is given by the average of column (4), and this turns out to be \bar{Y} showing that \bar{y} is unbiased for \bar{Y} . Since $E(y) = \bar{Y}$, an unbiased estimator of the population total is given by 4 times the sample mean. It may be noted that the probability of inclusion of a particular unit in the sample is 1/2 and it is the same for all the units in the population. It is of interest

to note that this procedure gives rise to only 2 of the 6 samples possible in the case of srs w.r. In fact the samples (U_1U_2) , (U_1U_4) , (U_2U_3) and (U_3U_4) have no chance of being selected at all.

Suppose we had 5 units in the population instead of 4 units. Now the inverse of the sampling fraction, $5/2$, is not an integer and in this case the interval may be taken as either 2 or 3. Taking it as 2 we see that the possible samples are $U_1U_3U_5$ and U_2U_4 and with 3 as the interval we get the samples U_1U_4 , U_2U_5 and U_3 . In these two cases, it may be seen that the sample size is not same for all samples and that the sample mean is not unbiased for \bar{Y} , for

$$\frac{1}{2} \left(\frac{Y_1 + Y_3 + Y_5}{3} + \frac{Y_2 + Y_4}{2} \right) \text{ and } \frac{1}{3} \left(\frac{Y_1 + Y_4}{2} + \frac{Y_2 + Y_5}{2} + Y_3 \right)$$

are not equal to \bar{Y} .

It is possible to get an unbiased estimator of \bar{Y} in this case by dividing the product of the sample total and k by N , that is, $\hat{Y} = \frac{k}{N} \sum_{i=1}^{n'} y_i$, n' being the number of units in the sample. The possible samples with the values of \hat{Y} , the modified estimator of \bar{Y} , are given in Table 5.2. From this table, it is clear that the expected value of the estimator \hat{Y} , which is the mean value of column (5), turns out to be \bar{Y} in both the cases considered here, showing that the estimator is unbiased for \bar{Y} .

TABLE 5.2. SYSTEMATIC SAMPLES OF SIZE 2 FROM 5 UNITS.

sampling interval	random start	sample composition	probabi- lity	estimator \hat{Y}
(1)	(2)	(3)	(4)	(5)
2	1	$U_1U_3U_5$	1/2	$2(Y_1 + Y_3 + Y_5)/5$
	2	U_2U_4	1/2	$2(Y_2 + Y_4)/5$
3	1	U_1U_4	1/3	$3(Y_1 + Y_4)/5$
	2	U_2U_5	1/3	$3(Y_2 + Y_5)/5$
	3	U_3	1/3	$3(Y_3)/5$

5.1b A MODIFIED SAMPLING SCHEME

It is possible to make the sample mean itself an unbiased estimator of \bar{Y} by slightly changing the procedure of systematic sampling considered above. One modification of the procedure consists in selecting the random start from 1 to N instead of from 1 to k and taking every k th unit in both forward and backward directions as constituting the sample (Cochran 1963). In the present case if k is taken as 2 this procedure gives rise to the samples $U_1 U_3 U_5$ and $U_2 U_4$ with probabilities $3/5$ and $2/5$ since the former sample gets selected whenever any one of the numbers 1, 3 or 5 is taken as the random start and the latter whenever the random start is 2 or 4. Hence the expected value of \bar{y} is

$$E(\bar{y}) = \frac{3}{5} \left(\frac{Y_1 + Y_3 + Y_5}{3} \right) + \frac{2}{5} \left(\frac{Y_2 + Y_4}{2} \right) = \bar{Y}$$

It may be noted that in this modified procedure also the sample size is not fixed, and that this procedure does not ensure equal probability of inclusion in the sample for every unit. Another modification of the systematic sampling procedure which makes y unbiased for \bar{Y} ensuring at the same time a given sample size and equal probability of selection in the sample for every unit in the population is given in Section 5.2.

5.1c SAMPLING OF n UNITS

The units in a systematic sample of n units drawn from a population of N units are given by

$$\{U_{r+jk}\}, \quad j = 0, 1, 2, \dots, r+jk \leq N \quad (5.1)$$

Case (i) $N = nk$

If N is a multiple of n , that is, if $N = nk$, then the number of units in each of the k possible systematic samples is n . In this case systematic sampling amounts to grouping the N units into k samples of exactly n units each in a systematic manner and selecting one of them with probability $1/k$. The k possible samples together with their means are given in Table 5.3. From this table, it is clear that each of the N units occurs once in only one of the k samples, thus ensuring equal probability of inclusion in the sample for every unit in the population.

TABLE 5.3. SYSTEMATIC SAMPLES OF n UNITS FROM $N (= nk)$ UNITS.

random start	sample composition	probabi- lity	sample mean \bar{y}
(1)	(2)	(3)	(4)
1	$U_1 \ U_{1+k} \dots \ U_{1+jk} \dots \ U_{1+(n-1)k}$	$1/k$	$\frac{1}{n} \sum_{j=0}^{n-1} Y_{1+jk} = \bar{y}_1$
2	$U_2 \ U_{2+k} \dots \ U_{2+jk} \dots \ U_{2+(n-1)k}$	$1/k$	$\frac{1}{n} \sum_{j=0}^{n-1} Y_{2+jk} = \bar{y}_2$
\vdots	\vdots	\vdots	\vdots
r	$U_r \ U_{r+k} \dots \ U_{r+jk} \dots \ U_{r+(n-1)k}$	$1/k$	$\frac{1}{n} \sum_{j=0}^{n-1} Y_{r+jk} = \bar{y}_r$
\vdots	\vdots	\vdots	\vdots
k	$U_k \ U_{2k} \dots \ U_{(d+1)k} \dots \ U_{nk}$	$1/k$	$\frac{1}{n} \sum_{j=0}^{n-1} Y_{(d+1)k} = \bar{y}_k$

Considering all the k possible samples, we find

$$\begin{aligned}
 E(\bar{y}) &= \frac{1}{k} \sum_{r=1}^k \bar{y}_r = \frac{1}{k} \sum_{r=1}^k \left(\frac{1}{n} \sum_{j=0}^{n-1} Y_{r+jk} \right) \\
 &= \frac{1}{nk} \sum_{r=1}^k \sum_{j=0}^{n-1} Y_{r+jk}, \\
 &= \frac{1}{N} (Y_1 + Y_2 + \dots + Y_N) = \bar{Y},
 \end{aligned}$$

showing that when $N = nk$, \bar{y} is unbiased for \bar{Y} . It may be noted that as in the case of srs wr, there is no possibility of repetition of any unit in a systematic sample also.

Case (ii) : $N \neq nk$.

Suppose N is not a multiple of n , then the number of units selected systematically with the sampling interval k equal to the integer nearest to N/n need not necessarily be equal to n , the required sample size, unlike in the previous case where N was taken as a multiple of n . If q and r' are respectively the quotient and the remainder obtained

on dividing N by n , then N can be written as $N = nq + r'$ and the sampling interval can be taken as q or $q+1$ according as $r' \leq n/2$ or $r' > n/2$. If k is q' ($= q$ or $q+1$), then the number of units n' that can be expected in the sample would be given by $[N/q']$ or $[N/q'] + 1$ with probabilities $[N/q'] + 1 - (N/q')$ and $(N/q') - [N/q']$ respectively. If $q' = q$, we get $n' = n + [r'/q]$ and $n + [r'/q] + 1$, with probabilities $[r'/q] + 1 - (r'/q)$ and $(r'/q) - [r'/q]$. Similarly, if $q' = q+1$, we get $n' = n - [(n-r')/(q+1)]$ and $n - [(n-r')/(q+1)] + 1$ with probabilities $[(n-r')/(q+1)] + 1 - (n-r')/(q+1)$ and $(n-r')/(q+1) - [(n-r')/(q+1)]$.

Examples

For instance, if $N = 129$ and $n = 20$, then $q = 6$ and $r' = 9$, and k can be taken as 6. In this case, the sample size for each of the samples with random starts 1, 2 and 3 would be 22 and for each of the other three samples, the sample size would be 21. Similarly, if $N = 131$ and $n = 20$, k may be taken as 7 and in this case the sample size for samples with random starts 6 and 7 would be 18 and for each of the other five samples it would be 19. The difference between the required sample size and that achieved can be substantial as is evidenced by sampling 80 units from a population of 199 units, since in this case the sample size actually achieved with the interval 2 can either be 99 or 100.

An unbiased estimator of \bar{Y} is provided by

$$\hat{Y} = \frac{l}{N} \sum_{i=1}^n y_i. \quad \dots \quad (5.2)$$

For

$$E(\hat{Y}) = \frac{1}{k} \sum_{r=1}^k \left(\frac{l}{N} \sum_{i=1}^n y_i \right)_r,$$

where r stands for the systematic sample with random start r ($= 1, 2, \dots, k$). It can be easily verified that each unit in the population occurs once in only one of the k possible samples and hence

$$\sum_{r=1}^k \left(\frac{n'}{N} y_r \right) = \sum_{i=1}^N Y_i \quad \text{and} \quad E(\hat{Y}) = \bar{Y}.$$

From this it follows that an unbiased estimator of \bar{Y} is given by $\hat{Y} = k \sum_{i=1}^{n'} y_i$. If $N = nk$, the estimator of \bar{Y} given in (5.2) reduces to the sample mean,

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \dots \quad (5.3)$$

since in this case $k/N = 1/n$ and $n' = n$ for all possible samples.

The difference between \hat{Y} given in (5.2) and \bar{y} , namely,

$$\frac{k}{N} \sum_{i=1}^{n'} y_i - \frac{1}{n} \sum_{i=1}^n y_i = \left(\frac{kn'}{N} - 1 \right) \bar{y},$$

is likely to be negligible if n is not very small and N is fairly large compared to n . Hence, the bias involved in using the sample mean as an estimator of \bar{Y} will be negligible in case of samples selected from a large population.

5.2 CIRCULAR SYSTEMATIC SAMPLING

The disadvantages of systematic sampling considered in Section 5.1 regarding the actual sample size being different from that required and the sample mean being a biased estimator of the population mean when N is not a multiple of n can be overcome by adopting a device, known as *circular systematic sampling* (css). This device consists in choosing a random start from 1 to N and selecting the unit corresponding to this random start and thereafter every k -th unit in a cyclical manner till a sample of n units is obtained, k being the integer nearest to N/n . That is, if r is a number selected at random from 1 to N , the sample consists of the units corresponding to the numbers

$$\text{and } \left. \begin{cases} \{r+jk\}, & \text{if } r+jk \leq N \\ \{r+jk-N\}, & \text{if } r+jk > N \end{cases} \right\}, j = 0, 1, 2, \dots, (n-1). \dots \quad (5.4)$$

This technique ensures equal probability of inclusion in the sample for every unit unlike the modified sampling scheme considered in Sub-section 5.1b. This procedure was suggested by D. B. Lahiri in

1952 (NSS Instructions to Field Workers) and since then this is being used in the Indian National Sample Survey for sampling villages, households, plots, etc. for different surveys. In contrast to this sampling technique, the usual procedure of selecting a random start r from 1 to k and including in the sample the units corresponding to $\{r+jk\} \leq N$ for $j = 0, 1, 2, \dots$ discussed in Section 5.1 may be termed *linear systematic sampling* (LSS). The procedures of LSS and CSS are illustrated in Figure 5.1.

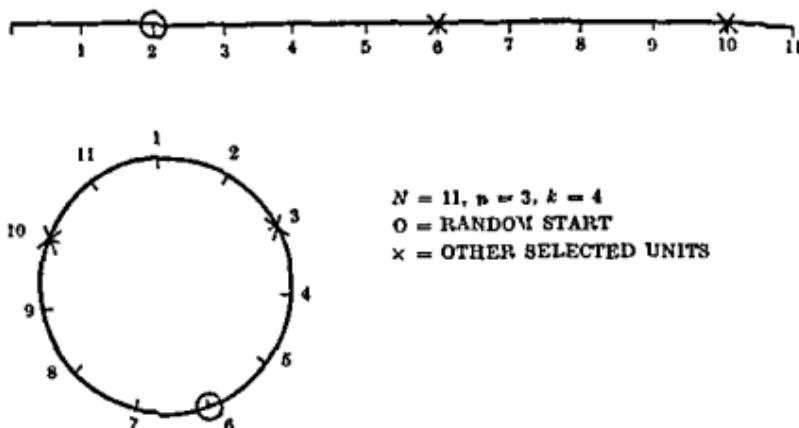


Figure 5.1. Diagrammatic representation of linear and circular systematic sampling schemes

5.2a SAMPLING OF TWO UNITS

The technique of CSS can be illustrated by applying it to the case of sampling 2 units from 5 units. Taking $k = 2$, the 5 possible samples together with the sample means are shown in Table 5.4

TABLE 5.4 CIRCULAR SYSTEMATIC SAMPLES OF 2 UNITS FROM 5 UNITS.

random start	sample composition	probability	sample mean
(1)	(2)	(3)	(4)
1	$U_1 U_3$	1/5	$(Y_1 + Y_3)/2$
2	$U_2 U_4$	1/5	$(Y_2 + Y_4)/2$
3	$U_3 U_5$	1/5	$(Y_3 + Y_5)/2$
4	$U_4 U_1$	1/5	$(Y_4 + Y_1)/2$
5	$U_5 U_2$	1/5	$(Y_5 + Y_2)/2$

It is to be noted that in case of random starts 4 and 5, the other unit in the sample is obtained by proceeding in a circular manner. That is, the sample with $r = 4$ consists of units U_4 and U_1 , since U_6 is taken as U_1 . Similarly for $r = 5$, we get the sample (U_5, U_2) , since U_7 is taken as U_2 . The expected value of \bar{y} is just the simple average of column (4), which turns out to be \bar{Y} , showing that \bar{y} is unbiased for \bar{Y} .

It is of interest to note that when this procedure is applied to selecting 2 units from a population of 4 units, the 4 possible samples, each selected with probability $1/4$, are U_1U_3 , U_2U_4 , U_3U_1 , U_4U_2 , which is effectively the same as selecting one of the samples U_1U_3 , U_2U_4 , with probability $1/2$. This shows that in this case css reduces to lss, that is, these two procedures become equivalent.

5.2b SAMPLING OF n UNITS

In selecting n units from N units with css, there are in all N possible samples, each having $1/N$ as its probability of selection. Clearly each unit in the population occurs in n of these N samples, since a unit can occur in a sample as the first, the second, ..., or the n -th unit. Hence, the expected value of the sample mean is

$$E(\bar{y}) = \frac{1}{N} \sum_{r=1}^N \left(\frac{1}{n} \sum_{i=1}^n y_i \right)_r = \frac{1}{nN} \sum_{r=1}^N \left(\sum_{i=1}^n y_i \right)_r = \bar{Y},$$

where r stands for the sample selected with random start r .

Another merit of css compared to lss is that this procedure retains the two advantages of a constant sample size from sample to sample and of the sample mean being an unbiased estimator of the population mean for *any* sampling interval. However k is usually taken as the integer nearest to N/n to ensure proper spread of the sample over the sampling frame. Further, since r is selected from 1 to N in this case, it is possible to estimate Y and \bar{Y} unbiasedly on the basis of the unit selected first or the first two selected units and so on.

5.3 USE OF FRACTIONAL INTERVAL

The disadvantages mentioned earlier in connection with lss with k as an integer can be overcome by selecting a linear systematic sample with N/n itself as k without rounding it off to the nearest integer. That is, the i -th unit in the sample is selected if $i-1 < r+jk \leq i$ for any $j = 0, 1, 2, \dots, (n-1)$. For instance, in

sampling 2 units from 5 units, l will be 2.5 ($= 5/2$) instead of 2 or 3 and in this case there are 25 possible samples, not all different. In fact, there are 5 different samples U_1U_3 , U_1U_4 , U_2U_4 , U_2U_5 and U_3U_5 occurring with equal frequency. The use of fractional interval in systematic sampling is equivalent to associating n different numbers with each unit such that the first unit gets the numbers 1 to n , the second unit from $n+1$ to $2n$ and so on and selecting units corresponding to a linear systematic sample of n numbers selected from 1 to Nn with N as the sampling interval.

If N is a multiple of n then sss and the procedure under consideration reduce to lss. Further, if N is very large compared to n , there is not much difference between the different procedures for all practical purposes. For the sake of simplicity, N is assumed to be a multiple of n in the subsequent discussions. The results obtained can, however, be generalized to the case where N is not a multiple of n .

5.4 OPERATIONAL CONVENIENCE

The operational simplicity of systematic sampling over srs is of considerable importance in large scale sampling. For instance, if a sample of 100 pages is to be selected from a book of 1000 pages for some philological studies it is easier, quicker and cheaper to select a systematic sample consisting of every 10th page than to select 100 pages by srs requiring the selection of 100 random numbers from 1 to 1000. Similarly in a population census if a 5% sample of persons is to be selected for purposes of quick tabulation of the results regarding age sex composition it would be operationally more convenient to take a systematic sample by distributing the person cards/slips into 20 pigeon holes in a specified order and selecting one of the 20 holes with equal probability than serially numbering all the thousands of cards/slips and selecting a simple random sample. Systematic sampling has particularly been found useful in forest surveys, where survey of randomly selected area units would be time consuming and costly due to difficulties of access and identification.

From the examples given above, it is clear that unlike srs, which requires prior serial numbering of all the units if not already existing, a systematic sample can be selected while progressively giving serial numbers, provided the sampling interval can be determined in advance on the basis of prior information. Of course, in giving progressive serial numbers special care is to be taken to avoid the possibility of any *bias*. In some cases, even the progressive serial numbering will not be necessary as it may conceptually be linked with counting, distance or time or some other measure. For instance, a sample of 100 punched cards from a deck of 10,000 cards occupying about 2 metres of space can be selected systematically by taking cards separated by a distance of 2 centimetres starting from a card at a point chosen at random from among the first 2 centimetres of space occupied by the cards. Similarly, in area surveys such as those involving determination of volume of timber in forest or area under a particular type of cultivation, it may be convenient to select and survey a systematic sample consisting of strips of specified width separated by a suitable length such as a kilometre.

Since the actual procedure of systematic sampling is quite simple, it would be easy to train persons in using it and hence it may be desirable to adopt this procedure whenever the sampling work has to be carried out by a large number of persons stationed in different areas or a very large sample is required to be selected at a central place. In fact, the sampling intervals and random starts to be used can be prespecified in a number of situations, thereby reducing the possibility of errors and biases in the process of sample selection. Further, considerable reduction in time and cost could be achieved by using mechanical devices in selecting systematic samples for large-scale surveys. It may be mentioned that systematic sampling has been used extensively in census work, examples of which are provided by (i) 2% sample of persons selected in the 1941 Indian population census to tabulate the data on means of livelihood, (ii) 1% sample of persons selected systematically from a 10% sample of 1950

census data in Japan to provide quick tabulation for important characteristics and (iii) 25% sample of households selected systematically in the 1960 United States population census for getting more detailed information. Because of the considerable potentialities of the usefulness of this method from purely operational considerations, it is necessary to study and understand the technical implications of the use of this procedure from the view point of sampling variability of the estimators commonly used in practice.

5.5 SAMPLING VARIANCE

In selecting n units from a population of $N (= nk)$ units with k as the interval, there are l possible samples and let \bar{y}_r be the sample mean of the r th possible sample ($r = 1, 2, \dots, l$). The sampling variance of the sample mean is, by definition,

$$V(\bar{y}) = \frac{1}{l} \sum_{r=1}^l (\bar{y}_r - \bar{Y})^2, \quad \bar{y}_r = \frac{1}{n} \sum_{i=1}^n y_{ri}, \quad (5.5)$$

where y_{ri} is the value of the characteristic for the i th unit in the r th systematic sample. Noting that

$$\sigma^2 = \frac{1}{nk} \sum_{r=1}^l \sum_{i=1}^n (y_{ri} - \bar{Y})^2$$

and that it can be written as the sum of σ_b^2 and σ_w^2 , which are the *between sample* and the *within sample* variances respectively,

$V(y)$ which is σ_b^2 can be expressed as

$$V(y) = \sigma^2 - \sigma_w^2, \quad \dots \quad (5.6)$$

where σ_w^2 is given by

$$\sigma_w^2 = \frac{1}{k} \sum_{r=1}^k \sigma_{wr}^2, \quad \sigma_{wr}^2 = \frac{1}{n} \sum_{i=1}^n (y_{ri} - \bar{y}_r)^2.$$

Since σ^2 is fixed for a given population, it is clear from (5.6) that in order to reduce $V(\bar{y})$ it is necessary to increase the within-sample variance as much as possible, which can be achieved by arranging the units in such a way that the units within each systematic sample are

as heterogeneous as possible with respect to the characteristic under consideration. Since a systematic sample is formed by systematically selecting one unit from each of n groups, into which the population units are conceptually divided, the heterogeneity between units within systematic samples can be achieved by ensuring heterogeneity between the n groups, which amounts to ensuring homogeneity between units within each of the n groups. This means that a good arrangement should ensure that units similar to one another with respect to the variable under consideration or an associated variable are put together.

The dependence of the sampling variance on the arrangement of units in systematic sampling is both an advantage and disadvantage in the sense that by effecting a *good* arrangement we can get better estimates and that a *bad* arrangement may lead to inefficient estimates. This shows that one has to be careful in using systematic sampling, and should at least ensure first that the existing arrangement does not lead to inefficient estimates before using systematic sampling, exploring at the same time the possibility of effecting a good arrangement.

An Example

Reverting back to the example of selecting 2 units from 4 units, let the values of the characteristic be

unit	U_1	U_2	U_3	U_4
value	1	2	3	4.

In sampling 2 units systematically from this population for estimating \bar{Y} , the sampling variance is 0.25, whereas the variances of sample mean in srs with and without replacement are respectively 0.625 and 0.417. The dependence of the variance on the arrangement of units in systematic sampling can be illustrated by considering the following arrangements

U_1	U_4	U_2	U_3	and	U_1	U_2	U_4	U_3
1	4	2	3		1	2	4	3

which give rise to the sampling variances of 1 and 0 respectively.

5.6 COMPARISON WITH SRS

By substituting $\frac{1}{n} \sum_{i=1}^n y_n$ for y_r in the expression for $V(\bar{y})$ given in (5 5), and expanding the squared term, $V(\bar{y})$ can be rewritten as

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n^2 k} \sum_{r=1}^k \left\{ \sum_{i=1}^n (y_n - \bar{Y}) \right\}^2 \\ &= \frac{1}{Nn} \sum_{r=1}^k \sum_{i=1}^n (y_n - \bar{Y})^2 + \frac{1}{Nn} \sum_{r=1}^k \sum_{i=1}^n \sum_{j \neq i}^n (y_{ri} - \bar{Y})(y_{nj} - \bar{Y}). \end{aligned}$$

The first term on the right hand side of the above expression is simply σ^2/n and the second term can be expressed as $(n-1)\rho_c\sigma^2/n$, where ρ_c , given by

$$\rho_c = \frac{\sum_{r=1}^k \sum_{i=1}^n \sum_{j \neq i}^n (y_{ri} - \bar{Y})(y_{nj} - \bar{Y})}{kn(n-1)\sigma^2}, \quad (5 7)$$

is a measure of correlation between pairs of sample units and it is termed *intraclass correlation coefficient*. Hence we have

$$V(\bar{y}) = \frac{\sigma^2}{n} \{1 + (n-1)\rho_c\} \quad (5 8)$$

It may be noted that since ρ_c is the correlation coefficient between pairs of units within the same systematic samples, and since the composition of the possible systematic samples depends on the arrangement of the units, the value of ρ_c would also depend much on the arrangement of the units in the population. Since σ^2 is fixed for a given population, it is clear from (5 8) that it is desirable to have such an arrangement for which ρ_c takes as large a negative value as possible. This again leads to the same principle regarding arrangement of the units as observed earlier namely, the units within the same systematic samples should be as heterogeneous as possible with respect to the characteristic under consideration. It is to be noted that since

$V(\bar{y}) < 0$, ρ_c cannot be less than $-1/(n-1)$. This result can also be obtained by equating (5.8) and (5.6), namely,

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

and expressing ρ_c in terms of σ_w^2/σ^2 , that is,

$$\rho_c = 1 - \frac{n}{n-1} \frac{\sigma_w^2}{\sigma^2}. \quad \dots \quad (5.9)$$

Noting that σ_w^2/σ^2 can at the most be equal to 1, we find that ρ_c cannot be less than $-1/(n-1)$. In fact, since $0 \leq \sigma_w^2/\sigma^2 \leq 1$, it follows from (5.9) that

$$-\frac{1}{n-1} \leq \rho_c \leq 1. \quad \dots \quad (5.10)$$

Comparing (5.8) with the variance of \bar{y} in srs with and without replacement (cf. (3.5) and (3.26) of Chapter 3), namely,

$$V'(\bar{y}) = \frac{\sigma^2}{n} \quad \dots \quad (5.11)$$

and

$$V''(\bar{y}) = \frac{N-n}{N-1} \frac{\sigma^2}{n}, \quad \dots \quad (5.12)$$

we find that

$$V(\bar{y}) < V'(\bar{y}), \quad \text{if } \rho_c < 0 \quad \dots \quad (5.13)$$

and

$$V(\bar{y}) < V''(\bar{y}), \quad \text{if } \rho_c < -\frac{1}{N-1}. \quad \dots \quad (5.14)$$

The relative efficiencies of systematic sampling compared to srs with and without replacement in estimating the population mean unbiasedly are respectively given by

$$E' = \frac{V'(\bar{y})}{V(\bar{y})} = \frac{1}{1 + (n-1)\rho_c} \quad \dots \quad (5.15)$$

and

$$E'' = \frac{V''(\bar{y})}{V(\bar{y})} = \frac{N-n}{N-1} E' = \frac{N-n}{N-1} \frac{1}{1 + (n-1)\rho_c}. \quad \dots \quad (5.16)$$

The values of the efficiencies E' and E'' for some specified values of ρ_e are given in Table 5 5

TABLE 5 5 RELATIVE EFFICIENCIES OF SYSTEMATIC SAMPLING

ρ_e	$E = V(y)/V(\bar{y})$	$E'' = V''(\bar{y})/V(\bar{y})$
(1)	(2)	(3)
$-\frac{1}{n-1}$	∞	∞
$-\frac{1}{N-1}$	$\frac{N-1}{N-n}$	1
0	1	$\frac{N-n}{N-1}$
1	$\frac{1}{n}$	$\frac{1}{n} \frac{N-n}{N-1}$

5.7 BEHAVIOUR OF SAMPLING VARIANCE

From the expression for variance in (5 8) and Table 5 5, we find that for systematic sampling to be very efficient, it is necessary to have as high a negative correlation as possible between pairs of units within the samples and that systematic sampling can turn out to be considerably inefficient if the intraclass correlation has a high positive value. One method of effecting high negative intraclass correlation would be to arrange the units in ascending or descending order of a characteristic associated with the variable under consideration. In case the auxiliary variable is a classificatory characteristic, then the units are to be so arranged that the units belonging to a particular class come together and that the groups which are not much different among themselves occur together. It may be noted that the behaviour of the sampling variance is not as regular as in the case of srs due to the fact that it not only depends on n and σ^2 but also on the intra-class correlation which generally varies with sample size and arrangement of units. In this section, some empirical examples are given to illustrate the behaviour of the sampling variance with increase

in sample size for different arrangements. The effect of different types of arrangements of units such as with linear trend and cyclical variation on the sampling variance is considered in Sections 5.9, 5.10 and 5.11.

5.7a SAMPLING OF STRIPS IN A FOREST

The data on volume of timber for 176 forest strips given in Annexure 4.2 of Chapter 4 have been used in calculating the sampling variances for systematic samples of size 2, 4, 8 and 16 strips by enumerating all possible systematic samples and the values of the efficiencies E' , E'' and ρ_c , the intraclass correlation, are shown in Table 5.6. The above experiment is repeated after rearranging the units in the population in increasing order of the strip-length by giving fresh serial numbers. As the volume of timber is expected to be related to the strip-length, the arrangement according to this is likely to be approximately similar to the arrangement according to volume of timber, the study variable. The results of this experiment are also given in Table 5.6.

TABLE 5.6. EFFICIENCIES OF SYSTEMATIC AND SIMPLE RANDOM SAMPLING IN ESTIMATING VOLUME OF TIMBER.

sample size n	relative variance			relative efficiency (%)			intraclass correlation ρ_c
	systematic sampling	srs swr	srs wor	E' (3)/(2)	E'' (4)/(2)		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	
1. arrangement as in frame							
2	0.0823	0.1510	0.1501	183.48	182.38		-0.4550
4	0.0748	0.0755	0.0742	100.94	99.20		-0.0031
8	0.0395	0.0377	0.0362	95.44	91.65		0.0068
16	0.0041	0.0189	0.0173	460.98	421.95		-0.0522
2. arrangement in increasing order of strip length							
2	0.0737	0.1510	0.1501	204.88	203.66		-0.5119
4	0.0413	0.0755	0.0742	182.81	179.66		-0.1510
8	0.0085	0.0377	0.0362	443.53	425.88		-0.1106
16	0.0041	0.0189	0.0173	460.98	421.95		-0.0522

From Table 5.6 it can be seen that the behaviour of sampling variance with increase in n in systematic sampling is rather irregular unlike in srs, where the sampling variance decreases with increase in n .

according to a known function of λ , n and σ^2 . It is of interest to note that in the example considered systematic sampling is almost as efficient as or more efficient than sampling of strips with srswr or srs wr for $n = 2, 4, 8$ and 16 when the existing arrangement is used. In this connection it may be mentioned that in forest surveys it is operationally convenient to use systematic sampling with the existing arrangement and that this usually turns out to be more efficient from the view point of sampling variance also. By comparing the relative efficiencies of systematic sampling we find that the arrangement of units in order of strip length has been of considerable help in reducing the sampling variance except possibly for large sample sizes. Another point needing attention is that the efficiency of systematic sampling tends to be substantial for larger sample sizes even if the intraclass correlation in such cases turns out to be small in magnitude provided of course it is negative.

5.7b SYSTEMATIC SAMPLING OF VILLAGES

The relative variances of estimates of total cultivated area and 1961 census population based on systematic samples of various sizes have been worked out for the population of 128 villages given in Annexure 41 of Chapter 4. This study has been carried out using three types of arrangement of the population units namely

- (i) existing arrangement as obtained in the frame
- (ii) increasing order of geographical area and
- (iii) increasing order of 1951 census population

The results of this study are given in Table 5.7. The arrangements (ii) and (iii) are considered since the cultivated area is likely to be related to geographical area and the 1961 census population to the 1951 census population.

From Table 5.7 the behaviour of the sampling variance with increase in n is seen to be rather irregular and hence it is difficult to predict in practice. Unlike in srs the behaviour of the systematic sampling variance varies with the characteristics and arrangements

TABLE 5.7. BEHAVIOUR OF RELATIVE VARIANCE IN SYSTEMATIC AND SIMPLE RANDOM SAMPLING.

sample size n	srswr	srs wor	systematic sampling with arrangement		
			as in frame	area	population
(1)	(2)	(3)	(4)	(5)	(6)
(i) cultivated area.					
2	0.1622	0.1609	0.2026	0.1249	0.1405
4	0.0811	0.0792	0.0797	0.0329	0.0701
8	0.0406	0.0383	0.0336	0.0139	0.0242
16	0.0203	0.0179	0.0173	0.0074	0.0083
32	0.0101	0.0077	0.0009	0.0018	0.0005
64	0.0051	0.0026	0.0001	0.0012	0.0003
(ii) 1961 census population					
2	0.1777	0.1763	0.2385	0.1512	0.1303
4	0.0889	0.0868	0.1263	0.0564	0.0409
8	0.0444	0.0420	0.0680	0.0215	0.0121
16	0.0222	0.0196	0.0194	0.0164	0.0033
32	0.0111	0.0084	0.0013	0.0093	0.0005
64	0.0056	0.0028	0.0007	0.0039	0.0003

of the population units. However, in the examples considered here, the relative variance has decreased with increase in n and sometimes this reduction in the sampling variance is substantial compared to the corresponding reduction in srswr and srs wor. Further, it is of interest to note that with the arrangement of units as in the frame, systematic sampling turns out to be worse than srs wor for smaller sample sizes. Comparison of the relative variances over the three types of arrangements shows that arrangement according to geographical area is generally more efficient than the arrangement in the frame or that according to the 1951 census population for estimating cultivated area, whereas the arrangement according to 1951 census population is more efficient than the other two arrangements.

5.7c SYSTEMATIC SAMPLING IN CENSUS

As mentioned earlier, systematic sampling, being operationally simple and convenient, is widely used in large-scale sampling work such as in population or other censuses for collecting detailed information or for quick tabulation of results. Some results of a

large scale experimental study (Lahiri, Poti and Banerjee, 1957) conducted on the 1941 census data with a view to studying the efficiency of systematic sampling are presented in Table 5.8

TABLE 5.8 EFFICIENCY OF SYSTEMATIC SAMPLING COMPARED TO SIMPLE RANDOM SAMPLING IN A POPULATION CENSUS

sr no	characteristic	region	percentage of total population	sampling fraction (%)	efficiency (%)
(1)	(2)	(3)	(4)	(5)	(6)
1	males	A	50.01	10	141.3
				4	180.5
				2	165.2
				1	140.8
2	married (males)	A	23.64	4	224.0
		B	24.54	0.5	98.0
				0.25	135.0
3	married (females)	A	24.21	4	125.1
		B	24.50	0.5	148.0
				0.25	161.0
4	self supporting persons	A	31.29	4	143.2
		B	23.38	0.5	148.0
				0.25	169.0
5	age group 15-24 (males)	A	7.12	4	134.5
		B	7.01	0.5	66.0
6	age group 15-24 (females)	A	7.20	4	98.5
		B	7.14	0.5	125.0
7	age group 45-54 (males)	A	4.09	4	129.3
		B	3.73	0.5	75.0
8	age group 45-54 (females)	A	3.88	4	78.8
		B	3.57	0.5	124.0
9	religion—Hindu (males)	A	43.46	4	163.7
		B	42.89	0.5	215.0
				0.25	181.0
10	religion—Hindu (females)	A	43.37	4	298.9
		B	42.46	0.5	167.0
				0.25	97.0
11	religion—Muslims	A	12.05	4	339.1
12	agricultural labourers	A	4.94	4	111.6
		B	2.50	0.25	117.0
13	persons engaged in agriculture other than (12)	A	72.35	4	410.1
		B	82.95	0.25	108.0
14	persons engaged in production other than cultivation	A	4.77	4	130.1
		B	5.62	0.25	130.0

Source Lahiri D B, Poti J and Banerjee S (1957) *Studies in Population Sampling*, Chapters 3 and 4 Mimeo graphed Indian Statistical Institute Calcutta.

In this study, confined to a part of Hazaribagh district in Bihar State, the 1941 census individual slips of about 35000 persons of Barhi police station area (region A), arranged almost in the order of enumeration, and those of about 665000 persons of Sadar sub-division (region B), arranged by sex, have been utilized to compare the efficiency of systematic sampling with that of srs for different characteristics for various sampling fractions. The sampling variances for systematic sampling have been estimated on the basis of a sample of the possible systematic samples. From Table 5.8 it can be seen that in most of the cases systematic sampling is more efficient than srs though it is slightly inefficient in a very few cases. Further, these results also bring out the irregularity of the behaviour of the efficiency of systematic sampling with increase in sample size, since it varies from characteristic to characteristic, and even for the same characteristic it differs from one arrangement to another.

5.8 ESTIMATION OF SAMPLING VARIANCE

In systematic sampling, it is not possible to estimate unbiasedly the variances of the estimators of population mean and total on the basis of a single sample. This is a disadvantage, as one important requirement for adopting any sampling method is that it should be able to provide an estimate of the sampling error. However, it is possible to build up some biased but useful variance estimators on the basis of a systematic sample. Before considering such variance estimators, let us briefly examine why it is not possible to estimate unbiasedly the sampling variance from a single systematic sample.

5.8a LINEAR SYSTEMATIC SAMPLING

It has been shown earlier that the variance of an unbiased estimator of \bar{Y} or Y can be estimated unbiasedly, if it is possible to estimate Y^2 unbiasedly, for if \hat{Y} is an unbiased estimator of Y , then $V(\hat{Y}) = E(\hat{Y}^2) - Y^2$. Hence, an unbiased variance estimator is given by $v(\hat{Y}) = \hat{Y}^2 - Est(Y^2)$, where $Est(Y^2)$ stands for an unbiased estimator of Y^2 , that is, of $\sum_{i=1}^N Y_i^2 + \sum_{i=1}^N \sum_{i' \neq i} Y_i Y_{i'}$. In the case of linear systematic

sampling, though it is possible to estimate $\sum_{i=1}^N Y_i^2$ unbiasedly by $L \sum_{i=1}^n y_i^2$,

the same is not possible for $\sum_{i=1}^N \sum_{j \neq i} Y_i Y_j$, since some of the $\binom{N}{2}$ pairs of units, such as any two neighbouring units, have no chance at all of getting selected in any sample when $k \geq 2$

One of the possible methods of getting an idea of the variance of a systematic sample estimator is to use biased estimators considered here. Since lss can be considered to amount to grouping the N population units into $n/2$ groups of $2k$ units each and selecting two units from each group in a systematic manner, the estimator of the population mean (namely, the sample mean) can be written as the mean of the $n/2$ group sample means, that is,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{2}{n} \sum_{i=1}^{n/2} \left(\frac{y_{2i} + y_{2i-1}}{2} \right). \quad \dots \quad (5.17)$$

Assuming that the two units have been selected with srs wr from the $2k$ units in the i th group, an unbiased estimator of the variance of the i th term in the brackets on the right hand side of (5.17) will be given by

$$\frac{k-1}{k} \left(\frac{y_{2i} - y_{2i-1}}{2} \right)^2$$

Hence, an estimator of $V(\bar{y})$ is given by

$$v_1(\bar{y}) = \frac{N-n}{Nn^2} \sum_{i=1}^{n/2} (y_{2i} - y_{2i-1})^2 \quad \dots \quad (5.18)$$

An alternative variance estimator based on the same principles as those considered above which takes into account successive differences of the sample values, is given by

$$v_2(\bar{y}) = \frac{(N-n)}{Nn} \sum_{i=1}^{n-1} \frac{(y_{i+1} - y_i)^2}{2(n-1)}. \quad \dots \quad (5.19)$$

Balanced Differences

Other variance estimators can be obtained by considering the overlapping differences of the type

$$\text{and } d_{3i} = \frac{1}{2} y_i - y_{i+1} + \frac{1}{2} y_{i+2}, \quad i = 1, 2, \dots, n-2,$$

$$d_{5i} = \frac{1}{2} y_i - y_{i+1} + y_{i+2} - y_{i+3} + \frac{1}{2} y_{i+4}, \quad i = 1, 2, \dots, (n-4),$$

which are termed *balanced differences*, instead of the successive differences used earlier, and they are given by

$$v_{31}(\bar{y}) = \frac{N-n}{Nn} \cdot \frac{1}{1.5(n-2)} \sum_{i=1}^{n-2} d_{3i}^2, \quad \dots \quad (5.20)$$

and

$$v_{32}(\bar{y}) = \frac{N-n}{Nn} \cdot \frac{1}{3.5(n-4)} \sum_{i=1}^{n-4} d_{5i}^2. \quad \dots \quad (5.21)$$

These variance estimators are based on the results obtained by Cochran (1946) and Yates (1948). These estimators are not unbiased and should be used with considerable caution, as inferences based on estimates subject to bias may be misleading in practice. Before using these estimators, it is desirable to get an idea of the bias in specific cases by comparing them with sampling variance obtained using the entire data for the same variable or an auxiliary variable that may be available for an earlier period.

5.8b CIRCULAR SYSTEMATIC SAMPLING

The above discussion applies to circular systematic sampling also in a substantial measure, since in this case also some of the possible ($\frac{N}{2}$) pairs of units do not have any chance of being selected in any sample, except in very special cases mentioned below. Hence, variance estimators given for linear systematic sampling may also be used for circular systematic sampling.

Unbiased Variance Estimator

In certain very special cases of css, it is possible to estimate the sampling variance unbiasedly. For instance, in sampling 3 units circular systematically from a population of 5 units with 2 as the sampling interval, it can be seen that all the $\binom{5}{2}$ or 10 pairs of units have some chance of being selected (each of the pairs (12), (15), (23), (34) and (45) occurs in one sample and each of the pairs (13), (14), (24), (25) and (35) occurs in two samples) and hence it is possible to estimate the variance unbiasedly. In fact,

if y_1, y_2, y_3 are the three sample values for one of the 5 possible samples the unbiased estimator of the variance of the sample mean \bar{y} is given by

$$v(\bar{y}) = \bar{y}^2 - \frac{1}{5} \left\{ \sum_{i=1}^3 \frac{y_i^2}{r_i} + \sum_{i=1}^3 \sum_{j \neq i} \frac{y_i y_j}{r_{ij}} \right\}$$

where r_i and r_{ij} are respectively the number of samples in which the unit U_i occurs and the units U_i and U_j occur together. In general this situation is obtained in sampling $(N+1)/2$ of the units circular systematically using the sampling interval 2 when N is an odd number. In this case the unbiased variance estimator of \bar{y} is given by

$$v(\bar{y}) = \bar{y}^2 - \frac{1}{N} \left\{ \sum_{i=1}^{\frac{N-1}{2}} \frac{y_i^2}{r_i} + \sum_{i=1}^{\frac{N-1}{2}} \sum_{j \neq i} \frac{y_i y_j}{r_{ij}} \right\} \quad (5.22)$$

where r_i and r_{ij} are as defined earlier.

5.8c RANDOM ARRANGEMENT OF UNITS

If all the units in the population are arranged at random, then systematic sampling is equivalent to srs wr, since in both the cases the probability of selecting a particular sample of n units is the same namely, $1/\binom{N}{n}$. For, in selecting n units with css from N units with interval k (which reduces to lss if $N = nk$), the number of possible systematic samples (not necessarily different) when all the $N!$ arrangements are considered is $(N!)N$. The number of systematic samples in which a particular set of n units occurs is given by $n!(N-n)!N$, since the number of arrangements where these n units are equally spaced at intervals of k in a particular random order is $(N-n)!$ and corresponding to each arrangement of the $(N-n)$ units, there are $n!$ arrangements for the n units and the first unit in the sample may occupy any one of the N places. Hence, the probability of getting a particular sample s of n units is

$$P(s) = \frac{n!(N-n)!N}{N!N} = \frac{1}{\binom{N}{n}}$$

In this case \bar{y} is unbiased for \bar{Y} with the variance $(1-f)\sigma^2/n$ and an unbiased estimator of the variance is given by

$$v_4(\bar{y}) = \frac{N-n}{Nn} s^2, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (5.23)$$

The variance estimator (5.23) may be used as an approximation to $V(\bar{y})$ in systematic samples drawn from a population, where the units are not strictly arranged at random, but there is evidence to believe that the given arrangement is independent of the variable under consideration and hence may be considered almost random with respect to that variable. An example of such a situation is provided by systematic sampling of villages for a population study using a list of villages arranged alphabetically.

For a given arrangement, the expected value of the variance estimator $v_4(\bar{y})$ is given by

$$\begin{aligned} E\{v_4(\bar{y})\} &= \frac{N-n}{Nn} E(s^2) \\ &= \frac{N-n}{Nn} \cdot \frac{1}{k} \sum_{r=1}^k \sigma'_{wr}^2 = \frac{(N-n)}{N(n-1)} \cdot \frac{1}{k} \sum_{r=1}^k \sigma_{wr}^2, \end{aligned}$$

where $\sigma'_{wr}^2 = n\sigma_{wr}^2/(n-1)$, σ_{wr}^2 being the variance between the units within the r -th systematic sample. Hence,

$$E\{v_4(\bar{y})\} = \frac{(N-n)}{N(n-1)} \sigma_w^2, \quad \dots \quad (5.24)$$

where σ_w^2 is as defined in (5.6). The variance estimator (5.23) over-estimates, is unbiased for, or under-estimates the systematic sampling variance σ_b^2 given in (5.6) according as

$$\frac{N-n}{N(n-1)} \sigma_w^2 \gtrless \sigma_b^2,$$

that is,

$$\frac{N-n}{N-1} \frac{\sigma^2}{n} \gtrless \sigma_b^2,$$

since $\sigma^2 = \sigma_b^2 + \sigma_w^2$. Thus, on the average this variance estimator over-estimates whenever systematic sampling is more efficient than srs wr, is unbiased if the two sampling schemes are equivalent and under-estimates when systematic sampling is less efficient. Because

of this result, the estimator (5.23) may be used to give an upper limit for the sampling variance in the cases where systematic sampling is expected to be more efficient than srs

5.8d USE OF INTERPENETRATING SUB-SAMPLES

A method of estimating the variance unbiasedly, which is resorted to often in practice because of its operational convenience, consists in selecting the sample of required size n in the form of two or more systematic sub samples of same size with independent random starts. If $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_m$ are estimates of the population mean based on m such sub samples each of size n/m , then an unbiased estimator of the variance of the combined estimator

$$\bar{y} = \frac{1}{m} \sum_{i=1}^m \bar{y}_i$$

is given by

$$v_s(\bar{y}) = \frac{1}{m(m-1)} \sum_{i=1}^m (\bar{y}_i - \bar{y})^2 \quad (5.25)$$

When $m = 2$, this expression reduces to $(\bar{y}_1 - \bar{y}_2)^2/4$

If the sampling interval is small, selection of m random starts with replacement may lead to repetition of samples and hence in such cases it is desirable to select the m interpenetrating systematic sub samples of n' ($= n/m$) units each with random starts selected from 1 to L' ($= mL$) without replacement. This amounts to selection of m of the L possible samples with equal probability without replacement and the variance of the sample mean is given by

$$V(\bar{y}) = \frac{L'-m}{L'm} \frac{1}{L'-1} \sum_{i=1}^{L'} (\bar{Y}_i - \bar{Y})^2, \quad (5.26)$$

where \bar{Y}_i is the i th sample mean, $i = 1, 2, \dots, L'$. If $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_m$ are m estimates based on m systematic samples with random starts selected without replacement, an unbiased estimator of $V(\bar{y})$ is given by

$$v_s(\bar{y}) = \frac{L'-m}{L'm} s^2, \quad s^2 = \frac{1}{m-1} \sum_{i=1}^m (\bar{y}_i - \bar{y})^2 \quad (5.27)$$

Though these variance estimators are unbiased, they are not likely to be precise if m is small and if the number of sub-samples is increased by decreasing the sub-sample size, the combined estimator of the population mean is likely to become less efficient than the estimator based on a single combined systematic sample. Hence, usually one has to strike a balance between the need for getting a good estimate of the population parameter and a good variance estimate. The variance estimators (5.25) and (5.27) may also be used as an approximation after dividing the selected systematic sample of n units into sub-samples in a systematic manner, as this avoids the need for selecting the sample in the form of sub-samples of smaller size, thereby retaining the efficiency generally obtained by taking a large systematic sample. For instance, if the sample is to be divided into two sub-samples, it can be done by taking the units with odd orders of selection in the sample to form one sub-sample and the other sample units to form the second sub-sample.

5.8e AN ILLUSTRATIVE EXAMPLE

The validity of the variance estimators (5.18) and (5.19) has been studied for systematic samples of different sizes drawn from the population of 128 villages given in Annexure 4.1 of Chapter 4 to estimate the total cultivated area and the total 1961 census population. The villages are arranged in order of geographical area for estimating the cultivated area and in order of 1951 census population for estimating the 1961 census population. The expected values of these variance estimators have been calculated for the two characteristics for different sample sizes to examine the sign and the magnitude of bias, if any, and the relative biases are presented in Table 5.9. In this connection it may again be pointed out that the variance estimators themselves are subject to sampling variance and that in practice one should attempt not only to get reliable estimates of the population parameter, but also to get sufficiently precise variance estimates, as otherwise the error estimate may be misleading for drawing inferences on the basis of the survey data. The relative mean square errors of the variance estimators are also shown in Table 5.9.

TABLE 5 9 RELATIVE BIAS AND RELATIVE MEAN SQUARE ERRORS OF THE VARIANCE ESTIMATORS

sample size	relative bias (B/V) for		relative mean square error ($M(v)/V^2$) for	
	v_1	v_2	v_1	v_2
(1)	(2)	(3)	(4)	(5)
<i>cultivated area</i>				
16	+0.043	-0.021	0.384	0.128
32	+0.113	+0.296	0.237	0.153
64	-0.142	-0.179	0.024	0.064
<i>1961 census population</i>				
16	-0.222	-0.322	0.324	0.146
32	+0.262	+0.112	0.072	0.037
64	-0.517	-0.522	0.268	0.274

From Table 5 9 it appears that the behaviour of the bias and the mse of the variance estimators with increase in sample size is irregular and that the estimator $v_2(y)$ has comparatively lesser mse than $v_1(y)$ though it has relatively a larger bias in some cases than the other estimator. These results are to be taken only as indicative of the validity and efficiency of these estimators. Further these two estimators and the estimators $v_3(\bar{y})$ based on balanced differences are likely to be useful only if a proper arrangement introducing a linear trend in the population has been effected. In case the arrangement of the units is almost random then the estimator $v_1(y)$ is likely to be efficient.

5 9 LINEAR TREND

From Table 5 7 we find that arrangements of the villages in order of geographical area and 1951 census population are very useful in increasing the efficiency of systematic sampling for estimating the mean and total of cultivated area and 1961 census population respectively. This increase in efficiency of systematic sampling for these arrangements is possibly due to the presence of a *trend* in the values of the units, as is evidenced by Figures 5 2 and 5 3.

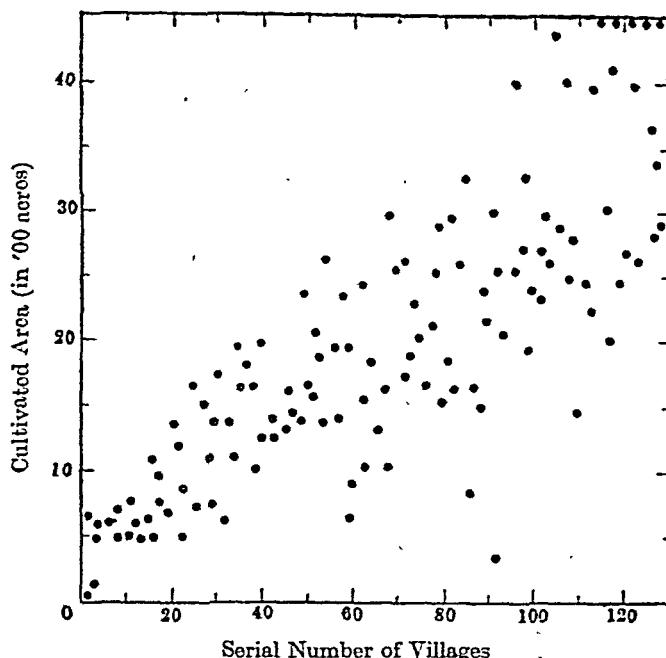


Figure 5.2. Scatter diagram of cultivated area with villages in increasing order of geographical area.

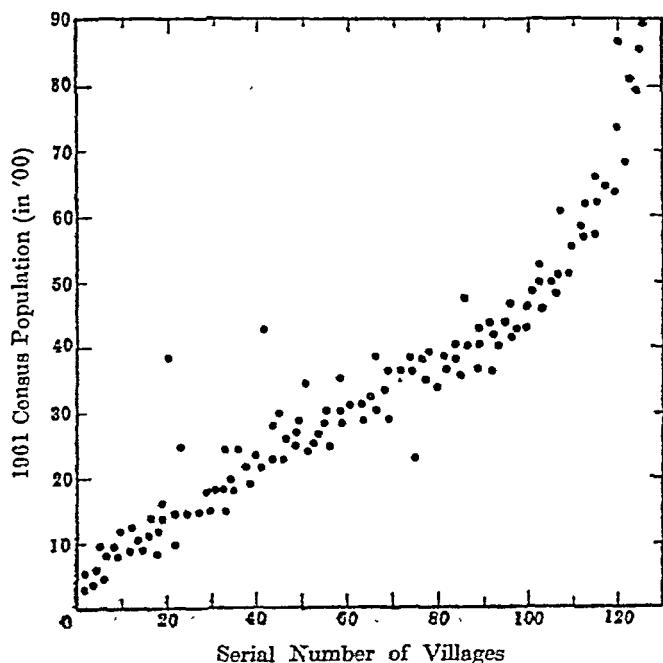


Figure 5.3. Scatter diagram of 1961 census population with villages

5.9a A HYPOTHETICAL POPULATION

To illustrate the effectiveness of linear trend in increasing the efficiency of systematic sampling, let us consider a hypothetical population, where the values of the N units are in *arithmetical progression*, that is,

$$Y_i = a + bi, \quad i = 1, 2, \dots, N, \quad \dots \quad (5.28)$$

where a and b are constants. In this case \bar{Y} and σ^2 are given by

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i = a + b \frac{N+1}{2} \quad \dots \quad (5.29)$$

and

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 \\ &= \frac{b^2}{N} \sum_{i=1}^N \left(i - \frac{N+1}{2} \right)^2 = \frac{b^2}{12} (N^2 - 1), \end{aligned} \quad \dots \quad (5.30)$$

since $\sum_{i=1}^N i = \frac{N(N+1)}{2}$ and $\sum_{i=1}^N i^2 = \frac{1}{6} N(N+1)(2N+1)$. Hence, the variance of the sample mean based on n units selected with srs wr and srs wor respectively are

$$V'(\bar{y}) = \frac{\sigma^2}{n} = \frac{b^2}{12n} (N^2 - 1) = \frac{b^2}{12} \left(k - \frac{1}{n} \right) (N+1) \quad \dots \quad (5.31)$$

and

$$V''(\bar{y}) = \frac{N-n}{N-1} \frac{\sigma^2}{n} = \frac{b^2}{12n} (N-n)(N+1) = \frac{b^2}{12} (k-1)(N+1), \quad \dots \quad (5.32)$$

where $k = N/n$.

When N/n is assumed to be an integer for the sake of simplicity, the values of the units in the systematic sample with r as the random start are given by

$$\{a + b(r+jk)\}, \quad j = 0, 1, 2, \dots, n-1,$$

and hence the sample mean based on the r -th systematic sample is

$$\bar{y}_r = \frac{1}{n} \sum_{j=0}^{n-1} \{a + b(r+jk)\} = a + b \left(r + \frac{n-1}{2} k \right). \quad \dots \quad (5.33)$$

The expected value of the sample mean \bar{y} is

$$\begin{aligned} E(\bar{y}) &= \frac{1}{k} \sum_{r=1}^k \bar{y}_r = a + b \left(\frac{k+1}{2} + \frac{n-1}{2} k \right) \\ &= a + \frac{1}{2}b(N+1) = \bar{Y}. \end{aligned}$$

The variance of \bar{y}_r is, by definition, given by

$$\begin{aligned} V(\bar{y}) &= \frac{1}{k} \sum_{r=1}^k (\bar{y}_r - \bar{Y})^2 \\ &= \frac{1}{k} \sum_{r=1}^k b^2 \left(r - \frac{k+1}{2} \right)^2, \end{aligned}$$

which after expansion and simplification becomes

$$V(\bar{y}) = \frac{b^2}{12} (k^2 - 1). \quad \dots \quad (5.34)$$

By comparing (5.34) with (5.31) and (5.32), we find that systematic sampling is more efficient than srs, that is,

$$V(\bar{y}) < V''(\bar{y}) < V'(\bar{y}) \quad \dots \quad (5.35)$$

for $2 \leq n < N$ and that in case N is large compared to n , the efficiency of systematic sampling is n times that of srs in estimating \bar{Y} , that is,

$$V(\bar{y}) : V''(\bar{y}) : V'(\bar{y}) = \frac{1}{n} : 1 : 1. \quad \dots \quad (5.36)$$

5.9b END CORRECTIONS

If a linear trend is present in the population, its estimator for \bar{Y} can sometimes be improved by giving the weights

$$\frac{1}{n} + \frac{(2r-k-1)}{2(n-1)k} \text{ and } \frac{1}{n} - \frac{(2r-k-1)}{2(n-1)k} \quad \dots \quad (5.37)$$

to the first and the last units in the sample respectively instead of the usual weight of $1/n$. These weights have been determined such that when applied to the special linear population considered in Sub section 5.9a the estimator turns out to be \bar{Y} giving rise to zero variance. For, letting the weights for the first and the last units to be $\left(\frac{1}{n} + x\right)$ and $\left(\frac{1}{n} - x\right)$ respectively we get the estimate for the r th systematic sample as

$$\begin{aligned}\bar{y}_r &= \frac{1}{n} \sum_{j=0}^{n-1} \{a + b(r+jL)\} + x(a+b r) - x[a+b(r+(n-1)L)] \\ &= a+b\left(r+\frac{n-1}{2}L\right) - xb(n-1)L\end{aligned}$$

Equating this to the population mean $a+\frac{1}{2}b(N+1)$ and solving for x , we get

$$x = \frac{2r-k-1}{2(n-1)L}$$

These corrections to the estimator based on a systematic sample drawn from a population exhibiting a linear trend in the population values are called *end corrections* (Yates, 1948). It may be pointed out that these end corrections may make the estimator slightly biased though the variance is likely to be reduced.

5.9c CENTRALLY LOCATED SAMPLE

Another situation where the estimator in the case of the hypothetical linear population equals \bar{Y} is obtained by considering only the systematic sample with the start $(k+1)/2$ if k is odd or the two systematic samples with starts $k/2$ and $(k+2)/2$ if k is even. For, if k is odd, substituting $r = (k+1)/2$ in \bar{y}_r , we get

$$\bar{y}_r = a+b\left(\frac{k+1}{2} + \frac{n-1}{2}L\right) = a+b\frac{N+1}{2}$$

and if k is even substituting $k/2$ and $(k+2)/2$ for r , we get

$$\bar{y}_{k/2} = a+\frac{1}{2}bN \text{ and } \bar{y}_{(k+2)/2} = a+\frac{1}{2}b(N+2),$$

the mean of which is \bar{Y} . Hence, it may be desirable to consider only the systematic sample with $(k+1)/2$ as the random start if k is odd and the systematic samples with $k/2$ or $(k+2)/2$ as random starts for selection with probability $1/2$ if k is even, whenever there is a linear trend present in the population. Such a sample is known as a *centrally located sample*. But in practice, it is not advisable to use such a sample, especially when one is in doubt about the presence of perfect linear trend in the arrangement used, since it is not a valid sample due to certain units not getting any chance at all of being included in the sample and hence it is subject to bias and it is not possible to estimate the error involved in the estimator.

5.9d BALANCED SYSTEMATIC SAMPLING

We consider here a systematic sampling technique, which reduces the variance of \bar{y} when there is a linear trend in the population and makes \bar{y} exactly equal to \bar{Y} in the case of the hypothetical population ($Y_i = a + bi$). This procedure, known as *balanced systematic sampling* (bss), consists in considering the population as divided into $n/2$ groups of $2k$ units each and selecting from each group a pair of units equidistant from the end units of that group in a systematic manner. For instance, the units corresponding to the serial numbers r and $2k-r+1$, where r is a number selected at random from 1 to k , are considered selected from the first group of $2k$ units. Then the two units to be selected from the second group of $2k$ units correspond to the serial numbers $r+2k$ and $4k-r+1$, and so on. Thus we see that the balanced systematic sample of n (even) units with random start r consists of the units corresponding to the numbers

$$\{r+2jk, 2(j+1)k-r+1\}, \quad j = 0, 1, 2, \dots, \left(\frac{n}{2}-1\right). \quad \dots \quad (5.38)$$

The sample mean based on the r -th balanced sample for the hypothetical population is given by

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{j=0}^{\frac{n}{2}-1} [a + b(r+2jk) + a + b\{2(j+1)k-r+1\}] \\ &= a + \frac{b}{n} \sum_{j=0}^{\frac{n}{2}-1} \{2(2j+1)k+1\} = a + b \frac{N+1}{2}, \end{aligned}$$

which shows that for this population the sample mean equals the population mean for all possible samples, thereby reducing the sampling variance to zero.

In the general case, the sample mean based on the r -th balanced systematic sample is

$$\bar{y}_r = \frac{1}{n} \sum_{j=0}^{\frac{n}{2}-1} \{Y_{r+2jk} + Y_{2(j+1)k-r+1}\}. \quad \dots \quad (5.39)$$

Since in this case also there are k possible samples and one of these is selected with probability $1/k$ the expected value of \bar{y} is given by

$$E(\bar{y}) = \frac{1}{nk} \sum_{r=1}^k \sum_{j=0}^{n-1} S \{ Y_{r+2jk} + Y_{2(j+1)k-r+1} \} = \bar{Y},$$

since each population unit occurs only in one of the k samples. This procedure is being used in the Indian National Sample Survey since 1955 for purposes of special studies and this has been considered in detail by Sethi (1965)

It is of interest to note that bss leads to optimum sampling in the case of selecting 2 units with equal probability from a population of N (even) units when the units are arranged in increasing or decreasing order of the values of the characteristic under consideration. Taking $N = 4$ for the sake of simplicity let Y_1, Y_2, Y_3, Y_4 be the values of the 4 units in the population such that $Y_1 < Y_2 < Y_3 < Y_4$. In selecting 2 units from these 4 units there are three possible ways of pairing the units namely $(Y_1, Y_2), (Y_3, Y_4), (Y_1, Y_3), (Y_2, Y_4)$. In each case there are two samples of 2 units each one of which can be selected with probability $1/2$. The absolute differences between the two sample means in these cases are given by half of

$$|(Y_1 + Y_2) - (Y_3 + Y_4)| \quad |(Y_1 + Y_3) - (Y_2 + Y_4)| \quad |(Y_1 + Y_4) - (Y_2 + Y_3)|$$

Rewriting these as

$$\begin{aligned} & |(Y_3 - Y_1) + (Y_4 - Y_2)| \quad |(Y_2 - Y_1) + (Y_4 - Y_3)| \\ & |(Y_4 - Y_2) - (Y_3 - Y_1)| \text{ or } |(Y_4 - Y_3) - (Y_2 - Y_1)| \end{aligned}$$

and comparing these we find that the last pair has the minimum "between sample difference" and hence that pairing of units may be considered optimum. Incidentally the second pairing is equivalent to lss and css considered earlier and this shows that for the arrangement considered here bss is more efficient than lss and css.

The above result can be generalized to the case of sampling n units from N units and this shows that in selecting n (even) units from N units ($N = nk$) with bss, optimum sampling is achieved in each

of the $n/2$ groups of $2k$ units. A better procedure for sampling n units from N would be to form first all balanced systematic samples of 2 units and then to use these $N/2$ pairs of units to form $N/4$ samples of 4 units by arranging the $N/2$ pairs in increasing order of the total value of the characteristic and by taking pairs of these pairs in a balanced manner. Proceeding similarly we can form all possible samples of size 2, 4, 8, 16, ..., by successive application of the bss technique and finally take one of the samples of a given size with equal probability. This procedure can be applied easily whenever the sample size is of the form 2^m . It may be pointed out that though the bss procedure has been discussed only for particular cases, such as N being even and multiple of twice the sampling interval, etc., it will be possible to apply this procedure to other cases also by adopting certain devices such as introduction of a few dummy units with values 0 or some other suitably specified constants.

5.9e SAMPLING FOR MULTIPLE CHARACTERISTICS

If parameters of more than one characteristic are to be estimated in a survey and if these characteristics are related to an auxiliary characteristic on which data are available, then the population totals and means of these characteristics can be efficiently estimated by taking a systematic or balanced systematic sample after arranging the units in increasing or decreasing order of the values of the auxiliary characteristic. For instance, if the characteristics under study are total consumer expenditure, expenditures on cereals, food, clothing, fuel, light, etc. for a region, then the sampling units (e.g. villages) may be arranged in increasing order of their population in a previous period before drawing a systematic or balanced systematic sample; or in an industrial survey if the characteristics under study are total wages, output, working capital, etc., then before systematic selection the factories may be arranged in increasing order of number of workers.

However, if all the characteristics under study are not related to one single auxiliary characteristic, then the arrangement according to a particular auxiliary variable will be efficient only for some

and not all the characteristics under study. For instance, population in a previous period and geographical area may be suitable auxiliary variables for estimating population characteristics and crop statistics respectively. In such a case, arrangement of the units according to population may not be efficient for estimating crop statistics and similarly arrangement in order of geographical area may not be efficient for estimating current population. If we have different arrangements, one according to population and the other according to area, then we get two different samples and hence the cost of the surveys may be high. However, it would be very desirable to integrate the two surveys, since it would be more economical to collect data on population and area from the same sample units than from two different samples.

In the above case the following procedure of having a compromise arrangement may be adopted. First, the units may be arranged in increasing order of one auxiliary characteristic (say area). With this arrangement all possible systematic or balanced systematic samples of size 2 are formed. Then these pairs of units are arranged in increasing order of their population. From this population of pairs of units all systematic or balanced systematic samples of 2 pairs (that is, samples of size 4) are formed. These samples are arranged in increasing order of their geographical area and all possible samples of size 8 are obtained. This procedure is continued till the samples of required size are realized. From these, one sample is selected with equal probability. Though this procedure can be directly applied only when two auxiliary characteristics are involved and the sample size is of the form 2^n , suitable modifications to this would enable the application of this principle to other cases also.

Suppose there are two auxiliary variables which are classificatory characteristics. For instance, in the above example, if the sampling units can be classified into two classes P and P' on the basis of their population and into A and A' on the basis of their geographical area, the four categories into which the units get classified can be arranged in the order PA , PA' , PA' , $P'A$, the advantage being that

the units belonging to the classes P , P' , A and A' come together if a circular systematic sample is selected. This procedure can be easily extended to the case of 3 or more classificatory characteristics, but in that case it may be necessary to give priority for certain characteristics and categories in respect of ensuring their occurrence together in the arrangement. This and the earlier procedures can be applied even in the case of uni-variate surveys when two or more suitable auxiliary variables are available. This pattern of arrangement of units on the basis of 2 or more auxiliary characteristics is being used in the Indian National Sample Survey (NSS Instructions to Field Workers, 1955, cf. Sub-section 15.3h of Chapter 15).

5.9f ILLUSTRATIVE EXAMPLES

To study the efficiencies of the centrally located sample and bss, the relative mse's of the estimator of \bar{Y} in these cases have been calculated using the data for the 128 villages, considered earlier for studying the efficiency of systematic sampling. In this study the successive arrangement of the units and samples of sizes 2, 4, 8, etc. alternatively according to ascending order of geographical area and 1951 census population, described earlier, has also been considered besides the three simpler arrangements, namely, arrangement as in the list, in order of geographical area and in order of 1951 census population. The results of this study are presented in Table 5.10. From this table it appears that bss is more efficient than lss when a good arrangement has been effected. But the possibility of the former being less efficient than the latter, when the arrangement is not favourable, is brought out by the results relating to the arrangement as in the frame. The performance of the centrally located sample is quite good in this example.

5.10 PERIODIC VARIATION

If in a population the units with large, medium and small values follow one another according to a regular repetitive pattern (cycle), then systematic sampling should be used with considerable care.

That is, suppose the arrangement of the units in the population is such that the values of the units follow a pattern similar to the curve in Figure 5.4

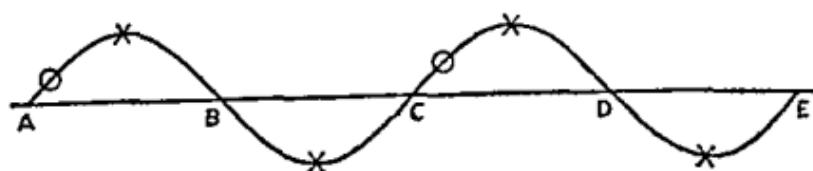


Figure 5.4 A specimen population with regular periodic variation

TABLE 5.10 EFFICIENCIES OF DIFFERENT TYPES OF SYSTEMATIC SAMPLING

arrangement of units	sampl ing methods*	relative mean square error for sample size					
		2	4	8	16	32	64
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>characteristic 1 cultivated area</i>							
as in frame	SS	0.2026	0.0797	0.0336	0.0173	0.0009	0.0001
	BSS	0.1797	0.1141	0.0488	0.0225	0.0083	0.0015
	CLS	0.7323	0.0471	0.0258	0.0027	0.0008	0.0001
increasing order of geographical area	SS	0.1249	0.0329	0.0139	0.0074	0.0018	0.0012
	BSS	0.0504	0.0216	0.0077	0.0045	0.0002	0.0002
	CLS	0.0111	0.0111	0.0041	0.0082	0.0033	0.0001
increasing order of area and 1951 population	SS	0.1249	0.0647	0.0207	0.0123	0.0009	0.0006
	BSS	0.0504	0.0284	0.0169	0.0061	0.0025	0.004
	CLS	0.0111	0.0024	0.0313	0.0094	0.0001	0.0006
<i>characteristic 2 total number of persons in 1961</i>							
as in frame	SS	0.2385	0.1263	0.0680	0.0194	0.0013	0.0007
	BSS	0.1564	0.1057	0.0335	0.0294	0.0113	0.0100
	CLS	1.7909	0.2779	0.0283	0.0286	0.0020	0.0007
increasing order of 1951 population	SS	0.1303	0.0409	0.0121	0.0033	0.0005	0.0003
	BSS	0.0310	0.0092	0.0030	0.0002	0.0001	0.0002
	CLS	0.0198	0.0055	0.0002	0.0007	0.0042	0.0003
increasing order of area and 1951 population	SS	0.1512	0.0481	0.0220	0.0046	0.0033	0.0007
	BSS	0.1686	0.2552	0.0104	0.0018	0.0002	0.0052
	CLS	0.0195	0.0160	0.0339	0.0031	0.0016	0.0007

* SS systematic sampling, BSS balanced systematic sampling, CLS centrally located sample

In a regular pattern AB, BC, CD and DE will be equal in length and the length AC ($= CE$, etc.) is called the *period* of the cycle. When the sampling interval is equal to or is a multiple of the period, the variation in the systematic sample estimates will be extremely large, since in that case the units in any sample have the same value (see circles in Figure 5.4). If the periodicity of the curve is known, then the interval may be taken as an odd multiple of half the period, in which case the variation in the sample estimates will be considerably reduced (see crosses in Figure 5.4.).

Periodic variation is likely to occur when the population consists of groups of equal or approximately equal number of units and the units within each group are arranged according to some definite pattern. An example of such a situation is provided by the population census enumeration slips, where the universe or population consists of households, within each of which the individuals are usually arranged according to a set pattern such as head of the household occurring first, then his wife and their children in order of their age. In such a case, systematic sampling with an interval equal to the group size or its multiple will result in inefficient samples, since the units selected in a sample will be having more or less similar characteristics. As populations with various kinds of periodicity (more or less regular) are not very uncommon in survey practice, one should carefully examine a population for the existence of possible periodic variation before using systematic sampling. This is important as the presence of periodicity can be of much help in reducing the variance of a systematic sample in some situations if the interval is properly chosen, and when this is not possible, the periodic pattern can be eliminated by a suitable re-arrangement of the units. The pitfalls involved in using systematic sampling in case of populations having periodic variation have been discussed in detail, among others, by Stephan, Deming and Hansen (1940) and Lahiri (1954).

5.11 SUITABLE ARRANGEMENT

The situation under which systematic sampling is likely to be more efficient, besides being operationally more convenient, than srs w^r may be described as that where the values of an associated characteristic, when plotted against the serial numbers of the population units, exhibit a linear trend or show periodic undulations without too much of fluctuations about the linear or curvilinear path and where the sampling interval used is smaller than the smallest period of the observed oscillations. Further, it would be desirable to select the systematic sample by first dividing the arranged population into broad groups of consecutive units and then selecting systematic samples in each of the groups with the same interval but with independently selected random starts, as this would reduce the possibility of getting *bad samples*.

So far we have been assuming the existence and availability of an auxiliary variable x highly correlated with the characteristics under study for effecting suitable arrangement before using systematic sampling. In the absence of such a variable, any qualitative or even subjective information, adequate for classifying the units into more or less homogeneous groups can be utilized in arranging the units into those groups with a view to increasing the efficiency of systematic sampling. For instance, in sampling families for a family budget enquiry in rural area, the agricultural families may be put together first and then the non agricultural families or the families may be arranged by per capita expenditure class, if that information were available. Such arrangements will ensure that the different categories of the population are represented in due proportions in the sample.

5.12 DISTRIBUTION OF SAMPLE PROPORTION

In Sub section 3.9c of Chapter 3 we have considered the sampling distribution of the sample mean based on a simple random sample and found that the distribution tends to get peaked at the population mean as the sample size increases even when sampling from highly skew populations. The sampling distribution of the sample mean based on a

systematic sample would also behave in the same manner provided the arrangement of the units is properly effected and this distribution would approach the normal distribution more rapidly than in srs as the sample size is increased, since systematic sampling is more efficient than srs for a suitable arrangement of the units. At this stage, it may be pointed out that the method of estimating the population mean and the sampling variance discussed so far for systematic sampling can readily be extended to the case of estimating the population proportion in a certain class by considering the value of each unit as 1 or 0 according as that unit belongs to or does not belong to that class.

In order to examine the sampling distribution of the estimator of the population proportion based on a systematic sample, the sample proportions of cultivators for the 400 possible systematic samples of about 840 persons obtained in selecting a 1 in 400 systematic sample from the population of about 336,000 males enumerated in the 1941 Indian Population Census in Sadar sub-division of Hazaribagh district in Bihar State as well as those for 1 in 200 and 1 in 100 systematic samples, given by Lahiri, Poti and Banerjee (1957), are considered. The sampling distributions of the sample proportion of cultivators for different sampling fractions are given in Figure 5.5, from which it is clear that as the sample size increases the concentration of the sample values about the population value 0.795 increases and that the distribution becomes symmetrical.

5.13 DETERMINATION OF SAMPLE SIZE

Since the sampling variance of an estimator based on a systematic sample depends much on the arrangement of the units in the population, we have seen that it is difficult to predict its behaviour with increase in sample size. When the arrangement is properly effected, the variance may be expected to decrease with increase in n . But even in this case, a simple relationship between sampling variance and n does not exist. Hence, it is difficult to determine the sample size required to ensure a specified precision for the estimates and this could only be done on the basis of extensive empirical studies using values of some related characteristics.

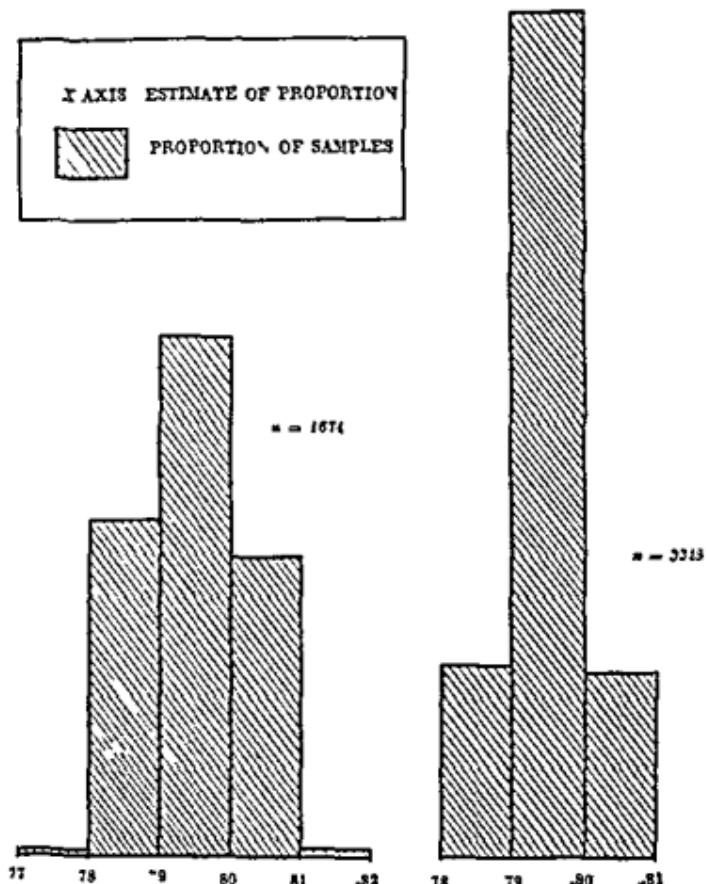
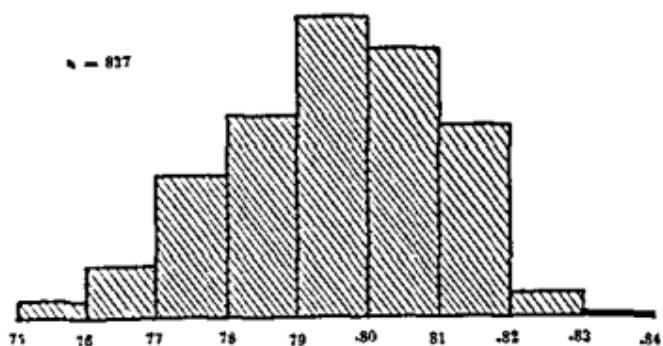


Figure 5.6. Histograms of the sampling distribution of the proportion of cultivators for different sample sizes

5.14 TWO-DIMENSIONAL SYSTEMATIC SAMPLING

So far we have considered uni-dimensional systematic sampling in the sense that the units in the population are arranged or considered to be arranged on a line. There may be situations where the population is such that the units are having a natural arrangement or can be considered as arranged on a plane instead of on a line and application of systematic sampling to such a situation is termed *two-dimensional systematic sampling* or *plane systematic sampling* (Das, 1949; Quenouille, 1949). For instance, if the universe consists of a specified region divided into N square or rectangular grids of equal size and if a sample of n such grids is to be selected so as to get representation from different parts of the region in the sample, then it would be desirable to divide the region into square or rectangular cells of equal size such that each cell consists of $k (= N/n)$ grids and to select the grid occupying the r -th place in each of the n cells, where r is a number chosen at random from 1 to k . It may be noted that it is convenient to locate the selected grid in each cell by two coordinate numbers (say, $i j$), instead of by one serial number, by considering the k grids in each cell as arranged in the form of l rows and m columns, such that $k = lm$, and any pair of numbers (i, j) , $i \leq l$ and $j \leq m$, determines a particular grid uniquely. Thus the random location of a grid in a cell can be specified by two numbers (i, j) selected at random from 1 to l and 1 to m . A representation of a two-dimensional systematic sample is given in Figure 5.6.

•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•

Figure 5.6. A representation of two-dimensional systematic sampling showing the selected grids in the cells by dots.

From this figure it is clear that there is considerable spread of the sample units over the region and hence this sample is likely to be more representative of the population units, thereby possibly having smaller variability, than a simple random sample. The expressions for estimators of the population mean, total and proportion, their variances and variance estimators can be worked out proceeding on lines similar to those in the case of uni dimensional systematic sampling (cf Problem 5.16, p. 182)

REFERENCES

- BUCKLAND, W R (1951) A review of the literature of systematic sampling, *J. Roy Stat. Soc. (B)*, 13, 208-215
- COCHRAN, W G (1946) Relative accuracy of systematic and stratified random samples for a certain class of populations, *Ann. Math. Stat.*, 17, 164-177
- COCHRAN, W G (1963) *Sampling Techniques*, Second Edition, Chapter 8, John Wiley & Sons New York
- DAS, A C (1949) Two dimensional systematic sampling, *Science and Culture*, 15, 157-158, *Sankhya*, 10 (1950) 95-108
- FINNEY, D J (1948) Random and systematic sampling in timber surveys, *Forestry* 22, 64-99
- HASEL, A A (1942) Estimation of volume in timber stands by strip sampling, *Ann. Math. Stat.*, 13, 179-206
- LAHIRI, D B (1954) On the question of bias of systematic sampling, *Proceedings of World Population Conference*, 6, 349-362
- LAHIRI, D B, POTI, J and BANERJEE, S (1957) Studies on population sampling—an experimental approach, Vol I and II, mimeographed, Indian Statistical Institute, Calcutta
- MADOW, W G and MADOW, L H (1944) Theory of systematic sampling, *Ann. Math. Stat.*, 15, 1-24
- MADOW, L H (1946) Systematic sampling and its relation to other sampling designs, *J. Amer. Stat. Assn.*, 41, 201-217
- NAIR, K R and BHARGAVA, R P (1951) Statistical sampling in timber surveys in India, *Indian Forest Leaflet* No 153, Forest Research Institute, Dehradun
- QUENOUILLE, M H (1949) Problems in plane sampling, *Ann. Math. Stat.*, 20, 355-375

- SETHI, V. K. (1965) : On optimum pairing of units; *Sankhyā*, 27, (B), 315-320.
- STEPHAN, F. F., Deming, W. E. and Hansen, M. H. (1940) : The sampling procedure of the 1940 population census, *J. Amer. Stat. Assn.*, 35, 615-630.
- SUKHATME, P. V., PANSE, V. G. and SASTRY, K. V. R. (1958) : Sampling techniques for estimating the catch of sea fish in India; *Biometrics*, 14, 78-96.
- YATES, F. (1948) : Systematic sampling; *Phil. Trans. Roy. Soc.*, 241, (A), 345-377.

COMPLEMENTS AND PROBLEMS

5.1 Suppose a sample of n units is to be drawn from a finite population of N units using css.

(i) If the sampling interval k is taken as the integer nearest to (N/n) , find the condition under which there would not be any need to repeat or cross the random start to achieve the required sample size.

(ii) State, giving reasons, whether you prefer the sampling interval $[N/n]$ or $[N/n]+1$ for estimating the population mean \bar{Y} .

(iii) Can the interval be fixed arbitrarily, and if so, what is its effect on the unbiased nature of the estimator and its variance, when the units have been arranged in ascending or descending order of a size measure?

(iv) If instead of using one fixed interval I , a sequence of prespecified intervals $\{I_j\}$, $j = 1, 2, \dots, n-1$, is used, I_j being the interval used for selecting the $(j+1)$ -th unit, derive an unbiased estimator for \bar{Y} .

5.2 Data for a small population exhibiting a fairly steady rising trend are given in Table 5.11.

TABLE 5.11. VALUES OF A VARIABLE y FOR A POPULATION OF 40 UNITS.

unit	y	unit	y	unit	y	unit	y
(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1	0	11	10	21	22	31	39
2	1	12	11	22	25	32	43
3	2	13	13	23	29	33	46
4	1	14	12	24	30	34	50
5	4	15	12	25	32	35	53
6	5	16	15	26	35	36	52
7	7	17	14	27	33	37	57
8	7	18	17	28	38	38	59
9	9	19	20	29	40	39	63
10	8	20	23	30	41	40	62

- (i) Calculate the relative efficiency of lss as compared to srs wr in estimating \bar{Y} when the sample size is 4.

(u) Determine the effect of reversing the order of observations in the second set and in the fourth set in the table on the relative efficiency. How do you account for the change in the efficiency?

53 A list of 108 villages in a tehsil arranged in ascending order of geographical area is given in Table 5.12 together with village wise area under winter paddy

TABLE 5.12 DATA ON GEOGRAPHICAL AREA (x) AND AREA UNDER WINTER PADDY (y) FOR 108 VILLAGES

no	x	y	no	x	y	no	x	y	no	x	y
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	103	41	28	240	108	55	370	98	82	658	230
2	106	33	29	241	94	56	379	270	83	678	396
3	120	87	30	243	116	57	389	79	84	681	239
4	120	78	31	244	58	58	396	99	85	682	166
5	121	56	32	246	47	59	397	147	86	691	83
6	121	62	33	248	69	60	400	187	87	698	232
7	124	58	34	249	44	61	404	273	88	710	282
8	128	19	35	251	56	62	410	118	89	716	191
9	135	64	36	259	160	63	418	130	90	716	303
10	137	61	37	264	102	64	433	158	91	727	303
11	145	74	38	264	102	65	445	116	92	730	288
12	147	13	39	266	187	66	453	194	93	738	286
13	151	81	40	271	23	67	460	161	94	805	239
14	153	41	41	273	129	68	462	222	95	803	242
15	160	58	42	274	51	69	467	223	96	864	146
16	166	44	43	280	161	70	501	96	97	873	445
17	176	65	44	287	179	71	503	164	98	897	487
18	178	69	45	292	76	72	514	318	99	910	354
19	185	29	46	313	137	73	515	272	100	924	340
20	206	46	47	320	127	74	541	155	101	1034	401
21	209	93	48	324	104	75	542	292	102	1117	261
22	216	38	49	327	115	76	543	214	103	1155	613
23	224	87	50	333	105	77	562	275	104	1195	227
24	229	72	51	349	245	78	570	100	105	1323	704
25	230	127	52	350	117	79	586	418	106	1419	682
26	235	114	53	364	170	80	601	189	107	1473	373
27	238	88	54	365	210	81	653	129	108	1495	164

(x and y in acres, 1 acre = 0.4047 hectare)

(i) Draw 5 circular systematic samples of 7 villages each with the following five independent random starts : 45, 3, 18, 62 and 37.

(ii) Making use of the 35 sample observations obtained in (i), estimate the relative efficiency of css as compared to that of srs wr for estimating the total area under paddy (Y) based on a sample of 7 villages.

(iii) Obtain a single combined estimate of Y based on all the 5 samples drawn in (i) and also estimate its rse.

5.4 The area under cultivation in 29 plots forming a systematic sample drawn from a region having 290 plots are given in Table (5.13).

TABLE 5.13. DATA ON CULTIVATED AREA (y) FOR 29 SAMPLE PLOTS.

plot		plot		plot	
(1)	(2)	(1)	(2)	(1)	(2)
1	0.0	11	2.8	21	2.3
2	0.9	12	2.6	22	2.9
3	0.0	13	2.3	23	2.1
4	0.0	14	3.5	24	6.3
5	0.3	15	2.4	25	8.2
6	0.1	16	3.8	26	5.4
7	0.5	17	4.1	27	6.5
8	3.1	18	4.9	28	6.6
9	2.8	19	6.0	29	4.1
10	2.7	20	5.4		

(y in hectares : 1 hectare = 2.471 acres).

(i) Estimate the total cultivated area Y in the region (\hat{Y} , say).

(ii) Obtain an estimate of the rse of \hat{Y} using the method of successive differences.

(iii) Compare the estimate of rse obtained in (ii) with the rse estimated on the assumption of srs wr and comment on the result.

5.5 Data on number of seedlings in every individual foot of sown bed, which is 80 feet in length, are shown in Table 5.14 in a rectangular form for convenience.

(i) Find the rse of the estimator of the total number of seedlings based on a systematic sample consisting of every 10th foot of the sown bed.

(ii) Determine the relative efficiency of systematic sampling as compared to that of srs wr when the sample size is 8 one-foot bed-lengths.

5.6 A pilot survey for investigating the possibility of estimating the catch of marine fish was conducted in a sample of fishing centres on the Malabar Coast of India. At each landing centre in the sample, a count was made of the number of boats landing every hour from 6 a.m. to 6 p.m. Out of the boats landing during each hour, the first one was selected for observation on weight of fish, the product of this with the number

TABLE 5.14 DATA ON NUMBER OF SEEDLINGS IN A 80 FEET BED

1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
26	16	27	37	4	36	20	21
28	9	20	14	5	20	21	26
11	22	25	14	11	43	16	16
16	26	39	24	9	27	14	18
7	17	24	18	25	20	13	11
22	39	25	17	16	21	9	19
44	21	18	14	13	18	25	27
26	14	44	38	22	19	17	29
31	40	55	36	18	24	7	31
26	30	39	29	9	30	30	29

of boats giving an estimate of the catch of fish during the hour. Data on the number of boats landing and the catch of fish at a particular centre on a particular day are given in Table 5.15.

TABLE 5.15 NUMBER OF BOATS (x) LANDING DURING EACH OF 12 HOURS (6 A M TO 6 P M) AND THE CATCH OF FISH IN MAUNDS (y)

hours	1	2	3	4	5	6	7	8	9	10	11	12
x	42	52	19	6	23	56	36	59	14	14	2	6
y	569	887	223	89	352	1295	934	1255	488	443	98	0

(1 maund = 82.5 lbs = 37,422 kilogrammes)

(i) Calculate the relative efficiencies of LSS as compared to those of SRS WOT for estimating the population totals of x and y when the sample sizes are 2, 3, 4 and 6, taking each hour as the sampling unit.

(ii) Also study the behaviour of the intraclass correlation coefficient for these two variables as the sample size increases.

(Sukhatme, P. V., Panse V. G. and Sastry, K. V. R., *Biometrika*, 45, (1958), 78-96)

5.7 (i) A 10% systematic sample is selected from a list of persons, in which families have been arranged by religion, and within families persons have been arranged by age. For which of the following characteristics do you expect this sample to be more efficient than one drawn with SRS WOT?

- (a) proportion of persons belonging to the different religions
- (b) average age of males, and
- (c) proportion of children going to school

(ii) How will the comparative picture in (i) change, if the arrangement of families in the list and that of persons within families are randomized before selection?

5.8 If the sampling interval k taken as $[N/n] + 1$ turns out to be a sub multiple of n , show that there would be a repetition of one unit in each circular systematic sample and that the LSS would give rise to samples of only $(n-1)$ units.

5.9 Show that the procedure of using fractional interval in systematic sampling described in Section 5.3 is equivalent to selection of units corresponding to the serial numbers $\left\{ n + \left[j \frac{N}{n} \right] \right\}, j = 0, 1, 2, \dots, (n-1)$, taken in a circular manner with r being a random start from 1 to N .

(Lahiri, D. B., (1954), unpublished).

5.10 Derive the result given in (5.22) relating to the variance estimate in the case of sampling $(N+1)/2$ units circular systematically using 2 as the interval when N is an odd number.

5.11 Express the variance of the mean based on a systematic sample of size n , drawn from a population of N units, in terms of the *serial correlation coefficients* ρ_a defined as

$$\rho_a = \frac{1}{k(n-\alpha)\sigma^2} \sum_{i=1}^k \sum_{i'=1}^{n-\alpha} (Y_{ii'} - \bar{Y})(Y_{i(i'+\alpha)} - \bar{Y}),$$

where $\alpha = 1, 2, \dots, (n-1)$ and σ^2 is the population variance.

(Madow, W. G. and Madow, L. H., *Ann. Math. Stat.*, 15, (1944), 1-24).

5.12 Suppose a finite population of N units having the values $\{Y_i\}$, $i = 1, 2, \dots, N$, are presumed to be drawn from the following two *super-populations* with the following models :

- (i) $E(Y_i) = \mu$, $V(Y_i) = \sigma^2$, and $\text{Cov}(Y_i, Y_{i'}) = 0$, $i' \neq i$;
- (ii) $E(Y_i) = \alpha + \beta i$, $V(Y_i) = \sigma^2$ and $\text{Cov}(Y_i, Y_{i'}) = 0$, $i' \neq i$.

For each of the above two cases, derive the expressions for the expected variances of the sample mean \bar{y} based on samples of size n selected (a) systematically, and (b) with srs w/o r when N is a multiple of n , and compare them.

(Cochran, W. G., *Sampling Techniques*, Ch. 8 (1963), 215-216).

5.13 Assuming that a finite population of N units is drawn from an *auto-correlated* super-population with the following model

$$E(Y_i) = \mu, \quad V(Y_i) = \sigma^2, \quad \text{and } \text{Cov}(Y_i, Y_{i+v}) = \rho_v, \text{ for all } i \text{ and } v \geq 1,$$

where $\rho_u \geq \rho_v \geq 0$ for $u < v$ and

$$\delta_i^2 = \rho_{i+1} - \rho_{i-1} + 2\rho_i \geq 0, \quad i = 2, 3, \dots, (N-2),$$

show that the expected variance of the sample mean in the case of systematic sampling of n units is less than that in the case of srs. N may be taken as a multiple of n .

(Cochran, W. G., *Ann. Math. Stat.*, 17, (1946), 164-177).

5.14 Show that the result mentioned in Problem 5.13 remains valid even if the first two conditions of the super-population are relaxed such that $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma_i^2$.

(Quenouille, M. H., *Ann. Math. Stat.*, 20, (1949), 355-375).

5.15 Suppose a population of N units is classified into g groups and we are interested in estimating the group means for a specified characteristic. Assuming the statistical model

$$Y_{ij} = \mu_i + e_{ij}, \quad j = 1, 2, \dots, N_i, \quad i = 1, 2, \dots, g, \quad \sum_{j=1}^{N_i} e_{ij} = 0, \quad \sum_{i=1}^g N_i = N,$$

find the condition for systematic sampling of n units to be more efficient than srs wr for estimating the group means. N may be assumed to be a multiple of n , and n to be a multiple of N .

5.16 A two dimensional population with a linear trend may be represented by the relation $Y_{ij} = i + j$ $i = 1, 2, \dots, N$, and $j = 1, 2, \dots, M$, where Y_{ij} is the value of the unit belonging to the i th row and the j th column of the rectangle into which the NM units of the population can be arranged. For selecting a two-dimensional systematic sample of nm units, where $n = (N/k)$ and $m = (M/l)$, k and l being integers, two independent random starts i and j from 1 to k and 1 to l respectively are chosen. Then a sample of size nm units consists of all units whose coordinates are of the form $(i + \alpha k, j + \beta l)$, $\alpha = 1, 2, \dots, (n-1)$ and $\beta = 1, 2, \dots, (m-1)$.

(i) Show that the sample mean is unbiased for the population mean.

(ii) Obtain the sampling variance of the sample mean and compare it with the variance of the sample mean in the case of selecting nm units with srs wr from the NM population units.

5.17 In a multi subject household survey covering a population of N households, it was found necessary to have varying intensity of sampling for the different subjects and sometimes with *sub frames* consisting of units relevant for particular subjects. Further, from operational considerations, it was decided not to select the same household for more than one enquiry. Obtain the expressions for unbiased estimators of the relevant population totals in the following cases.

(i) A sample of households is selected linear systematically for subject *A* with the sampling interval I_1 . Another sample of m households is selected circular systematically from a *sub frame* of N_1 households having a specified characteristic relevant for subject *B* with the modification that if a household selected for *B* has already been sampled for *A* then it will be replaced in the sample by the household next to it in the *sub frame* considering the first household in the *sub frame* as the next to the last one.

(ii) A combined sample of households is selected linear systematically with the sampling interval I . From this combined sample, linear systematic sub samples are selected with intervals I_1 and I_2 for subjects *C* and *D* respectively with the provision that every common sample household will be substituted by the next household in the combined sample for *D*. The remaining sample households of the combined sample are taken up for *A*.

(iii) In (ii) another sample of m households is selected for subject *B* from a *sub frame* consisting of N_1 households circular systematically as in (i) with the provision that if any household selected for *B* has already been sampled for *A*, *C* or *D*, it will be substituted by the next household in the *sub frame*, considering the first household as the next to the last one.

Varying Probability Sampling

6.1 MEASURE OF SIZE OF A UNIT

Under certain circumstances, selection of units with unequal probabilities provides more efficient estimators than equal probability sampling, and this type of sampling is known as *unequal* or *varying probability sampling*. In the most commonly used varying probability sampling scheme, the units are selected with probability proportional to a given measure of size (pps) where the size measure is the value of an auxiliary variable x related to the characteristic y under study and this sampling scheme is termed *probability proportional to size sampling*. For instance, the number of persons in some previous period may be taken as a measure of the size in sampling area units for a survey of socio-economic characters, which are likely to be related to population. Similarly, in estimating crop characteristics the geographical area or cultivated area for a previous period, if available, may be considered as a measure of size, or in an industrial survey, the number of workers may be taken as the size of an industrial establishment.

Since a large unit, that is, a unit with a large value for the study variable y , contributes more to the population total than smaller units, it is natural to expect that a scheme of selection which gives more chance of inclusion in a sample to larger units than to smaller units would provide estimators more efficient than equal probability sampling. Such a scheme is provided by pps sampling, size being the value of an auxiliary variable x directly related to y . It may

appear that such a selection procedure would give biased estimators as the larger units are over represented and the smaller units are under represented in the sample. This would be so, if the sample mean is used as an estimator of \bar{Y} . Instead, if the sample observations are suitably weighted at the estimation stage taking into consideration their probabilities of selection, it is possible to obtain unbiased estimators. Mahalanobis (1938) has referred to this procedure in the context of sampling plots for a crop survey and this procedure has been discussed in detail by Hansen and Hurwitz (1943).

6.2 SELECTION OF ONE UNIT WITH PPS

Suppose the population total Y of the variable y for a population of N units is to be estimated by sampling one unit with pps, size being the value of a related variable x . Let (X_i, Y_i) be the values of x and y for the i th unit U_i , $i = 1, 2, \dots, N$. The probability of selecting U_i will be $P_i = X_i/X$, ($X = \sum_{i=1}^N X_i$). Let the selected unit have the values y_1 and x_1 for the variables y and x respectively. An unbiased estimator of Y is given by

$$\hat{Y} = y_1/p_1 \quad (p_1 = x_1/X), \quad (6.1)$$

for $E(\hat{Y}) = \sum_{i=1}^N (Y_i/P_i)P_i = Y$. The sampling variance of \hat{Y} is given by

$$V(\hat{Y}) = \sum_{i=1}^N \left(\frac{Y_i}{P_i} - Y \right)^2 P_i = \sum_{i=1}^N \frac{Y_i^2}{P_i} - Y^2 \quad (6.2)$$

From this expression, it is clear that the variance will be small if P_i is roughly proportional to Y_i . In fact, the variance is zero when P_i and Y_i are exactly proportional. Hence, it may be expected that if the regression of y on x is found to be a straight line passing through the origin, pps sampling will be very efficient (cf. Sub section 6.4b).

It is of interest to note that srs is a particular case of pps sampling where $P_i = 1/N$, $i = 1, 2, \dots, N$, and that in this case the formulae (6.1) and (6.2) for the estimator and variance become $\hat{Y} = Ny_1$ and $V(\hat{Y}) = N^2\sigma^2$.

An Example

In estimating the total Y of a hypothetical population of four units given in Table 6.1 by selecting one unit with pps, size being the value of x , the sampling variance turns out to be 1, whereas the variance in the case of srs is 16.36. This illustration shows that there are cases, where pps sampling is considerably more efficient than srs.

TABLE 6.1. A HYPOTHETICAL POPULATION OF FOUR UNITS.

unit	U_1	U_2	U_3	U_4
size (x)	1	2	3	4
variable (y)	0.5	1.2	2.1	3.2

6.3 PPS SAMPLING WITH REPLACEMENT

As in srswr, sampling of n units with pps can also be done with replacement (ppswr). Suppose $\{y_i, p_i\}$ are the sample observation and the initial probability of the unit selected at the i -th draw, $i = 1, 2, \dots, n$. The n quantities $\{y_i/p_i\}$, $i = 1, 2, \dots, n$, are independent random variables, each taking the N values $\{Y_j/P_j\}$ with probabilities $\{P_j\}$, $j = 1, 2, \dots, N$, and they are unbiased for Y with the sampling variance given in (6.2). Hence, an unbiased estimator of Y in this case is

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} \quad \dots \quad (6.3)$$

with the sampling variance

$$\begin{aligned} V(\hat{Y}) &= \frac{1}{n^2} \sum_{i=1}^n V\left(\frac{y_i}{p_i}\right) = \frac{1}{n^2} \sum_{i=1}^n \left\{ \sum_{j=1}^N \left(\frac{Y_j}{P_j} - Y \right)^2 P_j \right\} \\ &= \frac{1}{n} \sum_{i=1}^N \left(\frac{Y_i}{P_i} - Y \right)^2 P_i = \frac{1}{n} \left(\sum_{i=1}^N \frac{Y_i^2}{P_i} - Y^2 \right), \quad \dots \quad (6.4) \end{aligned}$$

the covariance terms $\left\{ \text{Cov} \left(\frac{y_i}{p_i}, \frac{y_{i'}}{p_{i'}} \right) \right\}$, $i' \neq i$, being zero. This shows that the variance of the estimator is inversely proportional to the sample size n as in srswr.

Alternative Derivation

The expected value and the variance of \hat{Y} in (6.3) can alternatively be derived by rewriting it as

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^N r_i \frac{Y_i}{P_i}, \quad \sum_{i=1}^N r_i = n \quad (6.5)$$

where r_i is the number of occurrences of U_i in the sample and it is 0 when U_i is not selected. Noting that

$$E(r_i) = nP_i, \quad V(r_i) = nP_i(1-P_i) \quad \text{and} \quad \text{Cov}(r_i, r_j) = -nP_iP_j$$

we see that

$$E(\hat{Y}) = \frac{1}{n} \sum_{i=1}^N E(r_i) \frac{Y_i}{P_i} = Y$$

and

$$\begin{aligned} V(\hat{Y}) &= \frac{1}{n^2} \left\{ \sum_{i=1}^N V(r_i) \frac{Y_i^2}{P_i^2} + \sum_{i=1}^N \sum_{j \neq i}^N \text{Cov}(r_i, r_j) \frac{Y_i}{P_i} \frac{Y_j}{P_j} \right\} \\ &= \frac{1}{n} \left\{ \sum_{i=1}^N (1-P_i) \frac{Y_i^2}{P_i} - \sum_{i=1}^N \sum_{j \neq i}^N Y_i Y_j \right\} = \frac{1}{n} \left(\sum_{i=1}^N \frac{Y_i^2}{P_i} - Y^2 \right) \end{aligned}$$

Since in sampling with ppswr the values $\{y_i/p_i\}$, $i = 1, 2, \dots, n$, are n independent unbiased estimators of Y having the same variance, an unbiased estimator of $V(\hat{Y})$ is given by

$$v(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{p_i} - \hat{Y} \right)^2 = \frac{1}{n(n-1)} \left(\sum_{i=1}^n \frac{y_i^2}{p_i^2} - n\hat{Y}^2 \right) \quad (6.6)$$

Direct Derivation

This variance estimator can also be derived directly by estimating unbiasedly the two terms in the variance expression in (6.4) namely,

$$\frac{1}{n} \sum_{i=1}^N \frac{Y_i^2}{P_i} \quad \text{and} \quad \frac{Y^2}{n} \quad \text{by} \quad \frac{1}{n^2} \sum_{i=1}^n \frac{y_i^2}{p_i^2} \quad \text{and} \quad \frac{1}{n^2(n-1)} \sum_{i=1}^n \sum_{j \neq i}^N \frac{y_i}{p_i} \frac{y_j}{p_j}$$

respectively

6.4 EFFICIENCY OF PPS SAMPLING

As has been pointed out earlier, pps sampling is expected to be more efficient than srs whenever the size measure x is approximately proportional to the study variable y , that is, whenever x and y are linearly related and the line of regression passes through the origin

6.4a SUB-UNITS APPROACH

The circumstances under which pps sampling is expected to be more efficient than srs can be understood by introducing the concept of *sub-units*. Suppose the i -th unit U_i having the size X_i is considered as made up of X_i sub-units, each having the same value (Y_i/X_i) , that is, the j -th sub-unit of the i -th unit is taken as having the value $Z_{ij} = (Y_i/X_i)$, $j = 1, 2, \dots, X_i$. Then it can be seen that selection of one sub-unit with srs from the population of $X \left(= \sum_{i=1}^N X_i \right)$ sub-units and considering the unit to which it belongs as selected amounts to selecting that unit with pps, for the probability of selecting any unit is proportional to the number of sub-units in it. The equal probability estimator based on a selected sub-unit is the same as the pps estimator.

Let U_{ij} be the selected sub-unit. An unbiased estimator of the population total

$$Z = \sum_{i=1}^N \sum_{j=1}^{X_i} Z_{ij} = \sum_{i=1}^N \sum_{j=1}^{X_i} \frac{Y_i}{X_i} = Y$$

based on one sub-unit selected with srs is given by

$$\hat{Y} = Z_{ij} X = \left(\frac{Y_i}{X_i} \right) X = \frac{Y_i}{P_i}, \quad \left(P_i = \frac{X_i}{N} \right),$$

which is the same as the pps estimator. The sampling variance of \hat{Y} is given by

$$V(\hat{Y}) = X \sum_{i=1}^N \sum_{j=1}^{X_i} (Z_{ij} - \bar{Z})^2, \quad \bar{Z} = \frac{Z}{X} = \frac{Y}{X},$$

which is simply $X^2 \sigma_z^2$, where σ_z^2 is the variance of Z_{ij} for the population of sub-units. Substituting Y_i/X_i for Z_{ij} , we get

$$V(\hat{Y}) = X \sum_{i=1}^N \sum_{j=1}^{X_i} \left(\frac{Y_i}{X_i} - \frac{Y}{X} \right)^2 = \sum_{i=1}^N \left(\frac{Y_i}{P_i} - Y \right)^2 P_i,$$

which is the same as (6.2). Thus we see that pps sampling would be more efficient than srs if $N^2 \sigma_y^2 > X^2 \sigma_z^2$, where σ_y^2 is the variance

of the variable y for the population of units. Dividing both sides of the above inequality by Y^2 , we get

$$\sigma_y^2/\bar{Y}^2 > \sigma_z^2/\bar{Z}^2, \quad (6.7)$$

for $\bar{Y} = NY$ and $\bar{Z} = Y/X$. Hence, the condition for pps to be more efficient than srs is that the relative variance of $z (= y/x)$ for the population of sub units should be less than that of y for the population of units.

It may be noted that the concept of sub units has helped in considering pps sampling as a particular case of srs. Proceeding as before, it can be shown that the selection of a sample of n units with pps with replacement is the same as the selection of n sub units from X sub units with srs with replacement and treating the n units to which they belong as selected. The estimator, variance and variance estimator in pps sampling could be derived as a particular case of the corresponding expressions for srswr with the help of the concept of sub units. The use of the concept of sub units in pps sampling has been discussed by Sethi (1962).

6.4b LINEAR RELATIONSHIP

The conditions under which pps sampling is more efficient than srs, when the variable y is linearly related with the size measure x , are of practical interest as it is usually possible in practice to find auxiliary variables linearly related with the study variable. Suppose the relationship between y and x is of the form

$$Y_i = \alpha + \beta X_i + \gamma_i \quad i = 1, 2, \dots, N, \quad (6.8)$$

where α and β are constants and γ_i is a random variable with its conditional expected value and variance for a given X_i as

$$E(\gamma_i | X_i) = 0 \quad \text{and} \quad V(\gamma_i | X_i) = \gamma X_i^2, \quad (6.9)$$

where γ and g are constants. Noting that the variances of the estimators of Y based on pps sampling and srs with replacement are respectively given by

$$V_{pps} = \frac{1}{n} \left(\bar{X} \sum_{i=1}^N \frac{Y_i^2}{X_i} - \bar{Y}^2 \right) \quad \text{and} \quad V_{srs} = \frac{1}{n} \left(N \sum_{i=1}^N Y_i^2 - \bar{Y}^2 \right),$$

we get the difference in the variances as

$$D = V_{srs} - V_{pps} = \frac{N}{n} \sum_{i=1}^N Y_i^2 \left(1 - \frac{\bar{X}}{X_i} \right).$$

Substituting the value given in (6.8) for Y_i in the expression for D and taking its expected value using (6.9), we get after simplification

$$E(D) = \frac{N^2}{n} \left[\alpha^2 \left(1 - \frac{\bar{X}}{N} \sum_{i=1}^N \frac{1}{X_i} \right) + \beta^2 \sigma_x^2 + \gamma \operatorname{Cov}(X_i^{g-1}, X_i) \right],$$

where σ_x^2 is the variance of x . Hence, pps sampling will be more efficient than srs, if

$$\gamma \operatorname{Cov}(X_i^{g-1}, X_i) > \alpha^2 \left(\frac{\bar{X}}{N} \sum_{i=1}^N \frac{1}{X_i} - 1 \right) - \beta^2 \sigma_x^2. \quad \dots \quad (6.10)$$

From the condition (6.10), it is clear that even if y and x are perfectly linearly related, that is, $V(\gamma_i | X_i) = \gamma X_i^g = 0$, pps sampling is not necessarily more efficient than srs, for the condition becomes

$$\alpha^2 \left(\frac{\bar{X}}{N} \sum_{i=1}^N \frac{1}{X_i} - 1 \right) - \beta^2 \sigma_x^2 < 0, \quad \dots \quad (6.11)$$

which may not be satisfied always. Thus, linearity of regression is not a sufficient condition for pps sampling to be better than srs. If the line of regression passes through the origin, that is, if $\alpha = 0$, the condition (6.10) becomes

$$\gamma \operatorname{Cov}(X_i^{g-1}, X_i) > -\beta^2 \sigma_x^2, \quad \dots \quad (6.12)$$

which is obviously satisfied if $g > 1$, since $\gamma \geq 0$ and $\operatorname{Cov}(X_i^{g-1}, X_i) > 0$. Empirical studies conducted by different authors have shown that the value of g is likely to lie between 1 and 2 in practice. The above results are based on the work of Des Raj (1958) and the behaviour of the efficiency of pps sampling has also been considered by Zarkovich (1960).

6 5 COST ASPECT

The comparison of the efficiencies of pps sampling and srs for a fixed sample size, considered in Section 6 4, would be realistic only if the cost of survey is approximately proportional to the number of units n in the sample. However, in practice, the cost of survey is expected to depend on both n and the total size of the sample units. For instance, in estimating the total number of persons and other demographic characteristics in a region based on a sample of villages or urban blocks, the cost of survey consisting of cost of journey to sample points and collection and tabulation of data for all the persons in the sample would depend on the number of sample villages and on the number of persons in the sample.

The sampling efficiency of pps compared to srs for a given n is given by

$$E_s = \frac{V_{srs}}{V_{pps}} = \left(N \sum_{i=1}^N Y_i^2 - Y^2 \right) / \left(X \sum_{i=1}^N \frac{Y_i^2}{X_i} - Y^2 \right) \quad (6 13)$$

The form of the cost function can be taken as

$$C = C_0 + nC_1 + E \left(\sum_{i=1}^n x_i \right) C_2 \quad (6 14)$$

where C is the expected total cost. C_0 is the overhead cost, nC_1 is the cost of journey and other stages of work depending mainly on the number of sample units and the last term is the cost of data collection and other items of work depending mainly on the size of the sample units. $E\left(\sum_{i=1}^n x_i\right)$ in the case of srs and pps is respectively given by

$$n\bar{X} \text{ and } n \sum_{i=1}^N X_i P_i = n \sum_{i=1}^N \frac{X_i^2}{\bar{X}} \quad (6 15)$$

Hence for a given expected cost C , the sample sizes that can be used in srs and pps sampling are given by

$$n_1 = \frac{C - C_0}{C_1 + \bar{X} C_2} \text{ and } n_2 = \frac{(C - C_0)}{C_1 + \left(\sum_{i=1}^N X_i^2 / \bar{X} \right) C_2}$$

and

$$\frac{n_2}{n_1} = \frac{C_1 + \bar{X}C_2}{C_1 + \left(\sum_{i=1}^N X_i^2/X \right) C_2}, \quad \dots \quad (6.16)$$

which shows that $n_2 \leq n_1$. If $C_1 > 0$ and $C_2 = 0$, $n_1 = n_2$ and that when $C_1 = 0$ and $C_2 > 0$, $(n_2/n_1) = N\bar{X}^2 / \sum_{i=1}^N X_i^2 = 1/(1+C_x^2)$, where C_x is the coefficient of variation of x . Thus the cost-efficiency of pps sampling compared to srs is given by

$$E_c = \frac{V_{srs}/n_1}{V_{pps}/n_2} = \frac{n_2}{n_1} E_s, \quad \dots \quad (6.17)$$

which reduces to just E_s if $C_1 > 0$ and $C_2 = 0$ and which becomes $E_s/(1+C_x^2)$, when $C_1 = 0$ and $C_2 > 0$.

The cost-efficiency can also be measured in terms of relative expected costs for a fixed sampling variance. If V_0 is the fixed sampling variance, then the sample sizes required to attain V_0 in srs and pps sampling are respectively given by $n_1 = V_{srs}/V_0$ and $n_2 = V_{pps}/V_0$. Substituting these values in the cost function (6.14), we get the expected costs after some simplification as

$$C_{srs}^* = C_0 + \frac{V_{srs}}{V_0} \{ C_1 + \bar{X}C_2 \} \quad \dots \quad (6.18)$$

and

$$C_{pps}^* = C_0 + \frac{V_{pps}}{V_0} \left\{ C_1 + \left(\sum_{i=1}^N X_i^2/X \right) C_2 \right\}. \quad \dots \quad (6.19)$$

It can be easily verified that $C_{pps}^* < C_{srs}^*$, if $E_s > 1$, when $C_1 > 0$ and $C_2 = 0$, and if $E_s > (1+C_x^2)$ when $C_1 = 0$ and $C_2 > 0$.

6 6 EMPIRICAL STUDIES

To study empirically the efficiency of pps sampling compared to srs and the effectiveness of using a suitable size measure the village data given in Annexure 4 1 of Chapter 4 (p 127) have been used taking the geographical area (g) and the 1951 census population (p) as measures of size for estimating the total cultivated area, the 1961 census population number of cultivators and number of workers in household industry In Table 6 2, the relative variances of estimates of different characteristics in sampling one unit with pps and srs are presented together with the values of the efficiencies E_s and E_c of pps sampling compared to srs, when the cost is assumed to depend only on (i) number of units in the sample, and (ii) total size of the units in the sample

TABLE 6 2 EFFICIENCY OF PROBABILITY PROPORTIONAL TO SIZE SAMPLING COMPARED TO SIMPLE RANDOM SAMPLING

sr no	characteristic	relative variance per sample unit			efficiency (%)			
		srs	PPG		E_s	PPP		E_c
			(4)	(5)		(6)	(7)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	cultivated area	0.3244	0.1130	0.3422	287	206	95	70
2	number of persons	0.3554	1.1827	0.0233	30	21	1525	1110
3	number of cultivators	0.4493	0.4818	0.1319	93	67	341	250
4	number of workers in household industry	1.8001	8.5693	0.7682	21	15	234	172
		(1.01)	(0.96)	(0.62)	(106)	(76)	(162)	(119)

ppg : probability proportional to geographical area ppp : probability proportional to 1951 census population

(The figures in brackets for item 4 are obtained after excluding the village with serial number 100 which has a very high value for this characteristic)

Columns (6) and (8) of Table 6 2 show that there is definite gain over srs in using pps sampling with g as the size for estimating the cultivated area (a), and p as size for estimating the other characteristics relating to the year 1961, though the amount of gain varies from

characteristic to characteristic, being the maximum for the 1961 census number of persons (p'). It may be noted that the presence of one village (serial number 100) with a large figure for number of workers at household industry (w) and a small value for g makes pps sampling very inefficient for estimating total of w when g is taken as the size measure. From columns (7) and (9), it is seen that even when the cost is assumed to depend only on the total size of the units in the sample, the efficiency of ppp sampling is substantial except for estimating the cultivated area.

A study of the relationship between g and a and that between p and p' , shown in Figures 6.1 and 6.2, brings out clearly that in these cases the study variable and the size measure are linearly related and that the line of regression passes through the origin, thereby verifying empirically the conditions stated earlier for pps sampling to be efficient.

In Table 6.3 are given the relative variances of pps sampling and srs in estimating total of a and total of p' with those of systematic sampling for different arrangements. From this table, it may be noted that pps sampling is generally more efficient than srs and systematic sampling with both the arrangements for estimating the total of p' , whereas it is not consistently so for estimating the total of a , though in both the cases it is more efficient than srs.

An interesting example illustrating the possible effectiveness of a suitable function of an auxiliary characteristic instead of its value itself as the size measure for sample selection is provided by the sampling of branches for estimation of total number of fruits on a tree (Jessen, 1955; Pearce and Holland, 1957). This technique, termed *branch sampling*, consists in selecting one or more branches of a tree with srs or with pps, size being a suitable measure of size of the branches such as the square of girth or circumference. The particular example considered here relates to an orange tree with two main branches and five secondary branches. The data on number of fruits, estimates of fruit count and their relative variances in sampling one branch with pps, size being powers of the branch-girth are

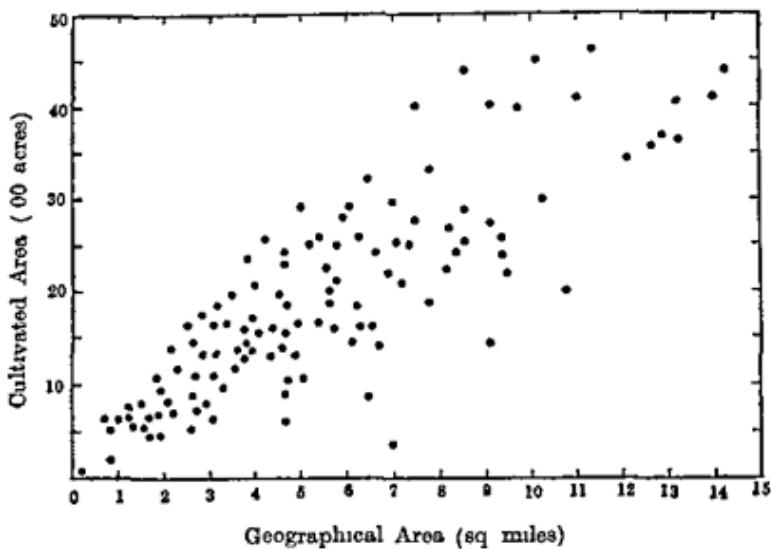


Figure 6.1 Relationship between cultivated area and geographical area for the villages in a tehsil

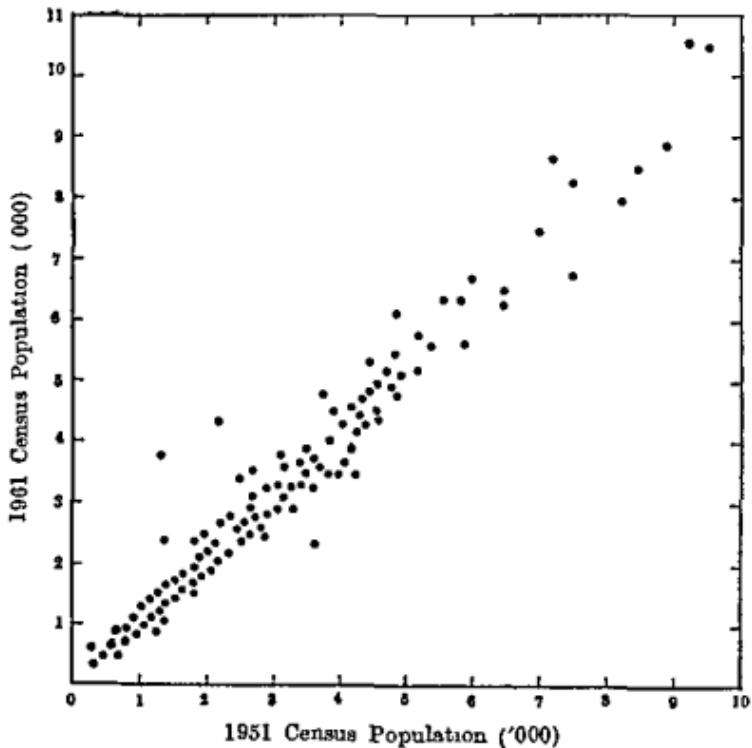


Figure 6.2 Relationship between 1951 and 1961 census populations for the villages in a tehsil.

TABLE 6.3. EFFICIENCY OF PROBABILITY PROPORTIONAL
TO SIZE SAMPLING COMPARED TO OTHER
SAMPLING SCHEMES.

sample size	srs	relative variance for the sampling scheme			
		systematic sampling*		pps sampling ⁺	
		(i)	(ii)	(i)	(ii)
(1)	(2)	(3)	(4)	(5)	(6)
<i>characteristic : cultivated area</i>					
2	0.1622	0.2026	0.1249	0.0565	0.0786
4	0.0811	0.0797	0.0329	0.0282	0.0393
8	0.0406	0.0336	0.0139	0.0141	0.0196
16	0.0203	0.0173	0.0074	0.0070	0.0098
32	0.0102	0.0009	0.0018	0.0035	0.0049
64	0.0051	0.0001	0.0012	0.0018	0.0024
<i>characteristic : 1961 census population</i>					
2	0.1777	0.2385	0.1303	0.0116	0.0158
4	0.0888	0.1263	0.0409	0.0058	0.0079
8	0.0444	0.0680	0.0121	0.0029	0.0040
16	0.0222	0.0194	0.0033	0.0014	0.0020
32	0.0111	0.0013	0.0005	0.0007	0.0010
64	0.0056	0.0007	0.0003	0.0004	0.0005

* arrangement (i) as in frame, (ii) in increasing order of (a) geographical area for estimating cultivated area and (b) 1951 census population for estimating total number of persons.

+ size is taken as geographical area in the case of cultivated area and as 1951 census population in the case of total number of persons; cost is taken as dependent on (i) number of sample units and (ii) total size of units in the sample.

given in Table 6.4. This table shows that pps sampling with size as the square of the branch-girth is considerably more efficient than srs and that by taking cube and fourth powers of the girth as size, pps sampling becomes even more efficient.

TABLE 6.4 EFFECTIVENESS OF POWERS OF THE GIRTH (g) AS SIZE,
FOR SAMPLING BRANCHES OF A TREE IN ESTIMATING
FRUIT COUNT

branch		no of fruits	estimate of fruit count by branch for			
main	subsidiary		srs	ppg ²	ppg ³	ppg ⁴
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	476	2380	1696	1625	1587
1	2	162	810	903	1117	1373
2	1	85	425	874	1164	1635
2	2	441	2205	1701	1427	1235
2	3	215	1075	1171	1194	1243
relative variance			0.317	0.068	0.020	0.014

Source Jessen R J (1955) Determining the fruit count on a tree by randomized branch sampling *Biometrics* 11 99-109 Pearce S C and Holland D A (1957) Randomized branch sampling for estimating fruit number *Biometrics* 13 127-130

The results of the empirical studies given in this section are to be taken as only indicative of the effectiveness of the use of pps sampling since they are based on rather limited data. In actual practice, extensive empirical studies may have to be carried out to assess the extent of gain in the use of pps sampling with different possible size measures.

6.7 GAIN DUE TO PPS SAMPLING

It is of interest to note that it is possible to estimate unbiasedly, the gain due to pps sampling as compared to srs from the pps sample itself. As an unbiased estimator of the variance of the pps estimator is already given in (6.6) we have to obtain an unbiased variance estimator for the srs estimator for purposes of comparison. The variance of the srs estimator is given by

$$V(\hat{Y}_{srs}) = \frac{N^2\sigma^2}{n} = \frac{1}{n} \left(N \sum_{i=1}^N Y_i^2 - \bar{Y}^2 \right)$$

and an unbiased estimator of this on the basis of a pps sample can be obtained by noting that unbiased estimators of the terms $\sum_{i=1}^N Y_i^2$ and Y^2 are respectively given by

$$\frac{1}{n} \sum_{i=1}^n \frac{y_i^2}{p_i} \text{ and } (\hat{Y}_{pps})^2 - v_{pps}(\hat{Y}_{pps}), \quad \hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}.$$

Thus, an unbiased estimator of $V(\hat{Y}_{srs})$ is given by

$$v_{pps}(\hat{Y}_{srs}) = \frac{1}{n^2} \left(N \sum_{i=1}^n \frac{y_i^2}{p_i} - n \hat{Y}_{pps}^2 \right) + \frac{1}{n} v_{pps}(\hat{Y}_{pps}). \quad \dots \quad (6.20)$$

Comparing this variance estimator with (6.6) the gain G due to pps sampling as compared to srs can be estimated as

$$G = v_{pps}(\hat{Y}_{srs}) - v_{pps}(\hat{Y}_{pps}) = \frac{1}{n^2} \sum_{i=1}^n \frac{y_i^2}{p_i} \left(N - \frac{1}{p_i} \right). \quad \dots \quad (6.21)$$

Estimators Based on SRS

Similarly, it is also possible to estimate the gain in efficiency by estimating unbiasedly $V(\hat{Y}_{pps})$ on the basis of a sample selected with srswr. Since

$$V(\hat{Y}_{pps}) = \frac{1}{n} \left(\sum_{i=1}^N \frac{Y_i^2}{P_i} - Y^2 \right)$$

and unbiased estimators of $\sum_{i=1}^N (Y_i^2/P_i)$ and Y^2 based on n units selected with srswr are given by

$$\frac{N}{n} \sum_{i=1}^n \frac{y_i^2}{p_i} \text{ and } (\hat{Y}_{srs})^2 - v_{srs}(\hat{Y}_{srs}),$$

we get

$$v_{srs}(\hat{Y}_{pps}) = \frac{N}{n^2} \left[\sum_{i=1}^n \frac{y_i^2}{p_i} - N(n\bar{y}^2 - s^2) \right], \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \quad (6.22)$$

6.8 AREA SAMPLING (A SPECIAL CASE)

In *area sampling*, the total area is divided into small well-defined, identifiable sub-areas, not necessarily of equal size, and a sample of such sub-areas (termed *area-units*) is selected with equal or unequal probability. This method of sampling is used when the small sub-areas are themselves units of observation or when a complete list of

units of observation is not available, but it is possible to associate each unit with one and only one of the sub areas. If the study variable is associated with the geographical area (g) of the area unit, then it would be desirable to select the area units with pps, size being g .

An example of area sampling is provided by sampling of n plots (parcels of land) with ppswr, size being g for estimating the area under a particular type of utilization (y). Suppose (G_i, P_i) are respectively the geographical area and the proportion of area under the specified type of utilization for the i th plot, $i = 1, 2, \dots, N$, and $\{g_i, p_i\}$, $i = 1, 2, \dots, n$, be the corresponding values for the sampled plots. In such a case, an unbiased estimator of Y is given by

$$\hat{Y} = \frac{G}{n} \sum_{i=1}^n g_i p_i, \quad \left(G = \sum_{i=1}^N G_i \right) \quad (6.23)$$

This is obtained by putting $x_i = g_i$ and $y_i = g_i p_i$ in the estimator $\frac{X}{n} \sum_{i=1}^n (y_i/x_i)$. The variance of \hat{Y} and an unbiased variance estimator are

$$V(\hat{Y}) = \frac{1}{n} \left(G \sum_{i=1}^N G_i P_i^2 - G^2 P^2 \right) = \frac{G^2}{n} \left(PQ - \sum_{i=1}^N \frac{G_i}{G} P_i Q_i \right), \quad (6.24)$$

where P is the overall proportion of area under that type of utilization, $Q_i = 1 - P_i$ and $Q = 1 - P$, ($i = 1, 2, \dots, N$), and

$$v(\hat{Y}) = \frac{G^2}{n(n-1)} \left(\sum_{i=1}^n p_i^2 - np^2 \right) = \frac{G^2}{n-1} \left(\bar{p} \bar{q} - \frac{1}{n} \sum_{i=1}^n p_i q_i \right), \quad (6.25)$$

where $p = \frac{1}{n} \sum_{i=1}^n p_i$, $\bar{q} = 1 - \bar{p}$ and $q_i = 1 - p_i$, $i = 1, 2, \dots, n$. When the plots are small, it is likely that the proportion of area under a particular type of utilization for each plot is either 1 or 0. In this case, $V(\hat{Y})$ and $v(\hat{Y})$ given in (6.24) and (6.25) reduce to

$$V(\hat{Y}) = G^2 P Q / n \quad (6.26)$$

and

$$v(\hat{Y}) = G^2 \bar{p} \bar{q} / (n-1), \quad (6.27)$$

since $P_i Q_i$ and $p_i q_i$ are zero for all i .

It may be noted that P here stands for the proportion of area under a type of utilization and should not be confused with the symbol P used for probability of selection or with the proportion of units having a specified characteristic. In this case, the probability of selecting the i -th plot is G_i/G .

In the case specified above, the variance of \hat{Y}_{srs} , the srs estimator of Y , is given by

$$V(\hat{Y}_{srs}) = \frac{1}{n} \left(N \sum_{i=1}^N G_i^2 P_i^2 - G^2 P^2 \right) \quad \dots \quad (6.28)$$

and if it is assumed $P_i = 1$ or 0 for all i , then $V(\hat{Y}_{srs})$ after some simplification becomes

$$\begin{aligned} V(\hat{Y}_{srs}) &= \frac{1}{n} \left[G^2 PQ + N \sum_{i=1}^N G_i^2 P_i - G^2 P \right] \\ &= V(\hat{Y}_{pps}) + \frac{N^2}{n} \text{Cov}(G_i P_i, G_i). \end{aligned} \quad \dots \quad (6.29)$$

Thus, we see that in this special case pps sampling is more efficient than srs if the correlation between g and y is positive.

6.9 AN ALTERNATIVE ESTIMATOR

As in the case of srswr, there is an alternative unbiased estimator in pps sampling also based only on the distinct units, which is more efficient than the pps estimator (6.3) (Pathak, 1962). But the alternative estimator is more difficult to calculate and does not admit of a simple variance estimator. Further, the gain in efficiency is likely to be small unless the sampling fraction is large. These disadvantages make the estimator less useful in practice than the usual estimator based on all the sample units including repetitions.

As an illustration, let us consider sampling three units with ppswr. Suppose two of the three units selected turn out to be distinct and let the values of the units in the sample and their initial probabilities be (y_1, p_1) , (y_1, p_1) and (y_2, p_2) . The usual estimator in this case is

$$\hat{Y} = \frac{1}{3} \left(\frac{2y_1}{p_1} + \frac{y_2}{p_2} \right).$$

The alternative estimator based only on the two distinct units which is obtained by weighting the usual estimators corresponding to the two possible samples

$$(y_1, p_1) (y_1 p_1) (y_2 p_2) \text{ and } (y_1, p_1) (y_2 p_2), (y_2, p_2)$$

by their respective probabilities of selection is given by

$$\hat{Y} = \frac{1}{3} \left[\frac{y_1}{p_1} + \frac{y_2}{p_2} + \frac{y_1 + y_2}{p_1 + p_2} \right]$$

Proceeding similarly, it can be shown that if in a sample of n units drawn with ppswr ($n-1$) units are distinct, then the alternative estimator will be given by

$$\hat{Y}' = \frac{1}{n} \left[\sum_{i=1}^{n-1} \frac{y_i}{p_i} + \left\{ \sum_{i=1}^{n-1} y_i \Big/ \sum_{i=1}^{n-1} p_i \right\} \right] \quad (6.30)$$

6.10 SELECTION PROCEDURES

The procedure of sampling with pps essentially consists in associating with each unit a number of numbers equal to or exactly proportional to its size and selecting the unit corresponding to a number chosen at random from the totality of numbers so associated. In this section, some procedures of selecting units with pps are considered, when the sampling frame is available in the form of a list or a map.

6.10a CUMULATIVE TOTAL METHOD

Let the size of U_i be X_i , $i = 1, 2, \dots, N$. A straight forward application of the principle of associating with each unit a number of numbers equal to its size is to associate the numbers 1 to X_1 with the first unit, the numbers X_1+1 to X_1+X_2 with the second unit and so on. The total number of numbers so associated is X , the sum of the sizes of all the units in the population. Then a number R is chosen at random from 1 to X and the unit with which this number is associated is considered as selected. This procedure of selection is termed *cumulative total method*, since this method needs cumulation of the sizes.

The steps involved in using the cumulative total method are as follows :

- (i) cumulation of sizes of the units and writing down of the cumulative totals for the units ($T_i = X_1 + X_2 + \dots + X_{i-1} + X_i = T_{i-1} + X_i$, $i = 1, 2, \dots, N$);
- (ii) choosing a number (R) at random from 1 to $T_N (= \bar{X})$; and
- (iii) selection of U_i if $T_{i-1} < R \leq T_i$.

The probability $P(U_i)$ of selecting the i -th unit is given by

$$P(U_i) = \frac{T_i - T_{i-1}}{T_N} = \frac{X_i}{\bar{X}},$$

which shows that the required probability of selection is achieved. For selecting a sample of n units with ppswr, the above operation is to be repeated n times.

An Example

An illustration of this procedure is given in Table 6.5 by considering selection of one factory from a population of 10 factories with pps, size being number of workers. If the random number chosen from 1 to 7120 is 3598, then the 6th factory, with which this number is associated, gets selected.

TABLE 6.5. SELECTING ONE FACTORY WITH PROBABILITY PROPORTIONAL TO WORKERS USING CUMULATIVE TOTAL METHOD.

factory	no. of workers	numbers associated	cumulative total	selected number
(1)	(2)	(3)	(4)	(5)
1	58	1 — 58	58	
2	908	59 — 966	966	
3	418	967 — 1384	1384	
4	442	1385 — 1826	1826	
5	615	1827 — 2441	2441	
6	1972	2442 — 4413	4413	3598
7	613	4414 — 5026	5026	
8	734	5027 — 5760	5760	
9	514	5761 — 6274	6274	
10	846	6275 — 7120	7120	

In this procedure, writing of the numbers associated with each unit as shown in column (3) of Table 6.5 is tedious, and it is unnecessary, since the cumulative totals given in column (4) are sufficient to indicate the numbers associated with each unit.

The main disadvantage of this method is that it involves cumulation of the sizes and writing down of the cumulative totals $\{T_i\}$, ($i = 1, 2, \dots, N$), which is time consuming and costly when N is large. For instance, if this method is used for selecting a sample of factories with probability proportional to the number of workers from the population of about 45 000 factories in India or for selecting a pps sample of farms or fields with area as the size from a large number of such units, the selection operation becomes prohibitively costly. A procedure which avoids the need for calculating cumulative totals for each unit, is considered in the next sub section. Of course, the work of cumulation is simple when the population is small.

6 10b LAHIRI'S METHOD

Lahiri (1951) suggested a method of pps selection which does not require cumulation of the sizes at all. This method consists in associating with each unit the same number of numbers, out of which only the number of numbers equal to its size are considered as effective numbers. That is, if X_i is the size of the i th unit, M numbers are associated with it, out of which X_i are considered to be effective, M being equal to or greater than the maximum of the sizes. Thus the total number of numbers associated is NM . If a number selected at random from 1 to NM is one of the effective numbers associated with a particular unit, that unit is considered selected and if it is one of the ineffective numbers, the procedure is repeated till one of the effective numbers is selected. One of the NM numbers can be selected at random by selecting one number at random from 1 to N and another from 1 to M .

Steps in Selection

Lahiri's method consists of the following steps

- (i) selection of a number at random from 1 to N (say, i),
- (ii) selection of another number at random from 1 to M (say, R), where M is the maximum of the sizes of the N units or some convenient number greater than the maximum size,

(iii) selection of U_i if $R \leq X_i$; and

(iv) rejection of U_i and repetition of the above process if $R > X_i$.
For selecting a sample of n units with ppswr, the above procedure is repeated till n units are selected.

To find the probability of selecting U_i in the first effective draw by this procedure, it is to be noted that this unit may be selected in the first draw or at some subsequent draw preceded by ineffective draws. The probability of selecting U_i in the first draw is $P_1(U_i) = (1/N)(X_i/M)$. Noting that a draw would be ineffective if one of the ineffective numbers is selected, the probability of rejecting a draw is

$$P(r) = \frac{1}{NM} \sum_{i=1}^N (M - X_i) = 1 - \frac{\bar{X}}{M}.$$

Hence, the probability of selecting U_i in the second draw, the first being ineffective, is $P(r)(1/N)(X_i/M)$. Proceeding similarly, we find that the probability of selecting U_i in the first *effective* draw is

$$\begin{aligned} P(U_i) &= P_1(U_i) + P_2(U_i) + P_3(U_i) + \dots \\ &= \frac{1}{N} \frac{X_i}{M} + \left(1 - \frac{\bar{X}}{M}\right) \frac{1}{N} \frac{X_i}{M} + \left(1 - \frac{\bar{X}}{M}\right)^2 \frac{1}{N} \frac{X_i}{M} + \dots \\ &= \frac{1}{N} \frac{X_i}{M} \sum_{j=1}^{\infty} \left(1 - \frac{\bar{X}}{M}\right)^j. \end{aligned}$$

Since $0 < \left(1 - \frac{\bar{X}}{M}\right) < 1$, the infinite series is convergent and we get,

$$P(U_i) = \frac{1}{N} \frac{X_i}{M} \left\{1 - \left(1 - \frac{\bar{X}}{M}\right)\right\}^{-1} = \frac{X_i}{\bar{X}}.$$

An Example

This procedure is applied to the example of selecting a factory with pps considered earlier and the relevant particulars are given in Table 6.6. It may be assumed that an equal number of two coordinate numbers have been associated with each unit, the first unit getting the numbers (1,1) to (1.1972) of which the numbers (1,1) to (1.59) are effective and the others ineffective, and so on. One of these numbers is chosen

TABLE 6.6 SELECTION OF ONE FACTORY WITH PROBABILITY PROPORTIONAL TO WORKERS USING LAHIRI'S METHOD

factory (1)	no of workers (2)	numbers associated	
		effective (3)	ineffective (4)
1	58	1(1 — 58)	1(59 — 1972)
2	908	2(1 — 908)	2(909 — 1972)
3	418	3(1 — 418)	3(419 — 1972)
4	442	4(1 — 442)	4(443 — 1972)
5	615	5(1 — 615)	5(616 — 1972)
6	1972	6(1 — 1972)	—
7	613	7(1 — 613)	7(614 — 1972)
8	734	8(1 — 734)	8(735 — 1972)
9	514	9(1 — 514)	9(515 — 1972)
10	846	10(1 — 846)	10(847 — 1972)

at random by selecting one number at random from 1 to 10 and another from 1 to 1972. Suppose the combined random number is (4 358) the 4th unit is selected since this number is an effective number associated with that unit. This procedure may seem to be complicated but operationally it is quite simple and does not even require the explicit specification of the numbers given in columns (3) and (4) of Table 6.6. A diagrammatic representation of this method is given in Figure 6.3.

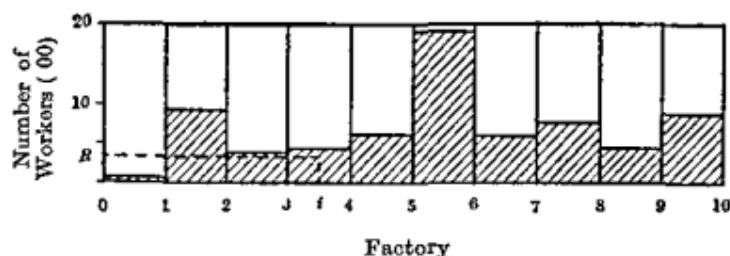


Figure 6.3 A diagrammatic representation of Lahiri's method of sampling with pps
(Shaded area denotes effective numbers)

The main advantage of this method is that it does not require writing down the cumulative totals for each unit. Further, for this method the sizes of all the units need not be known before hand. What is needed is only some number greater than the maximum size and the sizes only for the units selected by the choice of the first set

of random numbers from 1 to N . It is to be noted that though the sizes of all the individual units are not needed at the selection stage, the total of the sizes of all the units is required for estimating the population total or mean.

Split Method

A disadvantage of the above method is that some draws get rejected resulting in wastage of time and effort. In fact, the probability of rejection $P(r)$ is $1 - (\bar{X}/M)$ and the expected number of draws required to select one unit is M/\bar{X} , which will be large if M is much larger than \bar{X} . This difficulty can be, at least partially, overcome by using a device known as *split method*, which reduces $P(r)$. This device consists in splitting units having a large size into two or more *split units* and distributing their size over their split units. This procedure helps in reducing the gap between M and \bar{X} , as in this case M can be taken as the maximum of the sizes of the split units. Lahiri's method is then applied to select one unit from the totality of all split and unsplit units with the provision that whenever a split unit is chosen, the original unit to which it belongs is considered selected. It can be seen that the probability of selecting a unit remains proportional to its size as it is the sum of the probabilities of selecting its split units.

An Example

For instance, in the example considered earlier, $P(r)$ is 0.639, which is quite large. This is mainly due to the large unit having the size 1972. If this unit is split into 2 split units with 972 and 1000 as their respective sizes, then 1000 can be considered as M and $P(r)$ becomes 0.353, which is substantially less than what it was before.

Optimum Probabilities

In sampling units with pps, size being a function of the auxiliary variable x , it would be advantageous to use Lahiri's method. This can be illustrated by applying this method to selecting a unit with pps, size being \sqrt{x} . The procedure consists in selecting a number i at random from 1 to N and another random number R from 1 to M ($\geq \sqrt{\text{maximum } X_i}$), and selecting the unit U_i if $R^2 \leq X_i$. If $R^2 > X_i$, the draw is rejected and the operation is repeated till a unit is selected. It can be easily verified that the required probabilities for the units are achieved by this method. It is of interest to note that this method does not even require the computation of \sqrt{x} for

every unit unlike the cumulative total method which requires the calculation of \sqrt{x} and cumulative totals of \sqrt{x} for each unit. However the total of \sqrt{x} for all the units in the population is required for estimating population totals and for estimating ratios even this is not required.

6 10c CHOICE OF THE METHODS

Let C_1 and C_2 be the respective costs of cumulating the size and writing down the cumulative total per unit and let C_3 be the cost of selecting one random number. The cost involved in using the cumulative total method for selecting a sample of n units with ppswr from a population of N units is given by

$$C = N(C_1 + C_2) + nC_3 \quad (6.31)$$

The expected cost of using Lahiri's method is given by

$$C' = NC_1 + 2n(M/\bar{X})C_3 \quad (6.32)$$

since NC_1 is the cost of obtaining the total of the sizes required for estimation purposes and M/\bar{X} is the average number of draws required to select one unit and each draw involves selection of two random numbers. Comparing C and C' we find that Lahiri's method is to be used in preference to the cumulative total method when

$$\frac{M}{\bar{X}} < \frac{1}{2} \left(1 + \frac{C_2}{C_3} \frac{N}{n} \right) \quad (6.33)$$

The above inequality is likely to be satisfied in many cases, especially when M/\bar{X} is reduced by using split method since C_2 is likely to be greater than C_3 and N/n will generally be large.

However if the value of X is already available in the frame or if only ratios are to be estimated, the cost of using Lahiri's method reduces to

$$C' = 2n(M/\bar{X})C_3 \quad (6.34)$$

In this case Lahiri's method is to be preferred to the cumulative total method if

$$\frac{M}{\bar{X}} < \frac{1}{2} \left(1 + \frac{C_1 + C_2}{C_3} \frac{N}{n} \right), \quad (6.35)$$

which is more relaxed than (6.33). A working rule for the choice between the two methods is to use the cumulative total method when N is small and/or n is large and to use Lahiri's method when N is large and/or n is small compared to N .

6.10d AREA SAMPLING

If the frame consists of a map showing the boundaries of the units drawn to scale so that the areas occupied by the units on the map are proportional to their sizes, pps selection can be achieved by following a graphical procedure. The whole map is enclosed in a rectangle and two adjacent sides of the rectangle are taken as the X and Y axes. A pair of random numbers is taken as the coordinates on these axes to plot the random point on the map. The unit within which the point falls is selected. If the random point falls outside the mapped area, no unit is selected in that draw. This procedure is to be followed till the required number of units is selected in the sample. It may be noted that this method also does not envisage the use of cumulative totals. An example of the use of this procedure is provided by sampling of fields or plots in a region with pps, size being geographical area, with the help of a map showing the plots in a region (Figure 6.4).

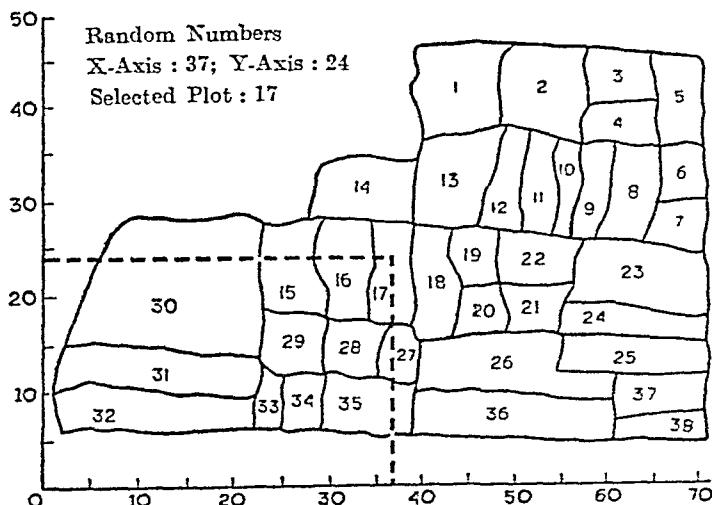


Figure 6.4. Selection of a field with probability proportional to area from a map.

6 10e SELECTION IN STAGES

Suppose one unit is to be selected with pps from a population consisting of K groups of units, the groups being formed on administrative or other considerations. For this purpose, it is usually convenient to select first one group with pps, size being the total of the sizes of units in it, and then to select one unit with pps from this group. This procedure avoids the need for giving continuous serial numbers and determining the cumulative totals for all the units in the population, and requires only a list of the groups with their sizes together with list of units with their sizes for each selected group. The probability of selecting the unit U_{ij} , the j th unit in the i th group, is given by the product of the probability of getting the i th group and the conditional probability of selecting the j th unit given the i th group, that is,

$$P(U_{ij}) = (X_i / \sum_{i=1}^K X_i) (X_{ij} / \sum_{j=1}^{N_i} X_{ij}) = X_{ij}/X,$$

since $X_i = \sum_{j=1}^{N_i} X_{ij}$ and $X = \sum_{i=1}^K X_i$, N_i being the number of units in the i th group. For selecting a sample of n units with ppswr, first n groups are selected with ppswr and from each selected group as many units as the number of times that group is selected are chosen with ppswr.

This method of selecting a unit in two stages may also be resorted to when a unit is to be selected with srs. In this case, first a group is to be selected with pps, size being the total number of units in the group, and then a unit is to be selected with srs from the selected group.

6 11 SAMPLING WITHOUT REPLACEMENT

It is generally observed that sampling without replacement provides a more efficient estimator than sampling with replacement, since the effective sample size is more in the former than in the latter. Considerable development has taken place in the field of sampling with varying probabilities without replacement since 1950. But most

of the suggested procedures, estimators and variance estimators are rather complicated and hence these are not commonly used in practice, especially in large-scale sample surveys with a small sampling fraction, since in such cases the efficiencies of sampling with and without replacement are not likely to differ much. However, it may be worthwhile to use these procedures of selection and estimation, if the sampling fraction is moderately large, as in that case the gain in efficiency in sampling without replacement is likely to be substantial.

6.11a PPS SAMPLING WITHOUT REPLACEMENT

Suppose n units are selected from N units with probability proportional to a size measure x at each draw without replacing the units selected in the previous draws (pps wor). The probabilities of selection at the first draw are given by $\{P_i\}$, $P_i = X_i/X$, $i = 1, 2, \dots, N$, those at the second draw when U_i has been selected in the first draw are $\{P_j/(1-P_i)\}$, $j \neq i$, those at the third draw when U_i and U_j ($j \neq i$) have been selected in the first two draws are $\{P_k/(1-P_i-P_j)\}$, $k \neq j \neq i$, and so on. With this set-up of sampling we shall consider some unbiased estimators of the population total and their variance estimators.

Horvitz and Thompson (1952) proposed a linear estimator of the form $\hat{Y}_{HT} = \sum_{i=1}^n a_i y_i$, where a_i is a constant depending on the i -th sample unit. The values of $\{A_i\}$, $i = 1, 2, \dots, N$, where A_i is the corresponding constant for the i -th unit in the population, can be found so as to make \hat{Y}_{HT} unbiased for Y . The expected value of \hat{Y}_{HT} is

$$E(\hat{Y}_{HT}) = \sum_s (\sum_{i=1}^n a_i y_i)_s P(s),$$

where Σ is the summation over all possible samples of n distinct units and $P(s)$ is the probability of getting the s -th sample. The above expression can be written as

$$E(\hat{Y}_{HT}) = \sum_{i=1}^N A_i Y_i \sum_{s \ni i} P(s) = \sum_{i=1}^N A_i Y_i \Pi_i,$$

where $\sum_{s \ni i} P(s)$, which is the sum of the probabilities of samples containing U_i , is the probability of inclusion of that unit in the sample, denoted by Π_i . Hence \hat{Y}_{HT} will be unbiased for Y if and only if $A_i = 1/\Pi_i$, that is, the estimator becomes,

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}, \quad . \quad (6.36)$$

where π_i is the probability of inclusion of the i th sample unit in the sample. It may be noted that the estimator (6.36) could be used for any sampling design, when the estimator is confined to only the distinct units in the sample (cf. Problem 6.11, p. 229).

The variance of \hat{Y}_{HT} is, by definition, given by

$$\begin{aligned} V(\hat{Y}_{HT}) &= E(\hat{Y}_{HT})^2 - Y^2 \\ &= \sum_i \left(\sum_{s \ni i} \frac{y_i^2}{\pi_i^2} + \sum_{i=1}^n \sum_{i \neq i} \frac{y_i}{\pi_i} \frac{y_i}{\pi_i} \right), \quad P(s) = Y^2 \\ &= \sum_{i=1}^N \frac{Y_i^2}{\prod_{i' \neq i}^n P(s)} + \sum_{i=1}^N \sum_{i \neq i'} \frac{Y_i}{\prod_i} \frac{Y_{i'}}{\prod_{i' \neq i}^n} \sum_{s \ni \{i, i'\}} P(s) - Y^2, \end{aligned}$$

since each unit occurs only once in a sample and any pair of units also occurs only once in a sample. Denoting $\sum_{s \ni \{i, i'\}} P(s)$, which is the probability of inclusion of U_i and $U_{i'}$ together in the sample, by $\Pi_{ii'}$ and expanding Y^2 , we get after some simplification

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \frac{1 - \Pi_i}{\Pi_i} Y_i^2 + \sum_{i=1}^N \sum_{i' \neq i}^N (\Pi_{ii'} - \Pi_i \Pi_{i'}) \frac{Y_i}{\Pi_i} \frac{Y_{i'}}{\Pi_{i'}}, \quad (6.37)$$

This can also be expressed as

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{i' > i}^N (\Pi_i \Pi_{i'} - \Pi_{ii'}) \left(\frac{Y_i}{\Pi_i} - \frac{Y_{i'}}{\Pi_{i'}} \right)^2, \quad (6.38)$$

since the coefficient of Y_i^2/Π_i^2 in (6.38) is

$$\prod_i \sum_{i' \neq i}^N \Pi_{i'} - \sum_{i' \neq i} \Pi_{ii'} = \Pi_i (n - \Pi_i) - (n - 1)\Pi_i = \Pi_i (1 - \Pi_i),$$

for $\sum_{i'=1}^N \Pi_{i'} = n$ and $\sum_{i' \neq i} \Pi_{ii'} = (n - 1)\Pi_i$, $i = 1, 2, \dots, N$.

Making use of the above two forms of $V(\hat{Y}_{HT})$, Horvitz and Thompson (1952), and Sen (1953) and Yates and Grundy (1953) have suggested the following unbiased variance estimators :

$$v_{HT}(\hat{Y}_{HT}) = \sum_{i=1}^n (1 - \pi_i) \left(\frac{y_i}{\pi_i} \right)^2 + \sum_{i=1}^n \sum_{i' \neq i}^n \frac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_{ii'}} \frac{y_i y_{i'}}{\pi_i \pi_{i'}} \quad \dots \quad (6.39)$$

and

$$v_{YGS}(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{i' > i}^n \frac{\pi_i \pi_{i'} - \pi_{ii'}}{\pi_{ii'}} \left(\frac{y_i}{\pi_i} - \frac{y_{i'}}{\pi_{i'}} \right)^2. \quad \dots \quad (6.40)$$

The main disadvantage of these variance estimators is that they take negative values for some samples, and this leads to difficulties in interpreting the reliability of the estimates. Two sampling schemes for which (6.40) is non-negative are given in (Problem 6.12, p. 229).

From the variance expression (6.38), it is clear that the variance does not reduce to zero even if the variable y and the size measure x are the same unless Π_i is made proportional to X_i . For instance, if a sample of two units is selected with pps wor, then

$$\Pi_i = P_i + \sum_{i' \neq i}^N P_{i'} \frac{P_i}{1 - P_{i'}} = P_i \left(1 + \sum_{i'=1}^N \frac{P_{i'}}{1 - P_{i'}} - \frac{P_i}{1 - P_i} \right),$$

which is not proportional to X_i . Narain (1951), Yates and Grundy (1953), Hanurav (1962) and Fellegi (1963) have given procedures of determining probabilities $\{P_i\}$ to be used at different draws such that Π_i becomes proportional to X_i . But these procedures are rather complicated and time-consuming. However, a fairly simple sampling scheme, which ensures the required probabilities of inclusion of the units in the sample, is discussed in Sub-section 6.11c.

Des Raj (1956a) suggested estimators based on the order in which the units are selected in the sample. Suppose $\{y_1, y_2, \dots, y_n\}$ and $\{p_1, p_2, \dots, p_n\}$ are respectively the values of the sample units and their initial probabilities in the order of their selection. The proposed estimator is given by

$$\hat{Y}_D = \frac{1}{n} \sum_{i=1}^n t_i, \quad \dots \quad (6.41)$$

where t_i is an unbiased estimator of Y based on the units selected in the first i draws and is given by

$$t_i = y_1 + y_2 + \dots + y_{i-1} + \frac{y_i}{p_i} (1 - p_1 - p_2 - \dots - p_{i-1}).$$

The expected value of t_i is clearly Y , for the conditional expected value over the i -th draw is

$$E(t_i | y_1, y_2, \dots, y_{i-1}) = (y_1 + y_2 + \dots + y_{i-1}) + \Sigma' Y_i = Y,$$

where Σ' stands for summation over all population units excluding those selected in the first $(i-1)$ draws.

Since \hat{Y}_D is the mean of n unbiased estimators, it is also unbiased for Y . It is of interest to note that the estimators $\{t_i, t_{i'}\}$, ($i' \neq i$), are uncorrelated. This is of importance, since this property gives rise to an unbiased variance estimator of the form

$$v(\hat{Y}_D) = \frac{1}{n(n-1)} \sum_{i=1}^n (t_i - \bar{t})^2, \quad \bar{t} = \hat{Y}_D \quad \dots \quad (6.42)$$

This variance estimator is clearly non-negative unlike those in the previous case. If $n = 2$, \hat{Y}_D and $v(\hat{Y}_D)$ become

$$\hat{Y}_D = \frac{1}{2} \left\{ \frac{y_1}{p_1} (1 + p_1) + \frac{y_2}{p_2} (1 - p_1) \right\} \quad \dots \quad (6.43)$$

and

$$v(\hat{Y}_D) = \frac{1}{4} (t_1 - t_2)^2 = \frac{1}{4} (1 - p_1)^2 \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2. \quad \dots \quad (6.44)$$

Roy Choudhury (1956) has shown that $V(\hat{Y}_D)$ is less than the variance of the usual unbiased estimator in ppswr. In the case of $n = 2$, $V(\hat{Y}_D)$ can be expressed as

$$V(\hat{Y}_D) = \left(1 - \frac{1}{2} \sum_{i=1}^N P_i^2\right) \left\{ \frac{1}{2} \sum_{i=1}^N \left(\frac{Y_i}{P_i} - Y \right)^2 P_i \right\} - \frac{1}{4} \sum_{i=1}^N \left(\frac{Y_i}{P_i} - Y \right)^2 P_i^2. \quad \dots \quad (6.45)$$

Das (1951) has also suggested an unbiased estimator based on the orders of selection of units in sampling with pps wr. But the variance estimator in this case is not always non-negative.

Unordered Estimator

Murthy (1957) has shown that corresponding to any estimator based on the order of selection of the units, there exists a more efficient estimator which ignores the orders of selection of the sample units and that the *unordered estimator* can be obtained by weighting all the possible *ordered estimators* (arrived at by considering all possible orders of selection of the given sample) with their respective probabilities. For instance, if a sample of two units u_1 and u_2 is selected with pps wr, then the ordered estimators $\hat{Y}_D(12)$ and $\hat{Y}_D(21)$ corresponding to the two possible orders of selection (u_1, u_2) and (u_2, u_1) are

$$\hat{Y}_D(12) = \frac{1}{2} \left\{ (1+p_1) \frac{y_1}{p_1} + (1-p_1) \frac{y_2}{p_2} \right\}$$

and

$$\hat{Y}_D(21) = \frac{1}{2} \left\{ (1+p_2) \frac{y_2}{p_2} + (1-p_2) \frac{y_1}{p_1} \right\}.$$

and their respective probabilities are $P(12) = p_1 p_2 / (1-p_1)$ and $P(21) = p_1 p_2 / (1-p_2)$. The unordered estimator \hat{Y}_M is given by

$$\begin{aligned} \hat{Y}_M &= \frac{\hat{Y}_D(12)P(12) + \hat{Y}_D(21)P(21)}{P(12) + P(21)} \\ &= \frac{1}{2-p_1-p_2} \left[(1-p_2) \frac{y_1}{p_1} + (1-p_1) \frac{y_2}{p_2} \right]. \quad \dots \quad (6.46) \end{aligned}$$

An unbiased non-negative variance estimator of this estimator given by

$$v(\hat{Y}_M) = \frac{(1-p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2. \quad \dots \quad (6.47)$$

In sampling n units with pps wr the unordered estimator corresponding to \hat{Y}_D can be shown to be of the form

$$\hat{Y}_M = \frac{1}{P(s)} \sum_{i=1}^n y_i P(s|i), \quad \dots \quad (6.48)$$

where $P(s)$ is the probability of getting the s th unordered sample and $p(s|s)$ is the conditional probability of getting the s th sample given that the unit u_i has been selected in the first draw. The variance estimator is of the form

$$\begin{aligned} v(\hat{Y}_{AS}) &= \frac{1}{\{P(s)\}^2} \left[\sum_{i=1}^n P(s|s) \{P(s|s) - P(s)\} y_i^2 + \sum_{i=1}^n \sum_{i' \neq i} \{P(s|s)P(s|i') - P(s)P(s|s)\} y_i y_{i'} \right] \\ &= \frac{1}{\{P(s)\}^2} \sum_{i=1}^n \sum_{i' > i} \{P(s)P(s|s) - P(s|s)P(s|i')\} p_i p_{i'} \left(\frac{y_i - y_{i'}}{p_i - p_{i'}} \right)^2. \quad \dots \quad (6.49) \end{aligned}$$

where $P(s|s)$ is the conditional probability of getting the s th sample given that the units $u_i, u_{i'}$ have been selected in the first two draws.

6.11b RANDOM GROUP METHOD

J N K. Rao, Hartley and Cochran (1962) have suggested a method of pps wr sampling, which consists in dividing the population of N units into n groups at random with N_i units in the i th group, $i = 1, 2, \dots, n$, and selecting one unit with pps from each of the random groups. This procedure may be termed the *random group method*. Let (y_1, y_2, \dots, y_n) and (p_1, p_2, \dots, p_n) be respectively the values of the study variable and the sizes or initial probabilities for the sample units selected from the n random groups. Then an unbiased estimator of Y is given by

$$\hat{Y}_G = \sum_{i=1}^n P'_i \frac{y_i}{p_i}, \quad \dots \quad (6.50)$$

where $P'_i = \sum_{j=1}^{N_i} P_{ij}$, P_{ij} being the initial probability of the j -th unit in the i th group.

Noting that $V(\hat{Y}_G)$ is given by (cf Section 2.8 of Chapter 2, p 41)

$$V(\hat{Y}_G) = E_1 V_2(\hat{Y}_G) + V_1 E_2(\hat{Y}_G),$$

where E_1 and V_1 are the unconditional expected value and variance over the formation of random groups and E_2 and V_2 are the conditional expected value and variance over the sampling of units within the random groups, it can be shown that

$$V(\hat{Y}_G) = \frac{1}{N(N-1)} \left(\sum_{i=1}^n N_i^2 - N \right) \sum_{i=1}^n \left(\frac{Y_i}{P_i} - Y \right)^2 P_i \quad \dots \quad (6.51)$$

Obviously the variance in (6.51) will be minimum if $N_i = N/n$ for all i . Then we get

$$V(\hat{Y}_G) = \frac{N-n}{N-1} V'(\hat{Y}) = \left(1 - \frac{n-1}{N-1} \right) V'(\hat{Y}), \quad \dots \quad (6.52)$$

where $V'(\hat{Y})$ stands for the variance of the estimator in sampling n units with ppsswr. This shows that this procedure of sampling and estimation is more efficient than ppsswr. An unbiased estimator of $V(\hat{Y}_G)$ given in (6.51) is

$$v(\hat{Y}_G) = \frac{\sum_{i=1}^n N_i^2 - N}{N^2 - \sum_{i=1}^n N_i^2} \sum_{i=1}^n \left(\frac{y_i}{p_i} - \hat{Y}_G \right)^2 P'_i. \quad \dots \quad (6.53)$$

If $N_i = N/n$, $i = 1, 2, \dots, n$, $v(\hat{Y}_G)$ reduces to

$$v(\hat{Y}_G) = \frac{(N-n)}{N(n-1)} \sum_{i=1}^n \left(\frac{y_i}{p_i} - \hat{Y}_G \right)^2 P_i. \quad \dots \quad (6.54)$$

Durbin (1953) had suggested a procedure similar to the random group method with the essential difference that in his procedure the groups are not formed at random.

6.11c PPS SYSTEMATIC SAMPLING

A simple selection procedure, which gives rise to specified probabilities for inclusion of units in a sample, consists in associating with each unit a number of numbers equal to its size as in the cumulative total method and selecting the units corresponding to a sample of numbers drawn systematically from all the numbers associated with the units. That is, in sampling n units with this procedure, the cumulative totals $\{T_i\}$, $i = 1, 2, \dots, N$, are determined and the units corresponding to the numbers,

$$\{r+jk\}, \quad j = 0, 1, 2, \dots, (n-1),$$

are selected, where $k = T/n = X/n$ and r is a random number from 1 to k . This procedure is known as *pps systematic sampling*. This method was suggested by W. G. Madow (1949) and later considered by Grundy (1954). It is being used in the Indian National Sample Survey and some other surveys for sampling units. The implications of using this method in preference to ppswr sampling have been discussed by Murthy and Sethi (1959) and Des Raj (1964). In contrast to this sampling scheme, systematic sampling with equal probability, discussed in Chapter 5, may be termed *simple systematic sampling*.

In pps systematic sampling, the unit U_i gets included in the sample, if $T_{i-1} < r+jk \leq T_i$ for some value of $j (= 0, 1, 2, \dots, (n-1))$. Since the random number, which determines the sample, is selected from 1 to k and since X_i of the numbers are favourable for inclusion of the i -th unit in a sample, the probability Π_i of inclusion of U_i is $X_i/k = nX_i/X$, provided $k > X_i$. This method can be applied even if $k < X_i$ for some units as is shown later in this sub-section. If X/n

is not an integer, the sampling interval k can be taken as the integer nearest to X/n and in this case the actual sample size differs from sample to sample and from the specified sample size at most by one unit if the remainder obtained on dividing X by n is less than k , which is likely to be the case in most of the situations. However, this difficulty can be overcome by selecting the sample in a circular fashion after choosing the random start from 1 to X instead of from 1 to k . The sampling scheme may be termed *pps linear systematic sampling* (pps lss) when the random start is taken from 1 to k and *pps circular systematic sampling* (pps css) when the random start is taken from 1 to X and sampling is done in a circular manner. However, if X/n is an integer, these two procedures will be equivalent. For the sake of simplicity X/n is assumed to be an integer in this section.

An Illustration

This technique can be illustrated by applying it to sampling of 2 factories from the population of 10 factories given in Table 6.5. The total size X in this case is 7120 and hence the sampling interval becomes 3560 ($= 7120/2$). Suppose the number chosen at random from 1 to 3560 is 2142, then the two units having the numbers 2142 and 5702 ($= 2142 + 3560$) associated with them, namely the 5th and the 8th units, are selected to form the sample.

An unbiased estimator of the population total Y is given by

$$\begin{aligned}\hat{Y}_{ps} &= \sum_{i=1}^n \frac{y_i}{\pi_i} = k \sum_{i=1}^n \frac{y_i}{x_i} \\ &= \frac{X}{n} \sum_{i=1}^n \frac{y_i}{x_i} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}, \quad \dots \quad (6.55)\end{aligned}$$

since $\pi_i = nx_i/X = np_i$. Thus we see that the estimator in this case has exactly the same form as the ppswr estimator. Pps lss is likely to be of considerable use in practice, since both the selection procedure and the estimator are simple in this case. The labour involved in obtaining cumulative totals can be avoided partly by continuing the cumulation till the points giving rise to the selection of sample units are reached without having to write down the cumulative totals for each unit.

Alternative Derivation

This estimator can be shown to be unbiased for Y by rewriting it in the form

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^N r_i \frac{Y_i}{P_i},$$

where r_i is the number of times the i -th unit occurs in the sample and it takes the value 0 if U_i does not occur in the sample and noting that r_i is a random variable taking the values $[nP_i]$ and $[nP_i+1]$ with probabilities $[nP_i+1]-nP_i$ and $nP_i-[nP_i]$ respectively. We find that $E(r_i) = nP_i$ and

$$E(\hat{Y}) = \frac{1}{n} \sum_{i=1}^N E(r_i) \frac{Y_i}{P_i} = \bar{Y}.$$

It may be noted that r_i is usually 1 or 0 if $k > X_i$ and that $r_i \geq 1$ if $k \leq X_i$. Thus we see that the estimator (6.55) is unbiased even if $k < X_i$.

As in simple systematic sampling, the variance of \hat{Y} depends much on the arrangement of the units and its behaviour with increase in sample size is also rather irregular. The efficiency of this technique can be increased appreciably by adopting a suitable arrangement of the units. Since pps systematic sampling is equivalent to simple systematic sampling of sub-units, formed such that U_i has X_i sub-units, each with the value Y_i/X_i , the principle to be followed in arranging the units is that the nearby sub-units are as homogeneous as possible and this situation is obtained by arranging the units in ascending or descending order of Y_i/X_i and not in the order of Y_i or X_i . But in practice Y_i/X_i would not be available and the values of Y_i/X_i for a previous period or a related variable may have to be used for this purpose. The main disadvantage of this method as in simple systematic sampling is that it is not possible to get an unbiased variance estimator on the basis of a single sample. However, the variance can be estimated unbiasedly by selecting m sub-samples of n/m units each pps systematically with m random starts selected with or without replacement.

Random Arrangement

Hartley and J. N. K. Rao (1962) have considered this procedure when the units are arranged at random and derived approximate expressions for the variance and variance estimator of \hat{Y}_{HR} , which is the same as the estimator given in (6.55), assuming that the population is at least moderately large and that $nP_i < 1$ for all i . The

approximate expressions for the variance and the variance estimator derived by them are given by

$$V(\hat{Y}_{HR}) = \frac{1}{n} \sum_{t=1}^N \left(\frac{Y_t}{P_t} - Y \right)^2 P_t \{1 - (n-1)P_t\} \quad (6.56)$$

and

$$v(\hat{Y}_{HR}) = \frac{1}{n^2(n-1)} \sum_{i=1}^n \sum_{i' > i} \left\{ 1 - n(p_i + p_{i'}) + n \sum_{t=t''=1}^N P_{t''}^2 \right\} \left(\frac{y_i}{p_i} - \frac{y_{i'}}{p_{i'}} \right)^2. \quad (6.57)$$

From (6.56), we see that even when the units are arranged at random pps systematic sampling is more efficient than ppswr sampling. Hence, it may be inferred that if a suitable arrangement of the units is effected, this procedure is likely to be considerably better than ppswr sampling. The variance estimator in (6.57) may be used to give an upper bound, when pps systematic sampling is used with a suitable arrangement of units.

6.11d PROBABILITY PROPORTIONAL TO TOTAL SIZE

Lahiri (1951) proposed the selection of a sample of n units with *probability proportional to the total of the sizes of the sample units* (ppts), since it makes the estimator

$$\hat{Y}_L = \left(\sum_{i=1}^n y_i / \sum_{i=1}^n x_i \right) X = (\bar{y}/\bar{x})X \quad \dots \quad (6.58)$$

unbiased for Y . A simple procedure of selecting a sample with pppts has been suggested by Midzuno (1952) and Sen (1952). This procedure consists in selecting one unit with pps in the usual manner and then drawing a sample of $(n-1)$ units from the remaining $(N-1)$ units with srs wr. The probability of getting a particular unordered sample s is the sum of the probabilities of getting it with the sample unit u_i , ($i = 1, 2, \dots, n$), being selected in the first draw. That is,

$$P(s) = \sum_{i=1}^n \frac{x_i}{X} \frac{1}{\binom{N-1}{n-1}} = \frac{1}{\binom{N}{n}} \frac{\bar{x}}{\bar{X}}, \quad \dots \quad (6.59)$$

since the probability of selecting the unit u_i in the first draw is x_i/X and that of selecting the remaining $(n-1)$ units of the sample is $1/\binom{N-1}{n-1}$.

The estimator \hat{Y}_L is unbiased, for $E(\hat{Y}_L)$ is by definition

$$E(\hat{Y}_L) = \sum_{s=1}^{\binom{N}{n}} \frac{\bar{y}_s}{\bar{x}_s} X P(s) = \frac{N}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \bar{y}_s = Y,$$

since each population unit occurs in $\binom{N-1}{n-1}$ of the possible $\binom{N}{n}$ samples. An unbiased variance estimator for \hat{Y}_L is given by

$$v(\hat{Y}_L) = \hat{Y}_L^2 - \frac{N \bar{X}}{n \bar{x}} \left\{ \sum_{i=1}^n y_i^2 + \frac{N-1}{n-1} \sum_{i=1}^n \sum_{i' \neq i} y_i y_{i'} \right\} \dots \quad (6.60)$$

for $E(\hat{Y}_L^2) = V(\hat{Y}_L) + Y^2$ and the expected value of the second term on the right hand side of (6.60) is Y^2 , since any unit occurs in $\binom{N-1}{n-1}$ samples and any two units occur together in $\binom{N-2}{n-2}$ samples.

6.11e COMPARISON OF VARIOUS ESTIMATORS

We have considered a number of estimators in the previous few sections and since it is difficult to compare their efficiencies theoretically, the sampling variances of these estimators are compared empirically in sampling two units from three hypothetical populations of four units each, studied by Yates and Grundy (1953) as being more extreme than situations usually met with in practice. The sampling variances together with their efficiencies compared to ppswr sampling are shown in Table 6.7. From this table, it is clear that there is no estimator with uniformly minimum variance for all the three populations, though some of the estimators are consistently more efficient than the ppswr estimator. In this table, the sampling variance and the efficiency of a selection procedure discussed later in Section 6.13, which may be considered as an extension of pps sampling, are also given and in this case the estimator (\hat{Y}_{RC}) turns out to be more efficient than almost all the other estimators for

populations A and B, though its performance is poor for population C. The hypothetical populations considered are given below :

A	$x = 0.1, 0.2, 0.3, 0.4$	B	$x = 0.1, 0.2, 0.3, 0.4$	C	$x = 0.1, 0.2, 0.3, 0.4$
	$y = 0.5, 1.2, 2.1, 3.2$		$y = 0.8, 1.4, 1.8, 2.0$		$y = 0.2, 0.6, 0.9, 0.8$

TABLE 6.7 EFFICIENCIES OF DIFFERENT SAMPLING SCHEMES

sr no	sampling scheme	estimator	A		B		C	
			Var	Eff	Var	Eff	Var	Eff
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	pps with replacement	\hat{Y}_{pps}	0.500	100	0.500	100	0.125	100
2	pps without replacement	\hat{Y}_{HT}	0.823	61	0.057	877	0.059	212
3	pps without replacement	\hat{Y}_D	0.365	137	0.365	137	0.088	142
4	pps without replacement	\hat{Y}_M	0.312	160	0.312	160	0.070	179
5	random group method	\hat{Y}_G	0.333	150	0.333	150	0.083	151
6	pps systematic sampling (increasing order of x)	\hat{Y}_{ps}	0.300	167	0.300	167	0	—
7	pps systematic sampling (random arrangement)	\hat{Y}_{HR}	0.367	136	0.367	136	0.033	379
8	pps sampling	\hat{Y}_L	0.363	138	0.510	98	0.101	124
9	extension of pps sampling	\hat{Y}_{RC}	0.100	500	0.100	500	0.168	74

$$\text{Eff } (\hat{Y}) = \{V(\hat{Y}_{pps})/V(\hat{Y})\}100$$

6.12 INTEGRATION OF SURVEYS

If all the characteristics under study in a survey are not related to one single measure of size, then selection with probability proportional to this size will be efficient only for some and not for all of them. For instance, geographical area (g) may be considered as a suitable size for sample selection in a crop survey, whereas the size to be used in the case of a population survey may have to be the population (p) according to an earlier census. If two independent samples are drawn with pps, size being g for crop survey and p for

population enquiry, the total number of sample points to be surveyed for both the surveys would almost be double that of each survey with subsequent increase in cost. Hence, it would be desirable to have a sampling scheme, which maximizes the number of common and adjacent units while retaining the required probabilities of selection for the two surveys. Some selection schemes, which help in integrating surveys, are discussed in this section.

Let x and x' be the size measures to be used in two surveys, which are to be integrated in the sense of having as many common units as possible. Suppose the sample with ppx is drawn for the first survey. The two surveys can be integrated by making the sample selection for the second survey with ppx' dependent on that for the first survey by adopting the following steps proposed by Keyfitz (1951), namely,

- (i) if $P'_i \geq P_i$, retain the selected unit U_i for the second survey ($P'_i = X_i/X$, $P_i = X'_i/X'$);
- (ii) if $P'_i < P_i$, retain the unit U_i with probability P'_i/P_i and reject the unit for the second survey with probability $1 - (P'_i/P_i)$; and
- (iii) if $P'_i < P_i$ and if the unit is rejected in (ii), select a unit from those units with $P'_j > P_j$ with probability proportional to $(P'_j - P_j)$.

This procedure of sampling can also be used when a sample selected in an earlier period with pps is to be revised on the basis of more recent information on the size so as to keep as many of the old sample units as possible in the new sample.

Proof: The fact that the required probabilities $\{P'_i\}$ for the units $\{U_i\}$, $i = 1, 2, \dots, N$, are achieved by this sampling scheme in respect of the second survey can be proved by noting that when $P'_i < P_i$, the unit U_i is selected in the second sample if it is selected in the first and it is retained with probability P'_i/P_i , that is,

$$P(U_i) = P_i(P'_i/P_i) = P'_i,$$

and that if $P'_i > P_i$, U_i is selected in the second sample if it is selected in the first sample or if some unit U_j , ($j \neq i$), with $P'_j < P_j$ is selected in the first sample and

it gets rejected with probability $1 - (P_j/P_i)$ and then U_i is selected with probability $(P_i - P_j)/\Sigma_1(P_j - P_i)$ Σ_1 standing for summation over units with $P_j > P_i$ that is in this case

$$P(U_i) = P_i + \Sigma_2 P_j \left(1 - \frac{P_j}{P_i} \right) \frac{P_i - P_j}{\Sigma_1(P_j - P_i)}$$

Σ_2 denot ng summation over units with $P_j < P_i$. Hence $P(U_i) = P_i$ for $\Sigma_1(P_j - P_i) = \Sigma_2(P_j - P_i)$. Further it can be seen that the probability of getting a common unit for the two surveys is $\sum_{i=1}^N \min(P_i, P_j)$ at each draw.

Lahiri (1954) suggested arrangement of the units in the sampling frame in a *serpentine* order so that any two geographically contiguous units occur next to each other in the sampling frame and then selection of units for the two surveys with the same random number chosen from 1 to X but using independent cumulative totals of the sizes x and x after adjusting them such that $\sum_{i=1}^N X_i = \sum_{i=1}^N X_j = X$

That is if $\{T_i\}$ and $\{T_j\}$, $i = 1, 2, \dots, N$, are the respective cumulative totals of sizes x and x and if R is a random number from 1 to X , then corresponding to this R , the units U_i and U_j will be selected for the two surveys respectively if $T_{i-1} < R \leq T_i$ and $T_{j-1} < R \leq T_j$. It is expected that U_i and U_j would be identical or adjacent units in a large number of cases because of the serpentine arrangement in the frame (cf Problem 6.24). In fact, Des Raj (1956b) has shown that the serpentine method minimizes the expected cost of travelling between units for both the surveys taken together under the assumption that the cost of travelling between the units U_i and U_j is proportional to $|i-j|$. The use of cumulative totals in this method can be avoided by using the method discussed in Sub section 6.10b with suitable modifications. Further, use of pps systematic sampling with a single random start after effecting a serpentine arrangement of the units is likely to result in a more efficient integration of the surveys.

The technique of sampling with ppts given in Sub section 6.11d can be applied to obtain a solution which amounts to almost complete integration of two or more surveys requiring the use of different sizes

at the selection stage. Roy Choudhury (1956) suggested selection of one unit with ppx (say U_i) for one survey and another with ppx' (say U_j) for the other survey and then selecting $(n-1)$ units from the remaining $(N-1)$ units with srs wr for both the surveys considering the places occupied by the two units selected earlier (U_i, U_j) as interchangeable at this subsequent selection. That is, for the subsequent selection of $(n-1)$ units from the remaining $(N-1)$ units, the units U_i and U_j are taken as one unit. If this unit is selected, U_j is considered as selected for the first survey and U_i is taken up for the second survey. The remaining $(n-2)$ units will also be common for the two surveys. If this unit is not selected, all the $(n-1)$ units will be common for the two surveys. It can be easily seen that in this method there can at most be one different unit between the two samples and that the probabilities of selection of the two samples are proportional to the total sizes $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x'_i$ respectively.

6.13 EXTENSION OF PPS SAMPLING

We have seen earlier that for pps sampling to be efficient, there should be a linear relationship between y and x and that the line of regression of y and x should pass through the origin. Roy Choudhury (1957) has extended pps sampling so as to make it efficient even in cases where the line of regression does not pass through the origin. The proposed method consists in dividing the units in the population into two groups according as $(X_i - \bar{X}) > 0$ or $(X_i - \bar{X}) < 0$ and selecting a sample of pairs of units from the population of all pairs of units, formed by pairing a unit of group 1 with a unit of group 2, with probability proportional to a suitable measure of size with replacement. If there are N_1 and N_2 units in the two groups, then there are $N_1 N_2$ possible pairs of units from which n pairs of units are selected with pps, size being $(P_{1i} + P_{2j})/N$, where $P_{1i} = (X_{1i} - \bar{X})/\sum_{i=1}^{N_1} (X_{1i} - \bar{X})$ and $P_{2j} = (X_{2j} - \bar{X})/\sum_{j=1}^{N_2} (X_{2j} - \bar{X})$, X_{1i} and X_{2j} denoting the values of size x for the i -th and the j -th units in the two groups respectively.

Let the n pairs, selected by the above procedure, have the values

$$\{y_{1k}, y_{2k}, x_{1k}, x_{2k}, p_{1k}, p_{2k}\}, \quad (k = 1, 2, \dots, n),$$

for the variables y , x and probability of selection defined above for the units in the pairs. Then an unbiased estimator of the population total is given by

$$\hat{Y}_{RC} = \frac{1}{n} \sum_{k=1}^n \frac{z_k}{p_k}, \quad (6.61)$$

where $z_k = p_{2k} y_{1k} + p_{1k} y_{2k}$ and $p_k = (p_{1k} + p_{2k})/N$, for $\{z_k/p_k\}$, $k = 1, 2, \dots, n$, are n independent estimates of Y and

$$\begin{aligned} E(z_k/p_k) &= \sum_{k=1}^{N_1 N_2} Z_k = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} (P_{2j} Y_{1i} + P_{1i} Y_{2j}) \\ &= \sum_{i=1}^{N_1} Y_{1i} + \sum_{j=1}^{N_2} Y_{2j} = Y \end{aligned}$$

An unbiased variance estimator in this case is given by

$$v(\hat{Y}_{RC}) = \frac{1}{n(n-1)} \left(\sum_{k=1}^n \frac{z_k}{p_k} - \hat{Y}_{RC} \right)^2 \quad (6.62)$$

It may be noted that if the regression of y on x is linear, the estimator \hat{Y}_{RC} reduces to Y even if the line of regression does not pass through the origin. Hence, this procedure is expected to be more efficient than ppswr especially when the line of regression does not pass through the origin. This procedure of sampling can also be adopted for use in the case of sampling without replacement. An empirical comparison of the efficiency of this estimator with those of other estimators has already been made in Table 6.7 and the results are briefly discussed in Sub section 6.11e.

REFERENCES

- DAS, A. C. (1951) : On two-phase sampling and sampling with varying probabilities; *Bull. Inter. Stat. Inst.*, 33, (2), 105-112.
- DES RAJ (1956a) : Some estimators in sampling with varying probabilities without replacement; *J. Amer. Stat. Assn.*, 51, 269-284.
- DES RAJ (1956b) : On the method of overlapping maps in sample surveys; *Sankhyā*, 17, 89-98.
- DES RAJ (1958) : On the relative accuracy of some sampling techniques; *J. Amer. Stat. Assn.*, 53, 98-101.
- DES RAJ (1964) : The use of systematic sampling with probability proportionate to size in a large-scale survey; *J. Amer. Stat. Assn.*, 59, 251-255.
- DURBIN, J. (1953) : Some results in sampling theory when units are selected with unequal probabilities; *J. Roy. Stat. Soc., (B)*, 15, 202-209.
- FELLEGI, I. P. (1963) : Sampling with varying probabilities without replacement : rotating and non-rotating samples; *J. Amer. Stat. Assn.*, 58, 183-201.
- GRUNDY, P. M. (1954) : A method of sampling with probability exactly proportional to size; *J. Roy. Stat. Soc., (B)*, 16, 236-238.
- HANSEN, M. H. and HUTCHINSON, W. N. (1943) : On the theory of sampling from finite populations; *Ann. Math. Stat.*, 14, 333-362.
- HANUBAV, T. V. (1962) : Some sampling schemes in probability sampling; *Sankhyā*, 24, (A), 421-428.
- HARTLEY, H. O. and RAO, J. N. K. (1962) : Sampling with unequal probability without replacement; *Ann. Math. Stat.*, 33, 350-374.
- HORVITZ, D. G. and THOMPSON, D. J. (1952) : A generalization of sampling without replacement from a finite universe; *J. Amer. Stat. Assn.*, 47, 663-685.
- JESSEN, R. J. (1955) : Determining the fruit count on a tree by randomized branch sampling; *Biometrics*, 11, 99-109.
- KEYFITZ, N. (1951) : Sampling with probability proportional to size—adjustment for changes in size; *J. Amer. Stat. Assn.*, 46, 105-109.
- LAHIRI, D. B. (1951) : A method of sample selection providing unbiased ratio estimates; *Bull. Inter. Stat. Inst.*, 33, (2), 133-140.
- LAHIRI, D. B. (1954) : Technical paper on some aspects of the development of the sample design; Indian National Sample Survey Report No. 5, reprinted in *Sankhyā*, 14, 264-316.
- MADOW, W. G. (1949) : On the theory of systematic sampling—II; *Ann. Math. Stat.*, 20, 333-354.
- MAHALANOBIS, P. C. (1938) : Statistical report on the experimental crop census, 1937; Indian Central Jute Committee.

- MIDZUNO, H (1952) On the theory of sampling with probability proportional to the sum of the sizes, *Ann Inst Stat Math*, 3, 99-107
- MURTHY, M N (1957) Ordered and unordered estimators in sampling without replacement, *Sankhya*, 18, 379-390
- MURTHY, M N and SETHI, V K (1959) Self weighting design at tabulation stage, Indian National Sample Survey Working Paper No 6, also *Sankhya*, 27, (B), (1965), 201-210
- NARAIN, R D (1951) On sampling without replacement with varying probabilities, *J Ind Soc Agr Stat*, 3, 169-174
- PATHAK, P K (1962) On sampling with unequal probabilities, *Sankhya*, 24, (A), 315-326
- PEARCE, S C and HOLLAND, D A (1957) Randomized branch sampling for estimating fruit number *Biometrics*, 13, 127-130
- RAO, J N K, HARTLEY, H O and COCHRAN, W G (1962) A simple procedure of unequal probability sampling without replacement, *J Roy Stat Soc*, (B), 24, 482-491
- ROY CHOUDHURY, D K (1956) Integration of several pps surveys, *Science and Culture*, 22, 119-120
- ROY CHOUDHURY, D K (1957) Unbiased sampling design using information provided by linear function of auxiliary variate, Chapter 5, Thesis submitted for Associateship of the Indian Statistical Institute
- SEN, A R (1952) Present status of probability sampling and its use in estimation of farm characteristics, (an abstract), *Econometrica*, 20, 130
- SEN, A R (1953) On the estimate of variance in sampling with varying probabilities, *J Ind Soc Agr Stat*, 5, 119-127
- SETHI, V K (1962) Some consequences of an interpretation of varying probability sampling, *Sankhya*, 24, (B), 215-222
- YATES, F and GRUNDY, P M (1953) Selection without replacement from within strata with probability proportional to size, *J Roy Stat Soc*, (B), 15, 253-261
- ZARKOVICH, S S (1960) On the efficiency of sampling with varying probabilities and the selection of units with replacement; *Metrika*, 3, 53-59

COMPLEMENTS AND PROBLEMS

6.1 Using the data given in Table 5.12 of Chapter 5 (p 178), calculate the relative efficiency of sampling villages with ppswr, size being geographical area, as compared to that of srswr for estimating the total area under paddy. How does the efficiency of ppswr compare with that of systematic sampling for a sample of 9 villages?

6.2 A sample of 10 villages was drawn from a tehsil with ppswr, size being the 1951 census population and the relevant data are given in Table 6.8

TABLE 6.8. 1951 CENSUS POPULATION (x) AND CULTIVATED AREA (y) IN ACRES FOR 10 SAMPLE VILLAGES.

village	x	y	village	x	y
(1)	(2)	(3)	(1)	(2)	(3)
1	5511	4824	6	7357	5506
2	865	924	7	5131	4051
3	2535	1918	8	4054	4060
4	3523	3013	9	1146	809
5	8368	7678	10	1165	1013

total population of the tehsil in 1951 = 415149.

(i) Estimate the total cultivated area \bar{Y} and its rse.

(ii) Obtain the sample size required to ensure an rse of 2%.

6.3 Data on number of workers, fixed capital and total output are given in Table 6.9 (p. 228) for a population of 80 factories situated in a region.

(i) Compare the efficiencies of sampling with probability proportional to (a) number of workers, and (b) fixed capital for estimating the total output.

(ii) Study the effect on sampling variance of using the number of workers as the size after rounding it off to the nearest multiple of 10 before selection.

6.4 There are 7 units in a population having the sizes 10, 20, 30, 40, 50, 60, and 70. A sample of 2 units is to be drawn with pps wr. Find the probabilities of inclusion in the sample for (a) each unit and (b) each pair of units, and verify that

$$\sum_{i=1}^N \Pi_i = 2 \text{ and } \sum_{i' \neq i} \Pi_{ii'} = \Pi_i.$$

6.5 For estimating the production of wheat in a region, a sample of n fields is drawn with ppswr, size being the area under wheat in them and yield per acre is determined for each of the sample fields. Suggest an unbiased estimator of the total wheat production in the region. Derive its sampling variance and also obtain an unbiased estimator of this variance.

6.6 Substantiate the statement : "Although two variables y and x may have perfect positive linear correlation, sampling for y with ppswr, size being x , may not necessarily be more efficient than srswr."

6.7 Show that in (6.18) and (6.19) $C_{ppw}^* < C_{srw}^*$, if $E_s > 1$ when $C_1 > 0$ and $C_2 = 0$, and if $E_s > (1 + C_x^2)$ when $C_1 = 0$ and $C_2 > 0$, E_s being the sampling efficiency of ppswr sampling compared to srswr (6.13).

TABLE 6.9 DATA ON NUMBER OF WORKERS (x_1), FIXED CAPITAL (x_2) AND OUTPUT (y) FOR 80 FACTORIES IN A REGION

sr no	x_1	x_2	y	sr no	x_1	x_2	y
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	51	106	1350	41	152	760	5124
2	51	162	1176	42	160	837	5236
3	52	235	1841	43	162	774	5113
4	52	325	2606	44	166	781	5230
5	53	244	2656	45	173	820	5330
6.	54	214	2546	46	185	672	4762
7.	57	310	2911	47	192	815	5420
8	60	370	3280	48	198	870	5562
9.	65	385	3425	49	211	885	5630
10.	67	390	3416	50	212	948	5582
11	68	367	3390	51.	253	770	5684
12	70	412	3395	52	285	1086	5790
13	71	407	3417	53	291	1073	5839
14	73	430	3290	54	314	1160	5920
15	74	435	3481	55	335	1320	6315
16	76	450	3520	56	352	1465	6510
17	78	463	3570	57	375	1530	6567
18	80	520	3740	58	387	1690	6719
19	81	470	3520	59	425	1670	6752
20	85	469	3601	60	443	1720	6660
21	87	510	3717	61	452	1780	6854
22	88	600	3750	62	466	1705	6760
23	92	584	3730	63	481	1650	6825
24	93	605	3767	64	495	1775	6940
25	97	590	3821	65	528	2370	7295
26	100	534	3886	66	544	1930	7070
27	107	618	3972	67	563	2290	7152
28	110	625	4065	68	585	2065	7186
29	113	630	4109	69	598	2170	7215
30	116	641	4216	70	641	2355	7288
31	119	720	4950	71	667	2500	7540
32	121	755	4302	72	705	2498	7416
33	125	663	4385	73	750	2500	7610
34	127	695	4426	74	775	2780	7894
35	127	680	4530	75	824	2750	8063
36	131	700	4689	76	870	2695	8180
37	134	750	5386	77	913	2880	8315
38	135	745	4961	78	951	2970	8576
39	139	732	4822	79	980	3000	8675
40	144	782	5097	80	1095	3485	9250

 x_2 and y in ('000) rupees

6.8 Noting that for the selection procedure given in Sub-section 6.10b, the proportion of effective draws is normally distributed when the number of draws is fairly large, indicate how you would proceed to determine the number of draws M_0 required to select n units with a confidence coefficient of $\alpha\%$. That is, find the value of M_0 such that $\text{Prob}(m < M_0) = \alpha\%$, where m is the actual number of draws resulting in n effective draws. If the probability of an effective draw is 60% and the sample size required is 100, how many draws should be made for ensuring the sample size with probability 99.7%?

(Haldar, A., *Sankhyā*, 23, (B), (1961), 329–330).

6.9 If in a sample of three units, drawn with ppswr, only two units are distinct, show that the estimators

$$(a) \frac{1}{3} \left(\frac{y_1}{p_1} + \frac{y_2}{p_2} + \frac{y_1+y_2}{p_1+p_2} \right) \text{ and } (b) \frac{y_1}{1-(1-p_1)^3} + \frac{y_2}{1-(1-p_2)^3}$$

are unbiased for the population total Y . If the size measure used for selection is approximately proportional to y , state, giving reasons, which of the two estimators you would prefer.

6.10 Suppose the population of N units is considered to be derived from a super-population with the following model :

$$E(Y_i|X_i) = aX_i, \quad V(Y_i|X_i) = \sigma_i^2 \quad \text{and} \quad \text{Cov}(Y_i, Y_{i'}|X_i, X_{i'}) = 0.$$

Derive the expected value of the variance of pps wr estimator in (6.37) and show that it reduces to

$$E\{V(\hat{Y}_{HT})\} = \sum_{i=1}^N \left(\frac{1}{\Pi_i} - 1 \right) \sigma_i^2,$$

when Π_i is made proportional to X_i , $i = 1, 2, \dots, N$.

(Godambe, V. P., *J. Roy. Stat. Soc.*, (B), 17, (1955), 269–278).

6.11 (i) Show that for any sampling design, the estimator $\hat{Y} = S'(y_i/\pi_i)$, where π_i is the probability of inclusion of the i -th distinct unit in a sample of n units and S' is the summation over all the distinct units in the sample, is unbiased for Y .

(ii) Derive the values of $\sum_{i=1}^N \Pi_i$ and $\sum_{i=1}^N \sum_{i' \neq i} \Pi_i \Pi_{i'}$.

(iii) Also derive the variance of \hat{Y} .

(iv) Obtain an unbiased estimator for $V(\hat{Y})$.

(Hanurav, T. V., *Sankhyā*, 24 (A), (1962), 429–436).

6.12 Derive the necessary and sufficient condition for the variance estimator in (6.40) to be non-negative for a sample of 2 units selected without replacement from a finite population of N units. Hence, show that the expression (6.40) is non-negative in sampling 2 units with pps wr. Also show that (6.40) is non-negative when one unit is selected with pps and then $(n-1)$ units are selected from the remaining $(N-1)$ units with srs wr.

(Sen, A. R., *J. Ind. Soc. Agr. Stat.*, 5, (1953), 119–127;

Vijayan, K., (1966), unpublished).

6.13 In sampling n units without replacement, derive the condition for the estimator $t/\lambda P(s)$ to be unbiased for Y , when t is the total of the n sample observations $P(s)$ is the probability of selecting the s th sample and λ is a constant. Hence, show that the estimator

$$\frac{(y_1+y_2)(1-p_1)(1-p_2)}{(N-1)p_1p_2(2-p_1-p_2)}$$

based on a sample of 2 units drawn with pps wr is unbiased for Y . Also prove that this estimator is more efficient than the *ordered estimator*

$$\frac{(y_1+y_2)(1-p_2)}{2(N-1)p_1p_2}$$

(Midzuno, H., *Ann Inst Stat Math*, 3, (1952), 99-107,

Murthy, M. N., *Sankhya*, 18, (1957), 379-390)

6.14 Show that the estimators t_i and t_j ($i \neq j$) in (6.41) are uncorrelated
 (Des Raj, *J Amer Stat Assn*, 51, (1956), 269-284)

6.15 Derive the results given in (6.48) and (6.49) relating to the unordered estimator

(Murthy M. N., *Sankhya*, 18, (1957), 379-390)

6.16 Show that the variance estimator $v(\hat{Y})$ in (6.49) is non negative

(Pathak P. K. and Shukla N. D., *Sankhya*, 28 (A) (1966), 41-46,
 Subrahmanya M. T., *Metrika*, 12, (1967))

6.17 Derive the results given in (6.51), (6.52), (6.53) and (6.54) regarding the sampling variance and the variance estimator in the random group method

(Rao, J. N. K., Hartley H. O. and Cochran W. G.,
J Roy Stat Soc (B), 24 (1962), 482-491)

6.18 Derive the results given in (6.56) and (6.57) relating to the variance and the variance estimator in pps systematic sampling with random arrangement of the units

(Hartley H. O. and Rao, J. N. K., *Ann Math Stat*, 33, (1962), 350-374)

6.19 Suppose the units in a population are grouped on the basis of the equality of their sizes and that each such group has at least n units. Then a sample of n units is chosen with ppsswr from the whole population and repeated units are replaced by units selected with srs wr from the respective groups. Suggest an unbiased estimator of the population total Y and derive its sampling variance. Also obtain an unbiased variance estimator. How does the sampling variance in this case compare with that of sampling with ppsswr?

(Stevens, W. L., *J Roy Stat Soc (B)*, 20, (1958), 393-397)

6.20 When ppswr sampling is continued till the $(r+1)$ -th draw, at which the $(n+1)$ -th distinct unit gets selected, show that the usual estimator

$$\hat{Y} = \frac{1}{r} \sum_{i=1}^N r_i \frac{Y_i}{P_i}$$

based on the first r draws producing n distinct units is unbiased for Y . Here r_i is the number of repetitions of the i -th unit in the population and $\sum_{i=1}^N r_i = r$, r_i being zero for units not selected in the sample. Obtain an unbiased variance estimator for \hat{Y} .

(Sampford, M. R., *Biometrika*, 49, (1962), 27-40).

6.21 Suppose in varying probability sampling the units are drawn at the n -th draw with the probabilities

$$P_i^{(n)} = \frac{P_i(K_n - K_{n-1})}{1 - K_{n-1}P_i}, \quad i = 1, 2, \dots, N,$$

where $K_0 = 0$, $K_1 = 1$ and

$$K_n = K_{n-1} + \sum_{i=1}^N \frac{P_i}{1 - K_{n-1}P_i}, \quad 2 \leq n \leq n',$$

n' being the minimum value of n for which $K_n > 1/\text{Max}\{P_i\}$. Show that this procedure results in the probability of inclusion of the i -th unit in a sample of n units (Π_i) being proportional to P_i , $i = 1, 2, \dots, N$, for $n \leq n'$.

(Hanurav, T. V., *Sankhyā*, 24, (A), (1962), 421-428).

6.22 Suppose two units are selected with probabilities $\{P_i\}$, $i = 1, 2, \dots, N$, with replacement and if the two units are distinct, the sample is retained; otherwise this sample is rejected and another sample of two units is selected with probabilities proportional to $\{P_i^2\}$ and with replacement. If the two units selected are distinct, this sample is retained; otherwise a further sample of two units is selected with probabilities proportional to $\{P_i^4\}$ and with replacement, and so on. Show that this procedure results in the probability of inclusion Π_i being proportional to P_i , $i = 1, 2, \dots, N$, when $P_N = P_{N-1}$, and that the estimator of the population total admits of a non-negative variance estimator.

(Hanurav, T. V., (1965), unpublished).

6.23 Suppose two units are selected without replacement from a population of N units with $\{P_i\}$, $i = 1, 2, \dots, N$, as the probabilities of selection in the first draw and with

$$P_{i'}^{(2)} = P_{i'} \left(\frac{1}{1-2P_i} + \frac{1}{1-2P_{i'}} \right) \left(1 + \sum_{i=1}^N \frac{P_i}{1-2P_i} \right)^{-1}, \quad i' \neq i,$$

as the conditional probabilities of selection at the second draw when U_i is selected in the first draw. Show that the probability of inclusion of U_i in the sample (Π_i) is $2P_i$ and derive the values of $\Pi_{ii'}$, $i' \neq i$. Also show that the unconditional probability of selecting the i -th unit at the second draw is P_i .

(Durbin, J., *Ann. Math. Stat.*, 36, (1965), 1327).

6.24 Suppose two surveys where the units are to be selected with two sets of initial probabilities $\{P_i\}$ and $\{P_j\}$ are to be integrated in the sense of ensuring a large number of common and adjacent units between the two samples. For this purpose a map showing the location of all the units is secured and the units are numbered in *serpentine* manner such that neighbouring units receive consecutive serial numbers. Taking two straight lines of equal length, suppose lengths proportional to $\{P_i\}$ are marked on one of them and those proportional to $\{P_j\}$ on the other line in the order of their serial numbers. After superimposing the two lines points are selected at random on the combined line and the units corresponding to the portions in which the points fall are taken up for the respective surveys.

Apply the above method to the population of 16 villages given in Table 6.10 so as to integrate the two samples to be selected with ppswrr sizes being area and population in the two cases. Find out the probabilities of selecting common villages and adjacent villages for the two surveys. A sketch map showing the boundaries of the villages is given in Figure 6.5 to facilitate the identification of neighbouring villages. By renumbering the villages in a different serpentine order, examine whether the probabilities of common and adjacent villages can be increased.

TABLE 6.10 DATA ON SIZE MEASURES PROPORTIONAL TO AREA (a) AND POPULATION (p) FOR 16 VILLAGES

village	a	p									
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	3	8	5	11	6	9	5	10	13	4	2
2	4	5	6	15	20	10	5	1	14	4	10
3	6	10	7	15	10	11	10	9	15	6	6
4	5	5	8	11	5	12	8	3	16	4	6

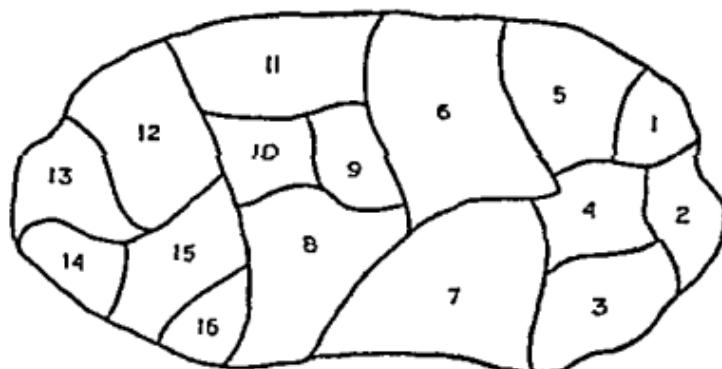


Figure 6.5 A sketch map showing village boundaries.

(Lahiri, D. B., *Sankhya*, 14, (1954), 264-316)

Stratified Sampling

7.1 NEED FOR STRATIFICATION

Stratified sampling consists in classifying the population units into a certain number of groups, called *strata*, and then selecting samples independently from each group or *stratum*. An appropriate estimator for the population as a whole is obtained by suitably combining the stratum-wise estimators of the characteristic under consideration. The division of the population into strata, termed *stratification*, is usually done in such a way as to reduce the variability of the strata estimators. This is generally achieved by forming the strata such that they are homogeneous within themselves with respect to a suitably chosen auxiliary variable termed *stratification variable*. Further, there is considerable flexibility in stratified sampling in the sense that the sampling and the estimation procedures may differ from stratum to stratum depending on the nature of supplementary information available. and that the demarcation of the strata boundaries and the allocation of the total sample size to the strata may be so done as to make the estimator most efficient from the points of view of sampling variability and cost. Some examples of stratification are provided by the division of the whole area of a region into coastal, plains and hilly areas in an area survey and by classification of retail stores in a city on the basis of the products sold or their total volume of sales in a trade survey.

Though the main advantage of using stratified sampling is the possible increase in efficiency per unit of cost in estimating the

population characteristics this method is likely to be useful in the following situations also

- (i) When estimates are required with specified margins of error not only for the population as a whole but also for certain groups of units then these groups forming *sub populations* or *domains of study* are themselves usually considered as strata
- (ii) If the sampling frame is available in the form of *sub frames* which may be for regions or for specified categories of units it may be operationally convenient and economical to treat the sub frames as strata for sample selection. Of course further stratification of the sub frames may be resorted to if found desirable
- (iii) The amount of supplementary information available, and hence the methods of selection and estimation to be used may differ from region to region or from one group of the population to another. In such cases it is very advantageous to treat these regions or groups as strata since this would enable the maximum possible utilization of the available supplementary information
- (iv) When a survey organization has field offices in several zones into which the country may have been divided for administrative purposes it might be desirable to treat the zones as strata so as to facilitate the organization of field work
- (v) Use of stratification is of considerable importance in case of skew populations (cf Sub section 2.2c of Chapter 2 p 29) since greater weightage may have to be given for the few extremely large units for reducing the sampling variability
- (vi) It may be noted that in Section 5.11 of Chapter 5 (p 172) the desirability of dividing the population into some parts (or strata) and selecting systematic samples from each part using the same sampling interval but with independently selected random starts was stressed for reducing the chance of getting bad samples and hence decreasing the sampling variability when there is some cyclical or periodic variation in the population

7.2 PRINCIPLE OF STRATIFICATION

Suppose a population of N units is divided into K strata. Let N_s be the number of units in the s -th stratum, and let Y_{si} be the value of the study variable for the i -th unit in the s -th stratum. The population mean \bar{Y} can be written as

$$\bar{Y} = \sum_{s=1}^K W_s \bar{Y}_s, \quad \left(\bar{Y}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} Y_{si}, \quad W_s = \frac{N_s}{N} \right). \quad \dots \quad (7.1)$$

An unbiased estimator of \bar{Y} can be obtained by estimating unbiasedly the stratum means $\{\bar{Y}_s\}$ on the basis of random samples drawn from each stratum with sampling schemes, which need not necessarily be the same for all the strata. Suppose \hat{Y}_s is an unbiased estimator of \bar{Y}_s , then an unbiased estimator of \bar{Y} is given by

$$\hat{Y} = \sum_{s=1}^K W_s \hat{Y}_s \quad \dots \quad (7.2)$$

and its sampling variance is

$$V(\hat{Y}) = \sum_{s=1}^K W_s^2 V(\hat{Y}_s), \quad \dots \quad (7.3)$$

since $\text{Cov}(\hat{Y}_s, \hat{Y}_{s'}) = 0$ for $s \neq s'$ due to the selection of samples independently from each stratum. This shows that in stratified sampling, the sampling variance depends only on the within-strata variation. Hence, for getting efficient estimators the strata should be so formed as to minimize the within-strata variation. Since the total variation, comprised of between- and within-strata variations, is a fixed quantity, minimization of within-strata variation amounts to maximization of the between-strata variation.

The estimator of the population total Y is given by $N\hat{Y}$, that is,

$$\hat{Y} = N \sum_{s=1}^K W_s \hat{Y}_s = \sum_{s=1}^K N_s \hat{Y}_s = \sum_{s=1}^K \hat{Y}_s.$$

The sampling variance of \hat{Y} is obtained by multiplying $V(\hat{Y})$ by N^2 , that is,

$$V(\hat{Y}) = N^2 \sum_{s=1}^K W_s^2 V(\hat{Y}_s) = \sum_{s=1}^K N_s^2 V(\hat{Y}_s) = \sum_{s=1}^K V(\hat{Y}_s)$$

Illustration

The utility of stratification in reducing the sampling variance may be illustrated by considering sampling of 2 units from the following hypothetical population of 5 units for estimating the population mean

unit	U_1	U_2	U_3	U_4	U_5
value	1	2	3	6	7

For this population the sampling variances of estimators of \bar{Y} based on a sample of size 2 selected with srswr srs wr and css with 2 as the sampling interval are respectively 2.68, 2.01 and 1.06. If we have the information that the first three units are homogeneous and are different from the other 2 units then two strata can be formed one stratum consisting of U_1, U_2, U_3 and the other stratum comprising U_4, U_5 . When one unit is selected from each of the two strata with srs $V(\hat{Y})$ given in (7.3) becomes 0.28 which is considerably less than the variances for the unstratified sampling schemes considered here.

In stratified sampling the following inter related points need careful consideration

- (i) choice of sampling design within strata
- (ii) choice of stratification variable
- (iii) allocation of sample size to strata,
- (iv) number of strata and
- (v) demarcation of strata

In fact, the best method of utilizing the technique of stratified sampling effectively consists in determining an optimum composite choice of the possible solutions to the five points mentioned above, since the solutions to these points are interdependent. The earlier developments in this field were mainly confined to questions relating to (iii), though some attention was also given to points (i) (ii) and (iv). Only in the recent past the points (iv) and (v) have been studied in greater detail.

7.3 DESIGN AND STRATIFICATION VARIABLE

One of the advantages of stratification is that it makes it possible to use different sampling designs in the different strata thereby enabling effective utilization of the available supplementary information in cases where the extent and nature of the available information vary between groups of units of the population. For instance, data on geographical area may be available for all the villages in a region but population figures may be available only for the villages in one part of the region and not for those in the other part of the region. In such a case, we may treat the two parts of the region as two sub-populations or strata for a population survey to enable the use of pps sampling with population as size in one stratum and with geographical area as the size in the other stratum. In the second stratum, srs may also be used instead of pps if found desirable. Once the sampling design to be used is decided, one may proceed to consider the possibility of stratifying the sub-populations further to reduce the sampling variability.

It may be noted that the available supplementary information may be used either for stratification purposes, or for selection of units, or for both. When the procedure of selecting units in the strata is specified, the problem of stratification may be considered to consist of determining the number of strata, allocation of sample size to the strata and demarcation of strata such that the sampling variance is minimized for a given cost or the cost is minimized for a specified precision for the estimator. The optimum solution to this problem would naturally depend on the values of the variable under consideration, termed *study variable* or *estimation variable*, which is usually not available. Hence, the solution to this problem has to be based, of necessity, on the data available for some suitable supplementary variable and on some knowledge or judgement regarding the relationship between the stratification and the estimation variables, in which case the solution cannot be optimum for the latter, but can at best be only near optimum for it.

Illustration

The fact that the choice of stratification variable would depend on the sampling design used within the strata can be illustrated by considering the two sampling designs srs and pps sampling.

If srs wr is used in each stratum an unbiased estimator of \bar{Y} is given by

$$\hat{\bar{Y}} = \sum_{s=1}^K \frac{W_s}{n_s} \sum_{t=1}^{n_s} y_{st}, \quad (7.4)$$

where y_{st} is the value of the t th sample unit in the s th stratum and its variance is

$$\begin{aligned} V(\hat{\bar{Y}}) &= \sum_{s=1}^K W_s^2 V(y_s) = \sum_{s=1}^K \frac{W_s^2 \sigma_s^2}{n_s} \\ &= \sum_{s=1}^K \frac{W_s^2}{n_s} \frac{1}{N_s} \sum_{t=1}^{N_s} (Y_{st} - \bar{Y}_s)^2 \\ &= \frac{1}{N^2} \sum_{s=1}^K \frac{1}{n_s} \sum_{t=1}^{N_s} \sum_{t' > t} (Y_{st} - Y_{st'})^2 \end{aligned} \quad (7.5)$$

Thus we see that to reduce $V(\hat{\bar{Y}})$ it is necessary to reduce the value of terms $(Y_{st} - Y_{st'})^2$, $s \neq t$. This could be achieved by grouping the units to form the strata such that the units belonging to the same stratum are as similar as possible with respect to a stratification variable x highly positively or negatively correlated with y , since in practice the values of y would not be available for stratification purposes.

If instead of srs wr, pps wr sampling is used in each stratum, an unbiased estimator of \bar{Y} is given by

$$\hat{\bar{Y}}' = \frac{1}{N} \sum_{s=1}^K \frac{X_s}{n_s} \sum_{t=1}^{n_s} \frac{y_{st}}{x_{st}}, \quad (7.6)$$

where x_{st} is the size measure and $X_s = \sum_{t=1}^{N_s} X_{st}$. Its variance is

$$\begin{aligned} V(\hat{\bar{Y}}') &= \frac{1}{N^2} \sum_{s=1}^K V(\hat{Y}'_s) = \frac{1}{N^2} \sum_{s=1}^K \frac{1}{n_s} \sum_{t=1}^{N_s} \left(X_s \frac{y_{st}}{x_{st}} - Y_s \right)^2 \frac{X_{st}'}{X_s} \\ &= \frac{1}{N^2} \sum_{s=1}^K \frac{1}{n_s} \sum_{t=1}^{N_s} \sum_{t' > t} \left(\frac{y_{st}}{x_{st}} - \frac{y_{st'}}{x_{st'}} \right)^2 X_{st} X_{st'} \end{aligned} \quad (7.7)$$

From (7.7) it is clear that to reduce $V(\hat{\bar{Y}}')$ it is necessary to group the units into strata such that those belonging to the same stratum are as similar or homogeneous to each other as possible with respect to the variable y/x . This can be achieved by grouping units similar to each other with respect to a supplementary variable positively or negatively correlated with y/x .

7.4 ALLOCATION OF SAMPLE SIZE

Intuitively it may be felt that the allocation of the sample size to the strata would depend on the stratum sizes and the within-strata variation. If the cost per unit is presumed to be constant for all the strata, the stratum or strata accounting for a substantial part of the variation should receive a larger allocation. On the other hand, if the contribution from each stratum to the sampling variance is almost constant, then the stratum or strata where the cost of survey per unit is large should get a smaller allocation.

7.4a VARIANCE AND COST FUNCTIONS

We have seen that the sampling variance in stratified sampling is of the form

$$V(\hat{Y}) = \sum_{s=1}^K W_s^2 V(\hat{Y}_s).$$

It may be mentioned that if units are selected with replacement in the strata, $V(\hat{Y}_s)$ in each stratum would be of the form V_s/n_s , where V_s is the variance of an estimator of \bar{Y}_s based on one sample unit and n_s is the sample size in the s -th stratum. In this case,

$$V(\hat{Y}) = \sum_{s=1}^K W_s^2 \frac{V_s}{n_s}. \quad \dots \quad (7.8)$$

For instance, in stratified srswr V_s is given by,

$$V_s = \sigma_s^2 = \frac{1}{N_s} \sum_{i=1}^{N_s} (\bar{Y}_{si} - \bar{Y}_s)^2, \quad \dots \quad (7.9)$$

and in stratified ppswr sampling it is

$$V_s = \frac{1}{N_s^2} \sum_{i=1}^{N_s} \left(\frac{\bar{Y}_{si}}{P_{si}} - \bar{Y}_s \right)^2 P_{si}, \quad \dots \quad (7.10)$$

where $P_{st} = X_{st}/X_s$, X_{st} and X_s are the sizes of the s th unit and the s th stratum. Even in the case of sampling without replacement in the strata, it may be possible for certain sampling designs and estimators to express the variance in the form

$$V(\hat{Y}) = \sum_{s=1}^K W_s^2 \frac{A_s}{n_s} + \sum_{s=1}^K W_s^2 B_s, \quad (7.11)$$

where A_s and B_s are population parameters independent of the sample size n_s . For example, in stratified srs w/o r $V(\hat{Y}_s)$ is given by

$$\begin{aligned} V(\hat{Y}_s) &= \frac{(N_s - n_s)}{(N_s - 1)} \frac{\sigma_s^2}{n_s} \\ &= \frac{1}{n_s} \left(\frac{N_s}{N_s - 1} \sigma_s^2 \right) + \left(\frac{-1}{N_s - 1} \sigma_s^2 \right), \end{aligned}$$

showing that the sampling variance is of the form given in (7.11).

The cost function in stratified sampling may be taken as

$$C = C_0 + \sum_{s=1}^K n_s C_s, \quad (7.12)$$

where C_0 is the overhead cost, which is a constant for certain broad ranges of the total sample size, and C_s is the average cost of surveying one unit in the s th stratum, which may depend on the nature and size of the units in the stratum. That is, C_s may be taken as $C_s = C_{s1} + a_s C_{s2}$, where C_{s1} is the cost per unit, which depends only on time for journey, contact, etc., and is independent of the size of the unit, and $a_s C_{s2}$ is the cost, which depends on the number of elements in the unit or on some other size measure affecting the cost of survey, a_s being the average value of the size of a selected unit. For example, in a population survey, where a sample of villages is selected with srs and all the persons in the sample villages are surveyed, then a_s is the average population per village and C_{s2} is the cost per person.

7.4b OPTIMUM ALLOCATION

The problem of optimum allocation consists in determining the number of units to be selected from the different strata with a view to minimizing (i) the sampling variance for a given cost, or (ii) the cost of the survey while ensuring a specified value for the sampling variance. That is, when the cost is fixed at C' , we have to find the values of $\{n_s\}$, which would minimize $V(\hat{Y})$ subject to the cost restriction (7.12). Alternatively, when the sampling variance to be achieved is specified as V' , then we have to find the values of n_s which would minimize the cost C given in (7.12) subject to the restriction that $V(\hat{Y}) = V'$.

Assuming that the variance can be expressed in the form given in (7.8), the optimum allocation for a fixed cost C' is obtained by equating to zero the partial derivatives of

$$\sum_{s=1}^K W_s^2 \frac{V_s}{n_s} - \lambda \left(C' - C_0 - \sum_{s=1}^K n_s C_s \right)$$

with respect to $\{n_s\}$, and solving for n_s , ($s = 1, 2, \dots, K$), and λ , the Lagrangian multiplier. That is,

$$\frac{1}{n_s^2} (W_s^2 V_s) = \lambda C_s, \quad \text{or} \quad n_s = \frac{W_s}{\sqrt{\lambda}} \sqrt{V_s / C_s}.$$

Solving for λ by substituting the values of $\{n_s\}$ in (7.12), we get the optimum allocation as

$$n_s = (C' - C_0) \frac{W_s \sqrt{V_s / C_s}}{\sum_{s=1}^K W_s \sqrt{V_s / C_s}}, \quad \dots \quad (7.13)$$

which shows that the allocation should be proportional to $W_s \sqrt{V_s / C_s}$. With this allocation, the sampling variance of the estimator becomes

$$V(\hat{Y}) = \frac{1}{C' - C_0} \left(\sum_{s=1}^K W_s \sqrt{V_s / C_s} \right)^2. \quad \dots \quad (7.14)$$

If the cost C_s is taken to be the same for all the strata, then the cost restriction amounts to fixation of the total sample size (n) and in that case the optimum allocation is proportional to $W_s \sqrt{V_s}$, which is basically the product of the number of units in the stratum and the standard error per one sample unit. In this case the sampling variance becomes $\frac{1}{n} \left(\sum_{s=1}^K W_s \sqrt{V_s} \right)^2$.

The solution for the problem of optimum allocation for a fixed sample size in stratified srs was suggested by Neyman (1934) and that for a general type of cost restriction was proposed by Mahalanobis (1944). Stuart (1954) has shown that the optimum allocation can be easily derived by using Cauchy's inequality.

Optimum Allocation for a Fixed Variance

If it is required to ensure a specified value V' for the variance of the estimator, then the optimum allocation is determined by equating to zero the partial derivatives of

$$C_0 + \sum_{s=1}^K n_s C_s + \lambda \left(\sum_{s=1}^K W_s^2 \frac{V_s}{n_s} - V' \right) \quad . \quad (7.15)$$

with respect to $\{n_s\}$ and solving for n_s , ($s = 1, 2, \dots, K$) and λ , the Lagrangian multiplier. That is,

$$n_s = \frac{1}{V'} \left(\sum_{s=1}^K W_s \sqrt{V_s C_s} \right) W_s \sqrt{V_s / C_s} \quad (7.16)$$

With this allocation, the cost of survey becomes

$$C = C_0 + \frac{1}{V'} \left(\sum_{s=1}^K W_s \sqrt{V_s C_s} \right)^2 \quad . \quad (7.17)$$

Here again we find that the optimum allocation is proportional to $W_s \sqrt{V_s / C_s}$. From (7.13) and (7.16), we find that the constant of proportionality for determining n_s depends on whether cost or variance is considered as fixed.

The main difficulty in having the optimum allocation is that it requires a knowledge of V_s , which information is usually either not available or difficult to obtain. But approximations to this allocation may be obtained by using estimates of V_s from a previous survey, or using the actual strata variances or suitable approximations to them.

for some variable related to the study variable. Sukhatme (1935) has shown that for effectively using the optimum allocation in stratified srs, estimates of the strata variances obtained in a previous survey or in a specially planned pilot survey based even on samples of moderate sample size would be adequate for increasing the precision of the estimator. Evans (1951) has also considered the problem of allocation based on estimates of strata variances obtained in an earlier survey. In this connection, it is useful to consider the possibility of conducting the main survey itself in a phased manner, utilizing the data collected in the first phase for ensuring better allocation in the second phase and so on (Mahalanobis, 1944). In the next few sub-sections we shall examine some simpler allocations based on certain approximations to the strata variances. The optimum allocation to some strata, as found by the above method, may be greater than the total number of units in them, in which case complete enumeration is resorted to in these strata and the remaining sample size is allocated to the rest of the strata on optimum or near optimum considerations.

7.4c PROPORTIONAL ALLOCATION

When no other information except $\{N_s\}$ is available, the allocation of a given sample size n may be done in proportion to N_s provided there is evidence to expect that the sampling variance in the smaller strata is less than that in the larger ones. In this allocation ✓

$$n_s = n W_s. \quad \dots \quad (7.18)$$

This allocation, known as *proportional allocation*, was originally proposed by Bowley (1926). This procedure of allocation is often resorted to in practice because of its simplicity. This allocation is likely to be nearly optimum for a fixed sample size when the strata are not very dissimilar among themselves with respect to the strata sampling variances, for in that case optimum allocation, that is, allocation proportional

to $W_s \sqrt{V_s}$, reduces to proportional allocation. This situation is likely to be obtained when strata are formed mainly for administrative or operational convenience without much emphasis on the question of reduction of sampling error. The sample size for a given cost C in the case of proportional allocation can be obtained by substituting nW_s for n_s in the cost function (7.12) and solving for n . It is given by

$$n = (C - C_0) / \sum_{s=1}^K W_s C_s \quad (7.19)$$

7.4d ALLOCATION PROPORTIONAL TO $W_s \bar{Y}_s$

When there is evidence to believe that the rse of \hat{Y} based on one sample unit does not vary considerably over strata it would be preferable to allocate the sample size in proportion to stratum total that is,

$$n_s = n (W_s \bar{Y}_s) / \bar{Y} \quad (7.20)$$

For, in that case the optimum allocation ($n_s \propto W_s \sqrt{V_s}$) reduces to the above allocation. This situation is again likely to be obtained when stratification is resorted to mainly for administrative or operational convenience. However in practice the allocation has to be in proportion to the strata totals of a suitably chosen supplementary variable, as the values of the estimation variable y would not be available. Mahalanobis (1952) proposed equalization of strata sizes and then having equal allocation as an approximation to optimum allocation. Hansen Hurwitz and W G Madow (1953) have demonstrated that this procedure is useful in situations where slight modifications of the strata boundaries do not disturb appreciably the strata sampling variances.

7.4e ALLOCATION PROPORTIONAL TO $W_s R_s$

In populations usually met with in practice, it is found that generally R_s , the range of the values of the variable, provides an approximation to the standard deviation. Hence, it may be useful to allocate the sample size to the strata in proportion to $W_s R_s$, that is,

$$n_s = n(W_s R_s) / \sum_{s=1}^K W_s R_s, \quad \dots \quad (7.21)$$

when the cost per unit is the same between strata and in proportion to $W_s R_s / \sqrt{C_s}$, that is

$$n_s = (C' - C_0) \frac{W_s R_s / \sqrt{C_s}}{\sum_{s=1}^K W_s R_s \sqrt{C_s}}, \quad \dots \quad (7.22)$$

when the per unit cost varies from stratum to stratum. It may be pointed out that R_s here stands for the range of the sampling distribution of \hat{Y} based on one sample unit. For instance, in stratified srs R_s happens to be the same as the range of the values of the units in the s -th stratum. This allocation is expected to be near optimum, since R_s is likely to be a good approximation to the standard error $\sqrt{V_s}$, provided the sampling distribution is not very skew. Since the values of the study variable are not available at the stage of allocation, it is necessary to use the values of a suitable supplementary variable in this case also. Ekman (1959) suggested formation of strata by equalizing the value of $W_s R_s$ and having equal allocation in the case of a stratified srs design.

7.5 STRATIFIED SRS WITH REPLACEMENT

Suppose n_s units are selected from the N_s units in the s th stratum ($s = 1, 2, \dots, K$) with srswr and let \bar{y}_s be the sample mean in the s th stratum. An unbiased estimator of \bar{Y} is given by

$$\hat{\bar{Y}}_{st} = \sum_{s=1}^K W_s \bar{y}_s \quad (7.23)$$

and its sampling variance is

$$V(\hat{\bar{Y}}_{st}) = \sum_{s=1}^K W_s^2 \frac{\sigma_s^2}{n_s}, \quad . \quad (7.24)$$

where the subscript st denotes stratified sampling. An unbiased estimator of $V(\hat{\bar{Y}}_{st})$ is given by

$$t(\hat{\bar{Y}}_{st}) = \sum_{s=1}^K W_s^2 \frac{s_s^2}{n_s}, \quad s_s^2 = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (y_{si} - \bar{y}_s)^2 \quad (7.25)$$

7.5a OPTIMUM ALLOCATION

Noting that in stratified srswr the sampling variance V_s of $\hat{\bar{Y}}_s$ based on one unit is σ_s^2 , we find that the value of n_s in optimum allocation for a fixed cost C is given by

$$n_s = (C - C_0) \frac{W_s \sigma_s / \sqrt{C_s}}{\sum_{s=1}^K W_s \sigma_s \sqrt{C_s}} \quad (7.26)$$

and that the optimum value of n_s for a fixed variance V is

$$n_s = \frac{1}{V} \left(\sum_{s=1}^K W_s \sigma_s \sqrt{C_s} \right) \left(W_s \sigma_s / \sqrt{C_s} \right), \quad (7.27)$$

If $C_s = C''$ is the same for all strata, then n_s in (7.26) becomes

$$n_s = n \frac{W_s \sigma_s}{\sum_{s=1}^K W_s \sigma_s}, \quad n = \frac{C' - C_0}{C''}. \quad \dots \quad (7.28)$$

The corresponding sampling variance $V_0(\hat{\bar{Y}}_{st})$ is given by

$$V_0(\hat{\bar{Y}}_{st}) = \frac{1}{n} \left(\sum_{s=1}^K W_s \sigma_s \right)^2. \quad \dots \quad (7.29)$$

The variance of the sample mean based on a sample of n units selected with srs wr from the whole population without stratification is

$$V(\hat{\bar{Y}}_{us}) = \frac{\sigma^2}{n}, \quad \sigma^2 = \frac{1}{N} \sum_{s=1}^K \sum_{i=1}^{N_s} (Y_{si} - \bar{Y})^2,$$

the subscript us standing for unstratified sampling. This variance can be rewritten as

$$V(\hat{\bar{Y}}_{us}) = \frac{1}{n} \sum_{s=1}^K W_s \sigma_s^2 + \frac{1}{n} \sum_{s=1}^K W_s (\bar{Y}_s - \bar{Y})^2. \quad \dots \quad (7.30)$$

Comparing (7.29) with (7.30), it can be shown that

$$V_0(\hat{\bar{Y}}_{st}) = V(\hat{\bar{Y}}_{us}) - \frac{1}{n} \sum_{s=1}^K W_s (\bar{Y}_s - \bar{Y})^2 - \frac{1}{n} \sum_{s=1}^K W_s (\sigma_s - \bar{\sigma})^2, \quad \dots \quad (7.31)$$

where $\bar{\sigma} = \sqrt{\frac{1}{K} \sum_{s=1}^K W_s \sigma_s}$. Hence, to achieve considerable gain over unstratified srs, the strata should be so formed as to maximize the variation between the strata means as well as between the strata standard deviations.

When the values or estimates of σ_s^2 are not readily available, allocations may be made proportional to the strata totals of a supplementary variable x . As mentioned earlier, this allocation would be near optimum when the coefficient of variation (σ_s/\bar{Y}_s) remains approximately the same over the strata. Alternatively, allocation proportional to $W_s R_s$, R_s being the range of x , is likely to be near optimum, since the range usually provides a fairly good idea of the standard deviation.

7.5b PROPORTIONAL ALLOCATION

In proportional allocation, where $n_s = nW_s$, $V(\hat{Y}_{st})$ becomes

$$V_p(\hat{Y}_{st}) = \frac{1}{n} \sum_{s=1}^K W_s \sigma_s^2, \quad \dots \quad (7.32)$$

where the subscript p denotes proportional allocation. Comparing (7.32) with (7.29) and (7.30), we get

$$V_p(\hat{Y}_{st}) = V_0(\hat{Y}_{st}) + \frac{1}{n} \sum_{s=1}^K W_s (\sigma_s - \bar{\sigma})^2, \quad \dots \quad (7.33)$$

and

$$V_p(\hat{Y}_{st}) = V(\hat{Y}_{us}) - \frac{1}{n} \sum_{s=1}^K W_s (\bar{Y}_s - \bar{Y})^2. \quad \dots \quad (7.34)$$

From (7.34) we find that the efficiency of proportional allocation can be substantially increased by forming strata in such a way that the strata means are as different as possible.

7.5c GAIN DUE TO STRATIFICATION

Comparing (7.30) and (7.24), we get the reduction in sampling variance due to stratification as

$$V(\hat{Y}_{us}) - V(\hat{Y}_{st}) = \frac{1}{n} \sum_{s=1}^K \left(1 - \frac{nW_s}{n_s} \right) W_s \sigma_s^2 + \frac{1}{n} \sum_{s=1}^K W_s (\bar{Y}_s - \bar{Y})^2. \quad \dots \quad (7.35)$$

From (7.35) it is clear that stratified srs is definitely more efficient than unstratified srs whenever n_s is taken proportional to (i) W_s or (ii) $W_s \sigma_s$, since the first term on the right hand side of (7.35) reduces to zero in case (i) and it becomes $\sum_{s=1}^K W_s (\sigma_s - \bar{\sigma})^2 > 0$ in case (ii). It is of interest to note that this term reduces to zero even if n_s is taken as proportional to $W_s \sigma_s^2$, which result is mainly of academic interest, since the same sampling variance is achieved with proportional allocation which is much simpler.

When the allocation deviates much from the above three allocations, there may be a few situations where the first term on the right hand side of (7.35) not only becomes negative but also exceeds the other term in absolute value thereby giving rise to a loss in efficiency due to stratification. Hence, one should be rather careful in allocating the sample size to the strata. However, in actual practice it is found that the sampling variance is not very sensitive to small or even moderate deviations in the allocations. Cochran (1963) has considered this problem and has demonstrated with an example that the optimum sampling variance is fairly stable for even moderate departures from optimum allocation.

In order to estimate unbiasedly the reduction in variance due to stratification, it is necessary first to get an unbiased estimator of $V(\hat{Y}_{us})$ on the basis of a stratified sample. This can be achieved by unbiasedly estimating σ^2 given by

$$\frac{1}{N} \sum_{s=1}^K \sum_{t=1}^{N_s} Y_{st}^2 - \bar{Y}^2.$$

An unbiased estimator of the first term is $\sum_{s=1}^K \frac{W_s}{n_s} \sum_{i=1}^{n_s} y_{si}^2$, and

\bar{Y}^2 is estimated unbiasedly by $\hat{Y}_{st}^2 - v_{st}(\hat{Y}_{st})$, where $v_{st}(\hat{Y}_{st})$ denotes the unbiased variance estimator of \hat{Y}_{st} based on the stratified sample (7.25). Hence, an unbiased estimator of $V(\hat{Y}_{us})$ is given by

$$v_{st}(\hat{Y}_{us}) = \frac{1}{n} \left[\sum_{s=1}^K \frac{W_s}{n_s} \sum_{i=1}^{n_s} y_{si}^2 - \hat{Y}_{st}^2 + v_{st}(\hat{Y}_{st}) \right]. \quad \dots \quad (7.36)$$

Use of Unstratified Sample

There may be situations where data are available for an unstratified simple random sample of size n selected with replacement and it may be required to find out on the basis of this information whether there would be reduction in variance in using stratified srs instead of unstratified srs. For this purpose, $V(\hat{Y}_{us}) = \sigma^2/n$ and

$$V(\hat{Y}_{st}) = \sum_{s=1}^K \frac{W_s^2 \sigma_s^2}{n_s} = \frac{1}{N} \sum_{s=1}^K \frac{W_s}{n_s} \sum_{t=1}^{N_s} Y_{st}^2 - \sum_{s=1}^K \frac{W_s \bar{Y}_s^2}{n_s} \quad \dots \quad (7.37)$$

are to be unbiasedly estimated on the basis of the data for the unstratified sample and it can be shown that their estimators are given by

$$\hat{\tau}_{us}(\hat{\bar{Y}}_{st}) = \frac{s^2}{n} - s^2 = \frac{1}{n-1} \sum_{s=1}^K \sum_{i=1}^{n_s} \frac{n_s}{S} (y_{si} - \bar{y}_s)^2$$

and

$$\hat{\tau}_{us}(\hat{\bar{Y}}_{st}) = \frac{1}{n} \sum_{s=1}^K \frac{W_s}{n_s} \sum_{i=1}^{n_s} \frac{n_s'}{S} + \frac{1}{n-1} \sum_{s=1}^K \frac{\bar{y}_s^2}{n_s} + \frac{1}{n} \sum_{s=1}^K \frac{s^2}{n_s}, \quad (7.38)$$

where n_s is the number of sample units falling in the s th stratum, n_s is the allocation for that stratum in stratified sampling and

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n_s} \frac{n_s}{S} (y_{si} - \bar{y}_s)^2 \quad \bar{y}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} y_{si}$$

It should be borne in mind that these are estimators of second order moments and hence they are themselves subject to possibly large sampling variations. Hence such estimators are to be used in practice only when the sample size is fairly large and there is evidence to believe that the sampling fluctuation would not be large enough to vitiate the inferences drawn from the results.

7.6 STRATIFIED SRS WITHOUT REPLACEMENT

Suppose in a stratified sampling design n_s units are selected from N_s units in the s th stratum ($s = 1, 2, \dots, K$) with SRSWOR. Then an unbiased estimator of the population mean \bar{Y} , its variance and an unbiased variance estimator are given by

$$\hat{\bar{Y}}_{st} = \sum_{s=1}^K W_s \bar{Y}_s = \sum_{s=1}^K W_s \bar{y}_s,$$

$$V(\hat{\bar{Y}}_{st}) = \sum_{s=1}^K W_s V(y_s) = \sum_{s=1}^K W_s^2 (1-f_s) \frac{\sigma_s^2}{n_s}$$

$$= \sum_{s=1}^K W_s \frac{\sigma_s^2}{n_s} - \frac{1}{N} \sum_{s=1}^K W_s \sigma_s'^2, \quad (7.39)$$

and

$$v(\hat{\bar{Y}}_{st}) = \sum_{s=1}^K W_s^2 (1 - f_s) \frac{s_s^2}{n_s}, \quad \dots \quad (7.40)$$

where $f_s = n_s/N_s$, $\sigma_s'^2 = N_s \sigma_s^2/(N_s - 1)$ and s_s^2 is as defined in (7.25).

7.6a ALLOCATION TO STRATA

The optimum allocation, which minimizes the variance (7.39) for a given cost C' , is given by

$$n_s = (C' - C_0) \frac{W_s \sigma_s'/\sqrt{C_s}}{\sum_{s=1}^K W_s \sigma_s' \sqrt{C_s}}. \quad \dots \quad (7.41)$$

With this allocation, the total sample size is

$$n = \sum_{s=1}^K n_s = (C' - C_0) \frac{\sum_{s=1}^K W_s \sigma_s'/\sqrt{C_s}}{\sum_{s=1}^K W_s \sigma_s' \sqrt{C_s}} \quad \dots \quad (7.42)$$

and $V(\hat{\bar{Y}}_{st})$ becomes

$$V(\hat{\bar{Y}}_{st}) = \frac{1}{C' - C_0} \left(\sum_{s=1}^K W_s \sigma_s' \sqrt{C_s} \right)^2 - \frac{1}{N} \sum_{s=1}^K W_s \sigma_s'^2. \quad \dots \quad (7.43)$$

When the cost of survey per unit does not vary from stratum to stratum, optimum value of n_s is given by

$$n_s = n \frac{W_s \sigma_s'}{\sum_{s=1}^K W_s \sigma_s'}. \quad \dots \quad (7.44)$$

The corresponding optimum sampling variance is

$$V_0(\hat{\bar{Y}}_{st}) = \frac{1}{n} \left(\sum_{s=1}^K W_s \sigma_s' \right)^2 - \frac{1}{N} \sum_{s=1}^K W_s \sigma_s'^2. \quad \dots \quad (7.45)$$

In proportional allocation, $V(\hat{Y}_{st})$ is given by

$$V_p(\hat{Y}_{st}) = \frac{N-n}{Nn} \sum_{s=1}^K W_s \sigma'_s{}^2. \quad \dots \quad (7.46)$$

Comparing (7.46) with (7.45), we get

$$V_p(\hat{Y}_{st}) = V_0(\hat{Y}_{st}) + \frac{1}{n} \sum_{s=1}^K W_s (\sigma'_s - \bar{\sigma}')^2, \quad \dots \quad (7.47)$$

where $\bar{\sigma}' = \sum_{s=1}^K W_s \sigma'_s$. Comparing (7.46) with the sampling variance in unstratified srs wrt, namely,

$$\begin{aligned} V(\hat{Y}_{us}) &= \frac{N-n}{N-1} \frac{\sigma^2}{n} \\ &= \frac{N-n}{N-1} \frac{1}{n} \left\{ \sum_{s=1}^K \left(W_s - \frac{1}{N} \right) \sigma'_s{}^2 + \sum_{s=1}^K W_s (\bar{Y}_s - \bar{Y})^2 \right\}, \end{aligned} \quad \dots \quad (7.48)$$

we get

$$V_p(\hat{Y}_{st}) = \frac{N-1}{N} V(\hat{Y}_{us}) - \frac{N-n}{Nn} \left\{ \sum_{s=1}^K W_s (\bar{Y}_s - \bar{Y})^2 - \frac{1}{N} \sum_{s=1}^K \sigma'_s{}^2 \right\}. \quad \dots \quad (7.49)$$

7.6b GAIN DUE TO STRATIFICATION

For estimating unbiasedly the gain due to stratification on the basis of a stratified sample selected without replacement, it is necessary to obtain an unbiased estimator of $V(\hat{Y}_{us})$ and this can be shown to be given by

$$v_{st}(\hat{Y}_{us}) = \frac{N-n}{N-1} \frac{1}{n} \left[\sum_{s=1}^K \frac{W_s}{n_s} \sum_{i=1}^{n_s} y_{si}^2 - \hat{Y}_{st}^2 + v_{st}(\hat{Y}_{st}) \right] \quad (7.50)$$

Comparing (7.50) with (7.40), the reduction in variance due to stratification can be estimated.

If the gain due to stratification is to be estimated on the basis of an unstratified sample selected with srs wor it is necessary to estimate unbiasedly $V(\hat{\bar{Y}}_{st})$ on the basis of a sample selected with srs wor and such a variance estimator is given by

$$v_{us}(\hat{\bar{Y}}_{st}) = \sum_{s=1}^K \left(\frac{W_s}{n_s} - \frac{1}{N} \right) \frac{W_s}{W_s - 1/N} \left[\frac{1}{n} \sum_{t=1}^{n'_s} y_{st}^2 - \frac{1}{W_s} \left\{ \bar{y}_s'^2 - \left(\frac{1}{n} - \frac{1}{N} \right) s_s'^2 \right\} \right], \quad \dots \quad (7.51)$$

where n'_s , n_s , \bar{y}_s' and $s_s'^2$ are as defined in (7.38). Comparing this with the unbiased estimator of $V(\hat{\bar{Y}}_{us})$, namely,

$$v_{us}(\hat{\bar{Y}}_{us}) = \left(\frac{1}{n} - \frac{1}{N} \right) s^2, \quad s^2 = \frac{1}{n-1} \left(\sum_{s=1}^K \sum_{t=1}^{n'_s} y_{st}^2 - n \bar{y}^2 \right), \quad \dots \quad (7.52)$$

we get an idea of the gain due to stratification.

7.7 ESTIMATION OF A PROPORTION

The theory developed in Sections 7.5 and 7.6 for estimating \bar{Y} on the basis of stratified sampling with srswr and srs wor in the strata can easily be applied to the estimation of a population proportion P by taking the value of Y_{st} as 1 or 0 according as the unit U_{st} belongs to that class or not. In this case, \bar{Y}_s and \bar{Y} become the s -th stratum proportion P_s and the overall proportion P . Similarly, it can be seen that $\sigma_s^2 = P_s Q_s$ and $\sigma^2 = PQ$, where $Q_s = 1 - P_s$ and $Q = 1 - P$. An unbiased estimator of

$$P = \sum_{s=1}^K W_s P_s, \quad \dots \quad (7.53)$$

based on a stratified srswr sample is provided by

$$\hat{P}_{st} = \sum_{s=1}^K W_s p_s, \quad \dots \quad (7.54)$$

where p_s is the sample proportion in the s -th stratum. The sampling variance of \hat{P}_{st} is

$$V(\hat{P}_{st}) = \sum_{s=1}^K W_s^2 \frac{\sigma_s^2}{n_s} = \sum_{s=1}^K W_s^2 \frac{P_s Q_s}{n_s} \quad \dots \quad (7.55)$$

and an unbiased variance estimator is given by

$$v(\hat{P}_{st}) = \sum_{s=1}^K W_s^2 \frac{p_s q_s}{n_s - 1}. \quad (7.56)$$

The optimum allocation for a fixed sample size n is given by

$$n_s = n \frac{W_s \sqrt{P_s Q_s}}{\sum_{s=1}^K W_s \sqrt{P_s Q_s}} \quad (7.57)$$

and $V(\hat{P}_{st})$ becomes

$$V_0(\hat{P}_{st}) = \frac{1}{n} \left(\sum_{s=1}^K W_s \sqrt{P_s Q_s} \right)^2. \quad (7.58)$$

When proportional allocation is used, the sampling variance becomes

$$V_p(\hat{P}_{st}) = \frac{1}{n} \sum_{s=1}^K W_s P_s Q_s \quad (7.59)$$

Comparing (7.58) with (7.59) we get,

$$V_0(\hat{P}_{st}) = V_p(\hat{P}_{st}) - \frac{1}{n} \sum_{s=1}^K W_s (\sqrt{P_s Q_s} - D)^2, \quad (7.60)$$

where $D = \sum_{s=1}^K W_s \sqrt{P_s Q_s}$. Comparing (7.59) with the sampling variance in unstratified srswr, namely,

$$V(\hat{P}_{us}) = \frac{PQ}{n} = \frac{1}{n} \left\{ \sum_{s=1}^K W_s P_s Q_s + \sum_{s=1}^K W_s (P_s - P)^2 \right\},$$

we get

$$V_p(\hat{P}_{st}) = V(\hat{P}_{us}) - \frac{1}{n} \sum_{s=1}^K W_s (P_s - P)^2 \quad (7.61)$$

This shows that for effective stratification it is necessary to form strata such that the strata proportions vary as much as possible.

In stratified srs w.r.t. an unbiased estimator of P is given by (7.54) and that its variance and unbiased variance estimator are

$$V(\hat{P}_{st}) = \sum_{s=1}^K W_s^2 \frac{N_s - n_s}{N_s - 1} \frac{P_s Q_s}{n_s}, \quad \dots \quad (7.62)$$

and

$$v(\hat{P}_{st}) = \frac{1}{N} \sum_{s=1}^K W_s (N_s - n_s) \frac{p_s q_s}{n_s - 1}. \quad \dots \quad (7.63)$$

7.8 STRATIFIED PPS SAMPLING

Suppose a sample of n_s units is selected from N_s units of the s -th stratum with ppswr, size being x . Let Y_{si} and $P_{si} = (X_{si}/X_s)$ denote the value and the probability of selection respectively of the i -th unit in the s -th stratum and let y_{st} and p_{st} be the corresponding sample values. Then an unbiased estimator of Y is given by

$$\hat{Y} = \sum_{s=1}^K \hat{Y}_s = \sum_{s=1}^K \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{y_{si}}{p_{si}} \quad \dots \quad (7.64)$$

with

$$V(\hat{Y}) = \sum_{s=1}^K V(\hat{Y}_s) = \sum_{s=1}^K \frac{1}{n_s} \sum_{i=1}^{N_s} \left(\frac{Y_{si}}{P_{si}} - \bar{Y}_s \right)^2 P_{st} \quad \dots \quad (7.65)$$

and

$$v(\hat{Y}) = \sum_{s=1}^K v(\hat{Y}_s) = \sum_{s=1}^K \frac{1}{n_s(n_s - 1)} \sum_{i=1}^{n_s} \left(\frac{y_{si}}{p_{si}} - \hat{Y}_s \right)^2. \quad \dots \quad (7.66)$$

From (7.65) it is clear that for stratification to be effective strata should be formed such that the units within each stratum are homogeneous with respect to the variable Y_{si}/P_{si} , (or Y_{si}/X_{si}) and not with respect to the variable y or x taken separately. Since the values of Y_{si} would not be available in practice, it becomes necessary to use the values of this ratio for a previous period, or the values of a related characteristic for purposes of stratification.

7.8a ALLOCATION OF SAMPLE SIZE

The sampling variance for proportional allocation can be obtained by substituting $n_s = nW_s$ in (7.65). But a more appropriate proportional allocation in this case would be to allocate n in proportion to the strata totals of the size measure, since the total size stands for the number of sub units according to the sub units approach explained in Sub section 6.4a of Chapter 6 (p 187). Substituting $n_s = nX_s / X = nP_s$ in (7.65), we get the variance for x proportional allocation as

$$\begin{aligned} V_{px}(\hat{Y}_{st}) &= \frac{1}{n} \sum_{s=1}^K \frac{1}{P_s} \sum_{i=1}^{N_s} \left(\frac{Y_{si}}{P_{si}} - Y_s \right)^2 P_{si} \\ &= \frac{1}{n} \sum_{s=1}^K \frac{1}{P_s} \left(\sum_{i=1}^{N_s} \frac{Y_{si}^2}{P_{si}} - Y_s^2 \right) \end{aligned} \quad (7.67)$$

Comparing (7.67) with the variance in unstratified ppswr sampling namely,

$$V(\hat{Y}_{us}) = \frac{1}{n} \left(\sum_{s=1}^K \sum_{i=1}^{N_s} \frac{Y_{si}^2}{P_{si}} - Y^2 \right), \quad (7.68)$$

where $P_{si} = X_{si}/X = P_{st}P_s$, we get

$$V_{px}(\hat{Y}_{st}) = V(\hat{Y}_{us}) - \frac{1}{n} \sum_{s=1}^K \left(\frac{Y_s}{P_s} - Y \right)^2 P_s \quad (7.69)$$

This shows that stratified ppswr sampling is always more efficient than unstratified pps sampling when the allocation is proportional to X_s .

Optimum Allocation

The optimum allocation for a fixed n in this case is given by

$$n_s = n \sqrt{r_s} / \sum_{s=1}^K \sqrt{r_s} \quad (7.70)$$

and the corresponding sampling variance is

$$V_0(\hat{Y}_{st}) = \frac{1}{n} \left(\sum_{s=1}^K \sqrt{v'_s} \right)^2, \quad \dots \quad (7.71)$$

where

$$v'_s = N_s^2 V_s = \sum_{t=1}^{K_s} \left(\frac{Y_{st}}{P_{st}} - Y_s \right)^2 P_{st}.$$

Comparing expression (7.71) with (7.67), we get

$$V_0(\hat{Y}_{st}) = V_{px}(\hat{Y}_{st}) - \frac{1}{n} \sum_{s=1}^K \left(\sqrt{v'_s} - D \right)^2 P'_s, \quad \dots \quad (7.72)$$

where $D = \sum_{s=1}^K \left(\sqrt{v'_s} \right) P'_s$. The optimum allocation for a fixed cost is given by

$$n_s = (C' - C_0) \frac{\sqrt{v'_s / C_s}}{\sum_{s=1}^K \sqrt{v'_s / C_s}}. \quad \dots \quad (7.73)$$

7.8b AREA SAMPLING FOR CROP SURVEY

For estimating the acreage under a specified crop in a region, suppose the region is divided into K strata and n_s area units such as plots, fields, etc., are selected with ppswr, size being geographical area from N_s area units in the s -th stratum, ($s = 1, 2, \dots, K$). Then an unbiased estimator of the total area under the specified crop in that region is given by

$$\hat{Y}_{st} = \sum_{s=1}^K \hat{Y}_s = \sum_{s=1}^K \frac{1}{n_s} \sum_{t=1}^{n_s} \frac{y_{st}}{(g_{st}/G_s)} = \sum_{s=1}^K \frac{G_s}{n_s} \sum_{t=1}^{n_s} p_{st},$$

where y_{st} is the crop area, p_{st} is its proportion to the geographical area g_{st} , and G_s is the total geographical area of the s -th stratum. The sampling variance of \hat{Y}_{st} is

$$V(\hat{Y}_{st}) = \sum_{s=1}^K V(\hat{Y}_s) = \sum_{s=1}^K \frac{1}{n_s} \left(G_s^2 P_s Q_s - G_s \sum_{t=1}^{n_s} G_{st} P_{st} Q_{st} \right) \quad \dots \quad (7.74)$$

which reduces to

$$V(\hat{Y}_{st}) = \sum_{s=1}^K G_s^2 P_s Q_s / n_s \quad \dots \quad (7.75)$$

when the proportion of area under the crop is 1 or 0 for every area unit, that is, when in each area unit the crop is either grown on the entire area or it is not grown at all.

By substituting $n_s = nG_s/G$ in (7.75), we get the variance in the case of allocation proportional to G_s as

$$V_p(\hat{Y}_{st}) = \frac{G}{n} \sum_{s=1}^K G_s P_s Q_s \quad (7.76)$$

The optimum allocation for fixed n can be easily verified to be $n_s \propto G_s \sqrt{P_s Q_s}$, and the optimum variance is given by

$$V_0(\hat{Y}_{st}) = \frac{1}{n} \left(\sum_{s=1}^K G_s \sqrt{P_s Q_s} \right)^2. \quad \dots \quad (7.77)$$

In this particular case it is possible to obtain an approximation to optimum allocation if the strata crop proportions for a previous period are known.

By comparing (7.76) and (7.77) with the variance in unstratified ppswr sampling, namely, $V(\hat{Y}_{st}) = G^2 PQ/n$, we get

$$V_p(\hat{Y}_{st}) = V(\hat{Y}_{st}) - \frac{G}{n} \sum_{s=1}^K G_s (P_s - P)^2, \quad \left. \right\} \quad (7.78)$$

and

$$V_0(\hat{Y}_{st}) = V_p(\hat{Y}_{st}) - \frac{G}{n} \sum_{s=1}^K G_s (\sqrt{P_s Q_s} - D')^2, \quad \left. \right\}$$

where $D' = \sum_{s=1}^K G_s \sqrt{P_s Q_s} / G$

7.9 ILLUSTRATIVE EXAMPLES

In this section three examples are given to illustrate the efficiencies of different types of stratification and allocation.

Example 1

For the purpose of this study the population of 128 villages given in Annexure 4.1 of Chapter 4 (p 127) is divided into 2, 3 and 4 strata for estimating the total 1961 census population (y), such that the villages within each stratum are as similar as possible with respect to the 1951 census population (x) and the number of villages in each stratum is approximately the same. It is presumed that units within each stratum would be selected with srs wor using proportional (equal, in this case) allocation to the strata. In Table 7.1, the relative variances of the estimator are given for unstratified srs wor and for

stratified srs wor with 2, 3 and 4 strata when (i) n is fixed at 24 presuming cost to be proportional to the number of sample villages and (ii) the expected total size of the villages in the sample ($n\bar{X}$ for unstratified srs and $\sum_{s=1}^K n_s \bar{X}_s$ for stratified srs) is fixed presuming the cost to be proportional to x . From the table it is clear that stratified sampling is considerably more efficient than unstratified sampling for a fixed cost. Some other types of stratification and allocation for this population are considered in Sub-section 7.10e.

TABLE 7.1. RELATIVE VARIANCES FOR UNSTRATIFIED AND STRATIFIED SAMPLING WHEN COST IS FIXED—I.

cost proportional to	unstrati-fied srs	number of strata		
		2	3	4
(1)	(2)	(3)	(4)	(5)
number of villages	0.012575	0.005665	0.003776	0.002793
expected total size of sample villages	0.012575	0.005308	0.003533	0.002617

total 1961 census population of the tehsil 443,319; number of villages in 1961 census 128.

Example 2

The four tehsils of a district considered in Table 4.2 of Chapter 4 are treated as strata in designing a sample for estimating the total 1961 census population of the district (Y). The allocation of sample size to the strata is taken as (i) proportional to number of villages, (ii) proportional to number of persons according to 1951 census and (iii) optimum allocation for a fixed sample size. The relative variances of the estimator of Y are given in Table 7.2 for unstratified srs wor and for stratified srs wor with the three allocations mentioned above when (i) the sample size is fixed at 100 presuming cost to be proportional to number of sample villages and (ii) the expected total size of the villages in the sample ($n\bar{X}$ for unstratified srs and $\sum_{s=1}^K n_s \bar{X}_s$

for stratified srs) is fixed presuming the cost to be proportional to the 1951 census population. From this table, it appears that geographical stratification has not been very effective.

TABLE 7.2 RELATIVE VARIANCES FOR UNSTRATIFIED AND STRATIFIED SAMPLING WHEN COST IS FIXED-II

cost proportional to (1)	unstrati- fied srs (2)	allocation proportional to		optimum allocation (5)
		number of villages (3)	1951 census population (4)	
number of villages	0.006486	0.006098	0.005930	0.005843
expected total size of sample villages	0.006486	0.006058	0.006523	0.006112

(total 1961 census population of the district 1225926 number of villages 806)

Example 3

This example relates to estimation of total area under autumn paddy (A) and jute (J) in Nadia district (West Bengal) treating the 13 Police Station areas comprising of 1408 villages, as strata and sampling villages from each stratum with ppswr, size being geographical area (g) taking the allocation as (i) proportional to geographical area G , (ii) proportional to $G_s \sqrt{P_s Q_s}$, P_s being the crop proportion and $Q_s = 1 - P_s$, and (iii) optimum allocation for a fixed n . The relative variances of the estimators of A and J are shown in Table 7.3 for unstratified and stratified ppswr sampling when (i) n is fixed at 100 presuming the cost to be proportional to the number of sample villages and (ii) the expected total of g for the sample villages is fixed presuming cost to be proportional to g . Here we find that geographical stratification has been of some help in reducing the sampling variability and that for autumn paddy the allocations proportional to G_s and $G_s \sqrt{P_s Q_s}$ are almost as good as the optimum allocation, while for jute crop the allocation proportional to $G_s \sqrt{P_s Q_s}$ is a better approximation for the optimum allocation than that proportional to G_s .

TABLE 7.3. RELATIVE VARIANCES FOR UNSTRATIFIED AND STRATIFIED SAMPLING WHEN COST IS FIXED—III.

cost proportional to	unstrati-fied pps	allocation proportional to		optimum allocation
		geographi-cal area G_e	$G_e \sqrt{P_e Q_e}$	
(1)	(2)	(3)	(4)	(5)
<i>paddy crop</i>				
number of villages	0.003667	0.002922	0.002937	0.002854
expected total geographical area of sample villages	0.003667	0.003065	0.003040	0.003028
<i>jute crop</i>				
number of villages	0.018102	0.015701	0.014881	0.012008
expected total geographical area of sample villages	0.018102	0.016470	0.015610	0.012344

(total geographical area of the district: 969,632 acres; area under paddy: 276,471 acres; area under jute: 21,688 acres; number of villages: 1,408).

7.10 DEMARCATON OF STRATA

In stratified sampling K strata are formed by specifying $(K-1)$ points of demarcation and the problem of optimum stratification consists in finding these $(K-1)$ points of stratification $(y_1, y_2, \dots, y_{K-1})$ such that the variance in stratified sampling is minimized when the sampling scheme, number of strata and allocation procedure are pre-specified. The sampling variance is a function of the points of stratification, since the values of W_s and $V(\hat{Y}_s)$ in the variance depend on them. This problem has been studied objectively by Dalenius (1950, 1952), Dalenius and Gurney (1951), Dalenius and Hodges (1959). Sethi (1963) has derived the solutions for optimum points of stratification in the case of finite populations and he has also tabulated the optimum points of stratification for normal, gamma and beta distributions. Mahalanobis (1952), Dalenius and Hodges (1957) and Ekman (1959) have suggested approximations to the theoretical solutions which are easier to apply in practice. Cochran (1961) and Hess,

Sethi and Balakrishnan (1966) have examined these approximations to optimum stratification through fairly extensive empirical studies. In this section we shall consider this problem only in the case of stratified srs. The results for stratified ppswr sampling can be similarly obtained by introducing the concept of sub units (cf Sub section 6.4a of Chapter 6, p 187)

7.10a THEORETICAL SOLUTIONS

As can be expected the optimum points of stratification (ops) would depend on the selection procedure and the method of allocation used. For instance, when proportional allocation is used in stratified srswr the problem consists in finding that set of points of stratification which minimizes the variance

$$V_p(\hat{Y}_{st}) = \frac{1}{n} \sum_{s=1}^K W_s \sigma_s^2$$

In this case the problem of determining the ops reduces to one of finding that set of points of demarcation of strata, which minimize

$$\sum_{s=1}^K N_s \sigma_s^2 = \sum_{s=1}^K \left(\sum_{i=1}^{N_s} Y_{si}^2 - N_s \bar{Y}_s^2 \right),$$

Since $\sum_{s=1}^K \sum_{i=1}^{N_s} Y_{si}^2$ is a constant independent of the points of stratification, the problem reduces to that of finding the $(K-1)$ points of stratification (y_1, y_2, \dots, y_{K-1}) such that $\sum_{s=1}^K N_s \bar{Y}_s^2$ or $\sum_{s=1}^K Y_s^2/N_s$ is maximized. Let the units be arranged in increasing order of y so that the points of demarcation can be taken as y_1, y_2, \dots, y_{K-1} . Assuming that all the points of demarcation except y_s are fixed, it can be shown that the value of y_s which maximizes $\sum_{s=1}^K Y_s^2/N_s$ is given by

$$y_s = (\bar{Y}_s + \bar{Y}_{s+1})/2 \quad (7.79)$$

Proof: The above result can be proved by noting that the value of y_s would affect only two terms in $\sum_{s=1}^K Y_s^2 / N_s$, namely, Y_s^2 / N_s and Y_{s+1}^2 / N_{s+1} . If the point of demarcation is optimum then

$$(Y_s^2 / N_s) + (Y_{s+1}^2 / N_{s+1})$$

should be greater than

$$\frac{(Y_s - y'_s)^2}{N_s - 1} + \frac{(Y_{s+1} + y'_s)^2}{N_{s+1} + 1} \quad \text{and} \quad \frac{(Y_s + y''_s)^2}{N_s + 1} + \frac{(Y_{s+1} - y''_s)^2}{N_{s+1} - 1},$$

where y'_s and y''_s are the values of the units just preceding and following the unit having the optimum value y_s . After simplification, we get

$$\text{and } (\bar{Y}_{s+1} - \bar{Y}_s)(\bar{Y}_{s+1} + \bar{Y}_s - 2y'_s) + \alpha' > 0$$

$$(\bar{Y}_{s+1} - \bar{Y}_s)(\bar{Y}_{s+1} + \bar{Y}_s - 2y''_s) + \alpha'' < 0,$$

where α' and α'' are of the order of $1/N_s$ or $1/N_{s+1}$ and can be neglected if the values of $\{N_s\}$ are not very small. This shows that

$$y'_s < (\bar{Y}_s + \bar{Y}_{s+1})/2 \quad \text{and} \quad y''_s > (\bar{Y}_s + \bar{Y}_{s+1})/2.$$

Hence y_s , which lies between y'_s and y''_s , is the optimum point of demarcation between the s -th and the $(s+1)$ -th strata and

$$y_s = (\bar{Y}_s + \bar{Y}_{s+1})/2.$$

Noting that the sampling variance in the case of optimum allocation and equal allocation becomes

$$\frac{1}{n} \left(\sum_{s=1}^K W_s \sigma_s \right)^2 \quad \text{and} \quad \frac{K}{n} \sum_{s=1}^K W_s^2 \sigma_s^2$$

respectively and proceeding as before, the optimum points of stratification in the two cases can be shown to be given by

$$\frac{\sigma_s^2 + (y_s - \bar{Y}_s)^2}{\sigma_s} \doteq \frac{\sigma_{s+1}^2 + (y_s - \bar{Y}_{s+1})^2}{\sigma_{s+1}} \quad \dots \quad (7.80)$$

and

$$W_s \{ \sigma_s^2 + (y_s - \bar{Y}_s)^2 \} \doteq W_{s+1} \{ \sigma_{s+1}^2 + (y_s - \bar{Y}_{s+1})^2 \} \quad \dots \quad (7.81)$$

for $s = 1, 2, \dots, K-1$. Thus we see that the ops can be obtained as the solution of a set of equations. But in practice the solution has to be obtained through a process of iteration. For instance,

one may start with a reasonable set of $(K-1)$ points of stratification $\{y_s\}$, $s = 1, 2, \dots, K-1$, and then revise them to $\{y'_s\}$ so as to satisfy the ops equations (7.79), (7.80) or (7.81) according as proportional, optimum or equal allocation is used. That is, if proportional allocation is used, the second approximation to the ops is given by

$$y''_s = (\bar{Y}_s + \bar{Y}_{s+1})/2,$$

and the third approximation by

$$y'''_s = (\bar{Y}_s + \bar{Y}'_{s+1})/2,$$

and so on, where the strata means are obtained on the basis of the previous approximation to the ops. This process is repeated till two consecutive sets of solutions become almost identical. This procedure may be rather tedious and time consuming if the number of strata to be formed is large. Hence, some simpler methods of forming strata, which lead to approximate solutions for the ops, are considered in Sub sections 7.10c and 7.10d.

In stratified srs w.r.t. the solutions for the ops can be obtained from (7.79), (7.80) and (7.81) by substituting $\sigma_s^2 = N_s \sigma_i^2 / (N_s - 1)$ in place of σ_i^2 .

7.10b EFFECT OF USING AUXILIARY VARIABLE

At this stage it is to be noted that the problem of stratification has been solved only in terms of the values of the estimation variable y . Of course this is not realistic, since in practice the values of y are not available at the stage of stratification. Hence, optimum stratification is generally effected using an auxiliary variable, x , and the strata so formed are used for sampling for the estimation variable. The effect of using x for determining optimum strata on the variability of the estimator of the study variable y has not been fully studied, but it is expected that this procedure would lead to efficient stratification for y provided that x and y are closely related. In fact Dalenius (1957) has shown that if y and x are linearly related

the ops obtained on the basis of x are also optimum for y in case proportional allocation is used. Sethi (1963) has derived the relationship between the sampling variances for y and x , when they are linearly related. That is, if the value of y for a given x is of the form

$$y = \alpha + \beta x + e,$$

where e is a random variable with 0 as the expected value and σ_e^2 as the variance, then the estimator of \bar{Y} in stratified srss becomes

$$\hat{\bar{Y}}_{st} = \sum_{s=1}^K W_s \hat{y}_s = \sum_{s=1}^K W_s (\alpha + \beta \bar{x}_s + \bar{e}_s)$$

and

$$V(\hat{\bar{Y}}_{st}) = \beta^2 V(\hat{\bar{X}}_{st}) + \sigma_e^2 \sum_{s=1}^K (W_s^2/n_s).$$

Since $\sigma_y^2 = \beta^2 \sigma_x^2 + \sigma_e^2$ and $\beta^2 = \rho^2 \sigma_y^2 / \sigma_x^2$, we get

$$V(\hat{\bar{Y}}_{st}) = \rho^2 (\sigma_y^2 / \sigma_x^2) V(\hat{\bar{X}}_{st}) + (1 - \rho^2) \sigma_y^2 \sum_{s=1}^K (W_s^2 / n_s).$$

Hence,

$$\frac{V(\hat{\bar{Y}}_{st})}{V(\hat{\bar{Y}}_{us})} = \rho^2 \frac{V(\hat{\bar{X}}_{st})}{V(\hat{\bar{X}}_{us})} + (1 - \rho^2) n \sum_{s=1}^K \frac{W_s^2}{n_s}. \quad \dots \quad (7.82)$$

From this, it is clear that if the absolute value of the correlation coefficient (ρ) between y and x is large, then stratification based on x would be efficient for y also (cf. Problem 7.20, p. 292). It may be noted that for proportional allocation the ops based on x are also valid for y , since in that case the second term in (7.82) becomes $(1 - \rho^2)$, which is independent of the points of stratification. It may be noted that the second term may not always be small relative to the first term even for moderately large values of $|\rho|$ and hence one should be very cautious in the choice of a stratification variable, as otherwise stratified sampling may even turn out to be inefficient.

7.10c NORMAL, GAMMA AND BETA DISTRIBUTIONS

Sethi (1963) has studied the normal, gamma and beta distributions from the point of view of evolving a good stratification system and has obtained the corresponding optimum or near optimum points of stratification. We have briefly described in Section 2.2 of Chapter 2 the normal and gamma distributions, which have the frequency functions

$$f_1(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < +\infty,$$

and

$$f_2(y) = \frac{\alpha^p}{\Gamma(p)} e^{-\alpha y} y^{p-1}, \quad 0 \leq y < \infty$$

The frequency function of the beta distribution is of the form

$$f_3(y) = \frac{1}{B(p+1, q+1)} y^{p-1} (1-y)^{q-1}, \quad 0 \leq y \leq 1$$

The values of the distribution functions $\{F(y') = \int_0^{y'} f(y) dy\}$ of these distributions at the ops have been tabulated for different allocations in the case of stratified srswr. The values of $\{y\}$ corresponding to these $\{F(y)\}$ are to be taken as the ops for the allocation used.

For the normal distribution with zero mean and unit standard deviation, the values of the distribution function corresponding to the ops for proportional, equal and optimum allocations are given in Table 7.4 for $K = 2, \dots, 10$. In practice, when a variable is known to be normally distributed, the ops can be obtained as those points in the frequency distribution where the cumulative frequencies equal the tabulated optimum values of the distribution function. Alternatively the ops $\{y_s\}$ corresponding to the tabulated values of the distribution function may be converted to $\{\bar{Y} + y_s \sigma\}$ to give the ops for the study variable, if it is distributed as $N(\bar{Y}, \sigma^2)$.

TABLE 7.4. DISTRIBUTION FUNCTION $F(y)$ OF THE NORMAL DISTRIBUTION $N(0,1)$ AT THE OPTIMUM POINTS OF STRATIFICATION.

number of strata	$F(y) (1000)$ at optimum stratification point								
	1	2	3	4	5	6	7	8	9
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. proportional allocation									
2	500								
3	271	729							
4	161	500	839						
5	107	352	648	893					
6	078	261	500	739	922				
7	056	195	394	606	805	944			
8	038	142	305	500	695	858	962		
9	033	121	255	413	587	745	879	967	
10	021	84	189	334	500	666	811	916	979
2. equal allocation									
2	500								
3	285	715							
4	184	500	816						
5	128	368	632	872					
6	095	277	500	723	905				
7	076	222	405	595	778	924			
8	059	182	333	500	667	818	941		
9	048	147	278	424	576	722	853	952	
10	042	126	239	366	500	634	761	874	958
3. optimum allocation									
2	500								
3	291	709							
4	188	500	812						
5	132	367	633	868					
6	100	281	500	719	900				
7	075	218	402	598	782	925			
8	062	181	333	500	667	819	938		
9	052	154	283	426	574	717	846	948	
10	044	131	242	367	500	633	758	869	956

Source : Sethi, V. K. (1963) : A note on optimum stratification of populations for estimating the population means; *Australian Journal of Statistics*, 5, 20-33.

For the gamma distribution, the variable y having the parameters $\alpha = 1/2$ and $p = \nu/2$ are considered and the values of the distribution

function corresponding to the ops have been tabulated for proportional and equal allocation and it is conjectured that the ops for equal allocation would be approximately the same as those for optimum allocation when v is not large. Here again the ops for any variable following gamma distribution with the parameters α and p can be obtained as the points where the cumulative frequencies equal the tabulated values of the distribution function for $v = 2p$. Alternatively the ops for this variable may be obtained by multiplying the ops corresponding to the tabulated values of the distribution function for $v = 2p$ by 2α . The values of the distribution function of the gamma distribution $G(\frac{1}{2}, v/2)$ for $K = 1, 2, \dots, 6$ and $v = 2(2) 10(5) 30$ are given in Table 7.5.

The ops for the beta distribution $B(p, q)$ in the case of proportion allocation are given by the points, which divide the beta distribution $B(p+1, q+1)$ into equal proportions. For equal and optimum allocation the ops are given by the points, which divide the cumulatives of $\sqrt{f(y)}$ into equal proportions.

An idea of the gain in precision that can be achieved by resorting to optimum stratification is provided by Table 7.6 giving the value of the variance ratio $V_{po}(\hat{Y}_{st})/V(\hat{Y}_{us})$, where the subscript po denotes optimum stratification with proportional allocation. From the table it is interesting to note that for a given number of strata the reduction in the sampling variance due to the use of optimum stratification with proportional allocation over that of an unstratified design is approximately the same for the normal distribution as for the gamma distribution with v varying from 1 to 30 and that the reduction in variance with increase in the number of strata is most at the initial stages and becomes marginal after a certain stage. However, it may be noted that the gain in precision given in Table 7.6 is on the basis of stratification based on the estimation variable itself which is impracticable and that if stratification is done on the basis of a supplementary variable x the variance ratios given in the table would get inflated by the factor $(1/\rho^2)$ where ρ is the correlation coefficient between x and y , as has been shown in Sub section 7.10b.

TABLE 7.5. DISTRIBUTION FUNCTION $F(y)$ OF THE GAMMA DISTRIBUTION $G(\frac{1}{2}, r/2)$ AT THE OPTIMUM POINTS OF STRATIFICATION.

number of strata	strati- fication points	values* of $F(y) (1000)$ at ops for gamma distribution with r as									
		1	2	4	6	8	10	15	20	25	30
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. proportional allocation											
2	1	862	798	733	697	674	658	631	614	603	594
3	1	746	632	537	494	464	439	405	360	353	346
	2	952	926	893	871	857	849	832	778	765	768
4	1	657	528	408	350	330	313	276	247	243	233
	2	879	817	751	715	690	673	644	579	570	564
	3	976	961	941	931	923	916	904	865	858	855
5	1	561	423	337	269	242	228	200	185	180	169
	2	808	713	620	557	547	533	497	478	463	456
	3	926	889	841	815	798	787	763	748	740	731
	4	986	976	964	957	951	948	939	935	931	928
6	1	520	362	264	217	191	176	150	140	130	126
	2	746	632	522	482	452	430	398	380	364	366
	3	879	817	751	715	690	673	644	626	613	604
	4	954	929	897	879	866	856	839	827	811	814
	5	991	985	977	972	969	966	960	957	954	952
2. equal allocation**											
2	1	808	740	680	641	623	610	591	579	570	564
3	1	657	551	475	430	397	400	375	360	353	346
	2	911	878	841	821	809	798	791	778	768	768
4	1	561	423	337	310	286	275	247	247	243	233
	2	808	740	680	641	623	610	591	579	570	564
	3	949	926	905	895	884	882	869	866	858	855
5	1	473	362	264	230	211	202	187	175	171	165
	2	706	632	557	518	495	487	467	452	440	430
	3	879	826	777	754	742	734	716	706	701	694
	4	968	953	939	928	926	921	913	910	904	903
6	1	416	295	228	191	181	168	150	141	137	119
	2	657	528	459	430	409	391	367	460	317	325
	3	808	740	680	641	623	610	591	579	570	559
	4	911	878	847	821	814	803	791	791	778	771
	5	976	967	952	946	943	940	933	930	926	928

*These are only close approximations to the optimum values.

**Optimum stratification for equal allocation is expected to be a good approximation to that for optimum allocation.

Source : Sethi, V. K. (1964) : Contributions to stratified sampling and some related problems; thesis submitted for the Ph.D. degree of the Indian Statistical Institute.

TABLE 7.6 VARIANCE OF OPTIMUM STRATIFIED SAMPLING
WITH PROPORTIONAL ALLOCATION RELATIVE TO THAT
OF UNSTRATIFIED SAMPLING

parent distribution	variance ratio $V_{po}(\bar{Y}_{st})/V(\bar{Y}_{us})$ when K is				
	2	3	4	5	6
(1)	(2)	(3)	(4)	(5)	(6)
$N(0, 1)$.393	.190	.117	.080	.058
$G\left(\frac{1}{2}, \frac{v}{2}\right), r=1$.348	.177	.107	.072	.052
2	.353	.179	.110	.075	.054
4	.357	.184	.113	.077	.046
6	.359	.186	.114	.077	.056
8	.359	.186	.115	.078	.057
10	.362	.187	.116	.079	.058
15	.363	.189	.118	.080	.058
20	.364	.191	.119	.081	.059
25	.364	.191	.119	.082	.059
30	.364	.191	.120	.083	.059

Source, Same as for Table 7.5

7.10d APPROXIMATIONS TO OPS

Since the theoretical solutions for obtaining the optimum points of stratification given in Sub-section 7.10a are difficult to apply to practical situations due to the heavy computational work involved, a number of approximations have been suggested.

(i) Equalization of $W_s \sigma_s$

Dalenius and Gurney (1951) conjectured that the formation of strata on the basis of equalization of $W_s \sigma_s$ and giving equal allocation to the strata would lead to optimum stratification. This method also is not simple, as it envisages the calculation of the values of σ_s for different sets of stratification points.

(ii) *Equalization of Strata Totals*

Mahalanobis (1952) proposed the equalization of the strata totals ($N_s \bar{X}_s$) with equal allocation. The main advantage of this rule is its simplicity. Hansen, Hurwitz and W. G. Madow (1953) demonstrated that this would lead to efficient stratification if the strata coefficients of variation are the same and remain about the same for slight modifications of the strata boundaries, as in that case it approximates to Dalenius-Gurney rule of equalizing $W_s \sigma_s$. Though this situation is likely to be obtained when stratification is done mainly on geographical or administrative considerations, this procedure does not necessarily lead to efficient stratification when applied to normal, gamma and beta distributions (Sethi, 1963).

(iii) *Equalization of $W_s R_s$*

Aoyama (1954) suggested the formation of strata on the basis of equalization of strata ranges and having equal allocation. This rule is based on the assumption that the distribution within a stratum is approximately rectangular, which is likely to be true only if the number of strata is large. Ekman (1959) suggested the equalization of $W_s R_s$, where R_s is the range of the variable in the s -th stratum for forming strata with equal allocation. This procedure is likely to be useful in practice, since it is fairly simple.

(iv) *Equalization of Cumulatives of $\sqrt{f(y)}$*

Dalenius and Hodges (1957) proposed formation of strata by equalizing the cumulative of $\sqrt{f(y)}$, where $f(y)$ is the frequency function. In deriving this rule, it is assumed that the distribution is bounded and that the number of strata is large. The first assumption is generally valid in practice. As regards the second assumption, it may be noted that when the number of strata is large, any *reasonable* rule of stratification would lead to near optimum stratification and hence the basic point needing attention in this case is whether this rule provides reasonably good approximation for smaller number of strata also. In fact the authors of this rule have pointed out that for some continuous distributions this gives a close approximation

to optimum stratification even when the number of strata is as small as 2 or 3 (Dalenius and Hodges, 1959). This observation has been supported by Sethi (1963) who has shown that this rule leads to more efficient stratification than equalization of strata totals or ranges for truncated gamma distributions.

(v) *Equalization of $\frac{1}{2}\{r(y)+f(y)\}$*

Durbin (1959) proposed the equalization of the cumulative frequencies of a distribution, $g(y)$, which is midway between the original distribution $f(y)$ and a rectangular distribution $r(y)$ over the range (y_0, y_K) of y . That is, $r(y)$ is taken as $F(y_K)/(y_K - y_0)$ and the ops are obtained by equalizing the cumulatives of the function

$$g(y) = \frac{1}{2}\{r(y)+f(y)\} \quad . \quad (7.83)$$

This rule has been arrived at as a first order correction to the rule of equalizing the strata ranges by considering simple departures from the rectangular distribution.

7.10e AN EMPIRICAL STUDY

For the purpose of the study, the 128 villages given in Annexure 4 1 of Chapter 4 (p 127) are arranged in increasing order of their 1951 census population (x) and 2 3 and 4 strata are formed by equalizing the strata totals for (i) N_s , the number of villages, (ii) $N_s\bar{X}_s$, the stratum total of x , and (iii) N_sR_s , R_s being the range of x in the stratum. The relative variances for the estimator of the total census population in 1961 are given in Table 7.7 for stratified srs w.r.t. allocation proportional to (i) N_s , (ii) $N_s\bar{X}_s$, (iii) N_sR_s and (iv) $N_s\sigma_s$, σ_s being the standard deviation of x , when (i) number of sample villages is taken as 24, presuming the cost to be proportional to the number of sample villages and (ii) expected total of sizes of sample villages is taken as fixed presuming the cost to be proportional to x . From the table, we find that formation of strata by equalizing N_sR_s and allocation proportional to N_sR_s provide good approximations to optimum stratification and allocation.

TABLE 7.7. RELATIVE VARIANCES FOR DIFFERENT METHODS OF STRATIFICATION AND ALLOCATION WITH COST FIXED.

number of strata	method of stratification: equalization of	relative variance for allocation proportional to			
		N_s	$N_s \bar{X}_s$	$N_s R_s$	$N_s \sigma_s$
(1)	(2)	(3)	(4)	(5)	(6)
(i) sample size is fixed at 24 sample villages					
2	N_s	.005665	.005266	.005071	.005005
	$N_s \bar{X}_s$.005047	.005052	.004773	.004773
	$N_s R_s$.005082	.004878	.004702	.004702
3	N_s	.003771	.003469	.003035	.002788
	$N_s \bar{X}_s$.002994	.003583	.003678	.002708
	$N_s R_s$.004918	.003690	.003108	.003107
4	N_s	.002793	.002444	.002077	.002011
	$N_s \bar{X}_s$.002100	.002908	.001943	.001915
	$N_s R_s$.001709	.001852	.001462	.001496
(ii) expected total of sizes of sample villages is fixed					
2	N_s	.005308	.006084	.005674	.005417
	$N_s \bar{X}_s$.004743	.005770	.004871	.004871
	$N_s R_s$.004800	.005544	.004983	.004983
3	N_s	.003533	.004206	.003618	.003344
	$N_s \bar{X}_s$.002789	.004327	.002843	.002774
	$N_s R_s$.004479	.004572	.003422	.003502
4	N_s	.002617	.002993	.002736	.002565
	$N_s \bar{X}_s$.001977	.003599	.001809	.001981
	$N_s R_s$.001642	.002282	.001640	.001732

7.11 DETERMINATION OF NUMBER OF STRATA

It is easy to see that the efficiency of stratification increases if the number of strata is increased by bifurcating the existing strata in any given system of stratification and allocation. There are, however, two limits to having an indefinitely large number of strata. The first limit, a natural one, is the sample size, since in stratified sampling at least one unit is to be selected from each stratum for getting unbiased estimators of the population mean, total or proportion. But if only one unit is selected from a stratum, it is not possible to obtain an unbiased variance estimator, though some approximate, but biased, estimators of the sampling variance can be obtained, (Problems 7.21 and 7.22 p 292). However, this difficulty can be easily overcome by selecting at least two units from each stratum, in which case the maximum number of strata that could be formed is $n/2$ if n is even and $(n-1)/2$ if n is odd. The second limit is set by the fact that the gain in efficiency, though substantial for initial increases in the number of strata, becomes marginal after a certain stage. In other words, the gain in efficiency in increasing the number of strata beyond a certain stage becomes incommensurate with the effort involved. In fact, the saturation point may be reached with a small number of strata if the stratification variable is only moderately correlated with the estimation variable. Hence, it appears that in large scale surveys, there may not be much advantage in increasing the number of strata up to the maximum possible extent unless the relevant information is available for several suitably chosen auxiliary variables. At this stage, it may be noted out that there may be situations, where increase in number of strata may even lead to less homogeneous strata and thereby decrease the efficiency of stratified sampling, if the principles of stratification are not properly adhered to.

Dalenius (1953) conjectured that the variance of an estimator of \bar{Y} based on K optimum strata is about $\{(K-1)/K\}^2$ times the variance in the case of $(K-1)$ optimum strata. That is,

$$V_K(\hat{\bar{Y}}) = \left(\frac{K-1}{K}\right)^2 V_{K-1}(\hat{\bar{Y}}) \quad (7.84)$$

This conjecture is based on the result that the variance is inversely proportional to the square of the number of strata in the case of rectangular population. Considering the populations with frequency functions e^{-y} and ye^{-y} , Dalenius has shown that the behaviour of the variance with increase in K conforms to (7.84). Cochran (1961) compared the variance ratios of the form $V_K(\hat{Y})/V_{K-1}(\hat{Y})$ with the values of $\{(K-1)/K\}^2$ for $K = 2, 3$ and 4 for eight distributions relating to different characteristics such as agricultural loans, bank reserve, college students, etc. and found that they are in broad agreement and that apparently there is no relation between skewness and rate of reduction in variance. Dalenius (1953) also suggested the use of a suitable cost function in determining the optimum number of strata.

Sethi (1963) has found that for optimum stratification with proportional and equal allocation in case of gamma distributions, the variance ratio $V_K(\hat{Y}_{st})/V(\hat{Y}_{us})$ can be expressed as the inverse of a quadratic function in the number of strata, K . That is, the proposed relation is of the form

$$\{V(\hat{Y}_{us})/V_K(\hat{Y}_{st})\} = aK^2 + bK + c, \quad \dots \quad (7.85)$$

where a, b and c are constants to be determined by considering the values of the variance ratio on the left hand side of (7.85) for $K = 1, 2$ and 3 . It is found that the values of the variance ratio obtained on the basis of this model for $K = 4, 5$ and 6 in case of different gamma distributions for both proportional and equal allocations agree well with the actual values of the variance ratio, demonstrating the usefulness of this model.

For determining the optimum number of strata, the following cost function may be used :

$$C = C_0 + KC_1 + nC_2, \quad \dots \quad (7.86)$$

where C_0 is the overhead cost and C_1 and C_2 are the costs per stratum and per unit respectively. Noting that the relationship between sampling variance and number of strata is of the form

$$V_E(\hat{Y}_{st}) = \frac{\sigma^2}{n} (aK^2 + bK + c)^{-1} \quad (7.87)$$

The optimum value of K for a fixed cost can be obtained as that value which minimizes (7.87) subject to the cost restriction (7.86). This can be achieved by first finding the sample sizes that could be had for different values of K when the cost is fixed and then locating the minimum value of the variance on the graph showing the value of the variance against K . For ensuring a specified precision with minimum cost a similar procedure can be followed to obtain the optimum number of strata by graphing the cost of the survey against the number of strata.

7.12 INTERPENETRATING SUB-SAMPLES

Suppose the sample from each stratum in a stratified sampling design is selected in the form of m independent interpenetrating sub-samples of same size according to the same sampling design such as srs systematic or pps and let \hat{Y}_{st} be the value of an unbiased estimator of the total of the s th stratum based on the i th sub sample. Then an unbiased estimator based on all the m sub samples is provided by the mean of the sub sample estimates namely

$$\hat{Y}_s = \frac{1}{m} \sum_{i=1}^m \hat{Y}_{st}$$

and an unbiased estimator of its sampling variance is given by

$$v(\hat{Y}_s) = \frac{1}{m(m-1)} \sum_{i=1}^m (\hat{Y}_s - \hat{Y}_{st})^2$$

Hence, an unbiased estimator of Y and its variance estimator are given by

$$\hat{Y} = \sum_{s=1}^K \hat{Y}_s = \frac{1}{m} \sum_{s=1}^K \sum_{i=1}^m \hat{Y}_{si} \quad \dots \quad (7.88)$$

and

$$v(\hat{Y}) = \sum_{s=1}^K v(\hat{Y}_s) = \frac{1}{m(m-1)} \sum_{s=1}^K \sum_{i=1}^m (\hat{Y}_{si} - \hat{Y}_s)^2. \dots \quad (7.89)$$

Another unbiased estimator of the variance, which is easier to calculate, can be obtained on the basis of the m estimates of the population total that can be built up by adding the estimates of \hat{Y}_s separately for the m sub-samples. Since the m sub-sample estimates of Y_s are independent unbiased estimates having the same sampling variance, the estimates,

$$\hat{Y}_i = \sum_{s=1}^K \hat{Y}_{si}, \quad i = 1, 2, \dots, m,$$

are m unbiased estimates of Y having the same variance. Hence the combined estimator \hat{Y} , the mean of these m estimates, is unbiased for Y and the variance of the estimator is of the form V/m , where V is the variance of the estimator based on one sub-sample. An unbiased variance estimator in this case is given by

$$v(\hat{Y}) = \frac{1}{m(m-1)} \sum_{i=1}^m (\hat{Y}_i - \hat{Y})^2; \quad \dots \quad (7.90)$$

which is simpler to compute than the variance estimator given in (7.89). Murthy (1962) has shown that the variance estimator (7.89) is more efficient than (7.90) and has derived an expression for the loss in precision in using (7.90) instead of (7.89), (cf. Problem 7.11, p.290). The author has also considered the question of setting up confidence intervals based on the variance estimators (7.89) and (7.90).

Other Short-cut Methods

If the number of strata is too large to enable quick computation of the variance estimator (7.89) and if the number of sub-samples m is too small to provide a reliable estimate of the variance based on (7.90), one of the two following methods of estimating the sampling variance unbiasedly, which are in a sense compromises between (7.89)

and (7.90), may be used. One method consists in pooling the stratum wise sub sample estimates over strata in each of $K' (< K)$ groups into which the K strata may conveniently be grouped and then using the variance estimator (7.89) treating the K' groups as strata. For instance, if in a survey each region is divided into some strata and if m sub samples are selected independently from each stratum, the variance estimator may be built up conveniently from the regional sub sample estimates instead of from the stratum level or from the overall sub sample estimates. The other procedure consists in first arranging the m sub sample estimates in each stratum at random and obtaining an estimate of variance on the basis of the m pooled estimates as in (7.90). Then this procedure is repeated r times and the mean of the r variance estimates is considered as the variance estimator. The efficiency of this variance estimator increases with r and in fact this variance estimator approximates to that given in (7.89), when r is sufficiently increased.

7.13 MULTIPLE STRATIFICATION

If the study variables are all related to one single supplementary variable (x) for which information is available for all population units, then stratification and allocation may be done in an optimum fashion using the data on x . But all the characteristics of interest may not be related to one supplementary variable but to two or more auxiliary variables. For instance, the area characteristics like acreage under crops are likely to be related to geographical or cultivated area, whereas the population characteristics like age distribution, distribution of labour force, etc. are likely to be related to population.

In such a situation, the units may first be grouped into primary strata with respect to the most important of the stratification variables and then within each of the primary strata so formed, secondary or sub strata may be constructed according to another supplementary variable, and so on. This procedure is known as *multiple stratification* or *deep stratification*.

Once the deep strata are formed, allocations may be made on the basis of each of the stratification variables and a compromise allocation is to be arrived at. For instance, the compromise allocation may be the weighted mean of the different allocations, the weights being specified on the basis of the relative importance of the stratification variables. Usually slight modifications in the allocations

do not affect the efficiency appreciably. But if it is found that the allocations arrived at on the basis of different stratification variables differ considerably, then it would not be desirable to attempt a compromise in the allocations. In such a situation, it may be useful to have a broad compromise allocation first and then to supplement it by additional samples in those strata where the individual allocation for any particular variable is considerably higher than the compromise allocation.

A number of methods have been suggested for arriving at a compromise allocation in a survey involving more than one stratification variable. One of the methods suggested is the maximization of the sum of the efficiencies for each variable of the compromise allocation as compared to the respective individual optimum allocation, that is, maximization of

$$E = \sum_{i=1}^t E_i, \quad E_i = V_o(\hat{\bar{Y}}_i)/V_c(\hat{\bar{Y}}_i), \quad \dots \quad (7.91)$$

where $\hat{\bar{Y}}_i$ stands for the estimator of the i -th characteristic and the subscripts o and c denote optimum and compromise allocations respectively. Dalenius (1957) has suggested the minimization of a weighted loss function

$$L = \sum_{i=1}^t a_i L_i, \quad L_i = \frac{V_c(\hat{\bar{Y}}_i) - V_o(\hat{\bar{Y}}_i)}{V_o(\hat{\bar{Y}}_i)}, \quad \dots \quad (7.92)$$

where $\{a_i\}$ are specified on the basis of the importance of the characteristics. Chakravarthy (1954) and S. P. Ghosh (1958) have proposed the minimization of the generalized variance, which is the determinant of the dispersion (or variance-covariance) matrix of the variables under study.

7.14 TECHNIQUE OF POST-STRATIFICATION

The technique of *post stratification* consists in dividing the population and the selected sample at the estimation stage into a certain number of strata (K' , say), termed *post strata*, and estimating \bar{Y} by the weighted mean of the estimators of the post strata means, the weights being the proportion of units in the post strata. That is, the estimator proposed is of the form

$$\hat{Y}_{pst} = \sum_{s=1}^{K'} W_s (\hat{Y}_s / \hat{N}_s) \quad (7.93)$$

where the subscript pst denotes post stratification, and \hat{Y}_s and \hat{N}_s are the unbiased estimators of the s th post stratum total of y and of the total number of units in that post stratum respectively. The estimators \hat{Y}_s and \hat{N}_s are respectively obtained from the usual unbiased estimator \hat{Y} by substituting the value of y and 1 respectively for sample units belonging to the s th post stratum and assigning the value 0 to those units not belonging to that post stratum. The estimator (7.93) is sometimes used in practice instead of the conventional estimator \hat{Y}_{us} which can be written as $\frac{1}{N} \sum_{s=1}^{K'} \hat{Y}_s$, since the post stratified estimator \hat{Y}_s / \hat{N}_s is expected to have a smaller sampling variance than the estimator \hat{Y}_s / N_s though it is biased due to the use of the *ratio estimator* \hat{Y}_s / \hat{N}_s . Hence the problem to be faced in the case of post stratification is whether the possible reduction in the sampling variance achieved through post stratification is adequate to offset the bias introduced in the estimator. The technique of post-stratification using a particular stratification variable can be resorted to even when stratified sampling based on another stratification variable has been used in drawing the original sample. This technique is also useful in obtaining estimates for domains of study by treating them as post strata, if they are not already treated as *selection strata*.

The problem of post-stratification has been discussed by Hansen, Hurwitz and W. G. Madow (1953) and Cochran (1963). Williams (1962) has given a simple procedure of finding approximations to the variance and the variance estimator of a post-stratified estimator on the basis of the variance and variance estimator of the original estimator. This procedure is based on the result that

$$V\left(\frac{\hat{Y}}{\hat{X}}\bar{X}\right) \doteq V\left(\hat{Y} - \frac{\bar{Y}}{\bar{X}}\hat{X}\right), \quad \dots \quad (7.94)$$

when the sample size is fairly large (cf. Chapter 10). The variance of the post-stratified estimator \hat{Y}' and its variance estimator are obtained from the variance of the usual estimator \hat{Y} and its variance estimator by substituting $\bar{Y}_s - \bar{Y}$ in place of \bar{Y}_{st} and $y_{st} - \hat{Y}$ in place of y_{st} respectively.

For instance, suppose post-stratification is resorted to when the original sampling design is unstratified srs wr. The variance of the usual estimator \hat{Y} and its variance estimator can be written as

$$V(\hat{Y}) = \frac{N-n}{N-1} \frac{1}{n} \frac{1}{N} \sum_{s=1}^{K'} \sum_{i=1}^{N'_s} (\bar{Y}_{st} - \bar{Y})^2, \quad \dots \quad (7.95)$$

and

$$v(\hat{Y}) = \frac{N-n}{Nn} \frac{1}{n-1} \sum_{s=1}^{K'} \sum_{i=1}^{n'_s} (y_{st} - \bar{y})^2, \quad \dots \quad (7.96)$$

where N'_s and n'_s are the number of population units and sample units falling in the s -th post-stratum. In this case the post-stratified estimator \hat{Y}' will be

$$\hat{Y}' = \sum_{s=1}^{K'} W'_s \bar{y}'_s, \quad \dots \quad (7.97)$$

where y_s is the mean of the n_s sample observations falling in the s th post stratum. Its variance and variance estimator are approximately given by

$$V(\hat{Y}) = \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{1}{N} \sum_{s=1}^K \sum_{i=1}^{N_s} (Y_{si} - \bar{Y}_s)^2 \quad (7.98)$$

and

$$v(\hat{Y}) = \frac{N-n}{Nn} \cdot \frac{1}{n-1} \sum_{s=1}^K \sum_{i=1}^{n_s} (y_{si} - y_s)^2 \quad (7.99)$$

These approximations to the variance and the variance estimator of the post stratified estimator are likely to be close to the actual values only when the sample size is fairly large (cf Problem 7.23 p 292)

7.15 CONTROLLED SELECTION

In this section a brief discussion is given on the devices available for effecting controls beyond stratification in selecting the sample so as to get estimates with greater precision per unit of cost than is possible in simple stratified sampling. R. Goodman and Kish (1950) suggested a process of selection termed *controlled selection*, which while retaining the probabilities of selection assigned to the units in a stratified sampling design ensures greater probabilities of selection for some or all preferred combinations of n out of N units and consequently less probabilities of selection for some or all non preferred combinations than in the original stratified sampling design. At this stage it may be mentioned that even stratified sampling and systematic sampling with a prespecified arrangement are also a kind of controlled selection as in these designs the probabilities of some or all preferred combinations of units are increased and those of some or all non preferred combinations are reduced as compared to an unstratified sampling design. However, in this section the term controlled selection is used in the sense of having controls beyond stratification for reducing the sampling variance of the estimator of the population parameter under study.

The fact that it is possible to introduce additional control in stratified sampling can be illustrated by the example of sampling one hospital from each of two size-strata given by Hess, Riedel and Fitzpatrick (1961). Let A, B, C, D be the four large hospitals forming stratum 1 and let a, b, c, d and e be the five small hospitals forming stratum 2. If one hospital is to be selected from each stratum with srs, then the probability of selection of any unit is 0.25 in stratum 1 and 0.20 in stratum 2. Suppose it is further known that the hospitals A, B, a and b have ownership code 1 and that the hospitals C, D, c, d and e have the ownership code 2. In controlled selection an attempt is made to increase the probability of getting samples with hospitals having different ownership codes retaining at the same time the originally assigned probabilities of selection. One way of achieving this is to re-arrange the units in the two strata such that in stratum 1 hospitals with ownership code 2 come first and then those with ownership code 1 and in stratum 2 the hospitals with code 1 come first and then those with code 2 and to select 2 hospitals systematically with probability proportional to their original probabilities of selection from all the units in the two strata taken together. All possible samples in stratified and controlled sampling are shown in Table 7.8, from which it is clear that the probability of getting a preferred sample having hospitals with different ownership codes is 0.90 in controlled selection as compared to 0.50 in stratified sampling.

As can be seen from the above example, in controlled selection an attempt is made to improve upon simple stratified sampling by making the selection in the different strata dependent on each other. It may be noted that in the above example we are conceptually trying to select 2 units from 4 deep strata formed on the basis of two stratification systems—one according to size and the other according to ownership—such that one unit is selected from each of the two size-strata and that the units in a sample are as heterogeneous as possible with respect to the second system of stratification. Thus the problem of controlled selection may be posed in a general way as that of devising a method of selecting a sample, when the

TABLE 7.8 ALL POSSIBLE SAMPLES IN STRATIFIED AND CONTROLLED SAMPLING WITH THEIR PROBABILITIES OF SELECTION

stratum	hospi- tal *	owner- ship code	proba- bility	stratified sampling				controlled selection	
				sample **	proba- bility	sample ***	proba- bility	sample **	proba- bility
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	O	2	25	C_2a_1	0.5	A_1a_1	0.5	C_2a_1	20
	D	2	25	C_2b_1	0.5	A_1b_1	0.5	C_2b_1	05
	A	1	25	C_2c_2	0.5	A_1c_2	0.5	D_2b_1	15
	B	1	25	C_2d_2	0.5	A_1d_2	0.5	D_2c_2	10
2	a	1	20	C_2e_2	0.5	A_1e_2	0.5	A_1e_2	10
	b	1	20	D_2a_1	0.5	B_1a_1	0.5	A_1d_2	15
	c	2	20	D_2b_1	0.5	B_1b_1	0.5	B_1d_2	05
	d	2	20	D_2c_2	0.5	B_1c_2	0.5	B_1e_2	20
	e	2	20	D_2d_2	0.5	B_1d_2	0.5		
				D_2e_2	0.5	B_1e_2	0.5		

* after re arrangement for controlled selection

** The subscript denotes the ownership code of the hospital A preferred sample is one with the hospitals having different ownership codes

number of multiple or deep strata is more than the sample size, such that the allocation to two or more systems of strata and the original probabilities of selection for the units are ensured Bryant, Hartley and Jessen (1960) have given an interesting and fairly simple solution to this problem and the procedure suggested by them is briefly described here

Two way Stratification

Suppose the population is stratified into K and K' strata on the basis of two stratification variables x_1 and x_2 respectively Let W_{is} , $s = 1, 2, \dots, K$ and $i = 1, 2, \dots, K'$, be the proportions of units in the KK' deep strata formed by combining the two systems of strata and let W_{is} and $W_{is'}$ be the proportions of units in the s th stratum of the first system and in the s' th stratum of the second system respectively In the case of proportional allocation, the sample sizes in the strata for the two systems are given by $\{nW_{is}\}$ and $\{nW_{is'}\}$ respectively In the present case, it is assumed that the sample

size n is not large enough to ensure positive integral allocations to each of the KK' multiple or deep strata, since if the sample size were large compared to the total number of strata the usual methods of allocation (e.g., $n_{ss'} = nW_{ss'}$) can be applied without much difficulty. Hence, instead of ensuring exact proportional allocation to the multiple strata, that is, $\{n_{ss'}\} = \{nW_{ss'}\}$, the suggested procedure ensures that the expected values of the allocations to the multiple strata are proportional to the number of units in them, that is, $E(n_{ss'}) = nW_{ss'}^*$ (or $n_s \cdot n_{s'} / n$).

The following steps are involved in this procedure :

- (i) constructing a square of n^2 cells with n rows and n columns,
- (ii) selecting n of the cells with equal probability such that no two selected cells belong to the same row or column,
- (iii) grouping the n rows into K strata such that the s -th stratum has an allocation of n_s units,
- (iv) grouping the n columns into K' strata such that the s' -th stratum has an allocation of $n_{s'}$ units, and
- (v) taking the allocation in the (ss') -th deep stratum as the number of cells (say, $n_{ss'}$) selected in the joint-group formed by the s -th group of (iii) and the s' -th group of (iv).

An unbiased estimator of the population mean in this case is given by

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{s=1}^K \sum_{s'=1}^{K'} G_{ss'} n_{ss'} \bar{y}_{ss'}, \quad \dots \quad (7.100)$$

where $\bar{y}_{ss'}$ is the sample mean in the (ss') -th stratum and $G_{ss'} = n^2 W_{ss'} / n_s \cdot n_{s'}$.

The above procedure may be illustrated by applying it to the example of sampling 2 hospitals considered earlier in this section. Suppose the 9 hospitals in the population are stratified by two systems of stratification on the basis of information on number of beds and ownership and it is desired to have one sample hospital per stratum in the two systems. In this case there are 4 deep strata formed on a joint consideration of size and ownership and two units are to be selected such that the marginal allocations of 1 unit per stratum for the two systems of stratification are achieved and that the expected allocation to the deep strata are equal to $n_s \cdot n_{s'} / n$ which is 1/2 in the present case. The application of the above procedure to this example would require the selection of one hospital from size-stratum 1 and if it happens to possess ownership code 1 (or 2), one hospital is to be selected from those having ownership code 2 (or 1) in size-stratum 2.

REFERENCES

- Aoyama, H. (1954) : A study of stratified random sampling; *Ann. Inst. Stat. Math.*, 6, 1-36.
- Bowley, A. L. (1926) : Measurement of the precision attained in sampling; *Bull. Inter. Stat. Inst.*, 22, (1), 1-62.
- Bryant, E. C., Hartley, H. O. and Jessen, R. J. (1960) : Design and estimation in two-way stratification; *J. Amer. Stat. Assn.*, 55, 105-124.

- CHAKRavarthy, I M (1954) On the problem of planning a multi-stage survey for multiple correlated characters, *Sankhya*, 14 211-216
- COCHRAN, W G (1961) Comparison of methods for determining stratum boundaries, *Bull Inter Stat Inst*, 38, (2), 345-358
- COCHRAN, W G (1963) *Sampling Techniques*, Second Edition, Chapters 5 and 54, John Wiley & Sons, New York
- DALENIUS, T (1950) The problem of optimum stratification—I, *Skand Alt*, 33, 203-213
- DALENIUS, T and GURNEY, M (1951) The problem of optimum stratification—II, *Skand Alt*, 34 133-148
- DALENIUS, T (1952) The problem of optimum stratification in a special type of design, *Skand Alt*, 35, 61-70
- DALENIUS, T (1953) Multivariate sampling problem, *Skand Alt*, 36, 92-122.
- DALENIUS, T (1957) *Sampling in Sweden*, Almqvist & Wiksell, Stockholm
- DALENIUS T and HODGES, J L (Jr) (1957) The choice of stratification points *Skand Alt*, 40, 198-203
- DALENIUS, T and HODGES, J L (Jr) (1959) Minimum variance stratification, *J Amer Stat Assn*, 54 88-101
- DURBIN, J (1959) Review of the book *Sampling in Sweden* *J Roy Stat Soc*, (A), 122, 246-248
- EKMAN, G (1959) An approximation useful in univariate stratification, *Ann Math Stat*, 30, 219-229
- EVANS, W D (1951) On stratification and optimum allocations, *J Amer Stat Assn*, 46, 93-104
- GHOOSH, S P (1958) A note on stratified random sampling with multiple characters, *Bull Cal Stat Assn*, 8, 81-90
- GOODMAN, R and KISH, L (1950) Controlled selection—A technique in probability sampling, *J Amer Stat Assn*, 45, 350-372
- HANSEN, M H, HURWITZ W N and MADOW, W G (1953) *Sample Survey Methods and Theory*, Volumes I & II, Chapter 5, John Wiley & Sons, New York
- HESS, I, RIEDEL, D C, and FITZPATRICK, T B (1961) *Probability Sampling of Hospitals and Patients* The University of Michigan Ann Arbor
- HESS, I, SETHI, V K and BALAKRISHNAN, T R (1966) Stratification—A practical investigation, *J Amer Stat Assn*, 61, 74-90
- MAHALANOBIS, P C (1944) On large-scale sample surveys, *Phil Trans Roy Soc*, 231, (B), 329-451
- MAHALANOBIS, P C (1952) Some aspects of the design of sample surveys, *Sankhya*, 12, 1-7
- MURTHY, M N (1962) Variance and confidence interval estimation *Sankhya*, 24, (B), 1-12
- NEYMAN, J (1934) On the two different aspects of the representative method, *J Roy Stat Soc*, 97, 558-625
- SETHI, V K. (1963) A note on optimum stratification of populations for estimating the population means, *Aust J Stat*, 5, 20-33.

STUART, A. (1954): A simple presentation of optimum sampling results; *J. Roy. Stat. Soc., (B)*, 16, 238-241.

SUKHATME, P. V. (1935): Contributions to the theory of the representative method; *J. Roy. Stat. Soc., Supplement*, 2, 253-268.

WILLIAMS, W. H. (1962): On the variance of an estimator with post-stratification; *J. Amer. Stat. Assn.*, 57, 622-627.

COMPLEMENTS AND PROBLEMS

7.1 For a socio-economic survey, all the villages in a region including the uninhabited ones were grouped into 4 strata on the basis of their altitude above sea-level and population density and from each stratum 10 villages were selected with srswr. The data on number of households in each of the sample villages are given in Table. 7.9.

TABLE 7.9. NUMBER OF HOUSEHOLDS FOR 40 SAMPLE VILLAGES.

stratum sr. no.	total no. of villages	total number of households in sample villages									
		1	2	3	4	5	6	7	8	9	10
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1.	1411	43	84	98	0	10	44	0	124	13	0
2.	4705	50	147	62	87	84	158	170	104	56	160
3.	2558	228	262	110	232	139	178	334	0	63	220
4.	14997	17	34	25	34	36	0	25	7	15	31

- (i) Estimate the total number of households (H) unbiasedly and estimate its rse.
- (ii) Examine whether there has been any gain due to use of stratification as compared to unstratified srswr.

(iii) Compare the efficiency of the present allocation with that of the optimum allocation keeping the total sample size fixed.

7.2 Using the data given in Table 7.10 (p.288) and considering the size classes as strata, compare the efficiencies of the following alternative allocations of a sample of 3000 factories for estimating the total output. The sample is to be selected with srs wr within each stratum :

- (a) proportional allocation;
- (b) allocation proportional to total output; and
- (c) optimum allocation.

7.3 A population of 112 villages has been divided into 3 strata having 51, 37 and 24 villages respectively on the basis of the type of available auxiliary information. We have a sample of 6 villages selected with srs wr from the first stratum, a ppsswr sample of 5 villages from the second stratum and two linear systematic samples (of four villages each) selected without replacement from the third stratum. For each selected village, the total area under wheat (y) is observed. The observed values and other relevant information are given in Table 7.11.

TABLE 7.10 DATA ON DISTRIBUTION OF FACTORIES BY NUMBER OF WORKERS AND ON AVERAGE OUTPUT AND STANDARD DEVIATION

sr no	size class no of workers	no of factories	output per factory (in 000 Rs)	standard deviation (in 000 Rs)
(1)	(2)	(3)	(4)	(4)
1	1 — 49	18260	100	80
2	50 — 99	4315	250	200
3	100 — 249	2233	500	600
4	250 — 999	1057	1760	1900
5	1000 & above	567	2250	2500

TABLE 7.11 AREA UNDER WHEAT y FOR ALL THE SAMPLE VILLAGES AND CULTIVATED AREA x FOR SAMPLE VILLAGES OF STRATUM 2

sample village	stratum 1 y	stratum 2		stratum 3 (y)	
		x	y	sample 1	sample 2
(1)	(2)	(3)	(4)	(5)	(6)
1	75	729	247	427	335
2	101	617	238	326	412
3	5	870	359	481	503
4	78	305	129	445	348
5	78	569	223	—	—
6	45	—	—	—	—

(total cultivated area in stratum 2 = 26 912 acres x and y are in acres)

(i) Estimate the total area under wheat in each stratum separately and also in all the 3 strata taken together

(ii) Obtain estimates of rses of these estimates estimating unbiasedly their variances

7.4 In a demographic survey it is proposed to have stratified sampling using the districts in a region as strata. The relevant data are given in Table 3.8 of Chapter 3 (p 91).

(i) Assuming the cost of enumeration and tabulation per person is 1/4th of a rupee and the overhead cost to be Rs 10 000 determine the optimum values of n_s s that would minimize the sampling variance of the estimator of the overall population mean for a given expected total cost of Rs 80 000 when villages are selected with srswr from each stratum.

(ii) For the same value of the total sample size n obtained in (i) find the values of n_s s when the allocation is made in proportion to $N_s \sigma_s$ and obtain the cost-efficiency of this procedure as compared to that of (i).

7.5 A survey is to be conducted for estimating the total number of literates in a town having three communities, some particulars of which are given in Table 7.12 based on the results of a pilot study.

TABLE 7.12. A ROUGH IDEA OF THE TOTAL NUMBER OF PERSONS AND PROPORTIONS OF LITERATES.

community	total number of persons	percentage of literates
1	60 000	40
2	10 000	80
3	30 000	60

(i) Treating the communities as strata and assuming srswr in each stratum, allocate a total sample size of 2000 persons to the strata in an optimum manner for estimating the overall proportion of literates in the town.

(ii) Estimate the efficiency of stratification as compared to unstratified sampling.

7.6 The analysis of variance of a population of 340 villages divided into 4 unequal strata is given in Table 7.13. Calculate the efficiency of stratification with proportional allocation and with srswr in each stratum as compared to unstratified sampling for estimating the area under wheat.

TABLE 7.13. ANALYSIS OF VARIANCE FOR AREA UNDER WHEAT.

source of variation	degrees of freedom	sum of squares	mean square (3)/(2)
(1)	(2)	(3)	(4)
between strata	3	$\Sigma_s N_s (\bar{Y}_s - \bar{Y})^2$	5400
within strata	336	$\Sigma_s \Sigma_{st} (Y_{st} - \bar{Y}_s)^2$	24
total	339	$\Sigma_s \Sigma_{st} (Y_{st} - \bar{Y})^2$	71.58

7.7 Suppose a region is divided into two sub-regions having 1000 and 1500 persons and the proportions of workers in manufacturing industries in the two sub-regions are likely to be about 30% and 70% respectively. Determine the total sample size required for estimating the overall population proportion within 5% of the true value with 95% confidence, if srswr is adopted in the two strata (sub-regions) using optimum allocation of the total sample size.

7.8 Assuming the number of units in the strata to be equal, show that in stratified sampling with srswr and with equal allocation, the sampling variance V_{et} of the estimator of \bar{Y} can be expressed as

$$V_{et} = V_r - \frac{1}{Kn} \sum_{s=1}^K (\bar{Y}_s - \bar{Y})^2,$$

where V_r is the variance in the case of unstratified srswr in estimating \bar{Y} .

7.9 In stratified srs wr the surveyor allocated the total sample size of n units by mistake in proportion to $N_s \sigma_s^2$ instead of the usual optimum allocation proportional to $N_s \sigma_s$. How does this allocation compare with proportional and optimum allocations?

7.10 Suppose the objective is to estimate the difference between the rates of incidence of a particular disease in two villages, one a model village having N_1 persons and the other a neighbouring village having N_2 persons. Let P_1 and P_2 be the proportions of persons having the disease and let C_1 and C_2 be the average costs of medically examining a person in the two villages. Assuming the cost to be fixed at C , determine the optimum allocation of the total sample size to the two villages when srswr is adopted in each village.

7.11 Suppose t_{s1} and t_{s2} are unbiased estimates of the s th stratum total ($s = 1, 2, \dots, K$) based on 2 independent samples. Show that the following two estimators are unbiased for the variance of the combined estimator $\Sigma_s(t_{s1} + t_{s2})/2$ and compare their variances :

$$(i) v_1 = \Sigma_s(t_{s1} - t_{s2})^2/4$$

and

$$(ii) v_2 = (\Sigma_s t_{s1} - \Sigma_s t_{s2})^2/4$$

(Murthy, M. N., *Sankhya*, 24, (B), (1962), 1-12)

7.12 A population of N units is divided into two strata of sizes N_1 and N_2 units and samples of n_1 and n_2 units are selected from each of the two strata with srswr. Show that the efficiency of the estimator of \bar{Y} in this case compared to that of optimum allocation of the total sample size $(n_1 + n_2)$ is not less than $4t/(1+t)^2$, where $t = n_1 n_2' / n_2 n_1'$, n_1' and n_2' being the optimum allocation to the two strata.

(Cochran, W. G., *Sampling Techniques*, (1953), Ch 5, p 79)

7.13 In stratified srs wr for estimating the overall population mean, obtain the sampling variances for the estimators in case of (i) proportional allocation and (ii) optimum allocation based on a total sample size of n units, assuming that N_s is large enough for $N_s/(N_s - 1)$ to be approximately unity. Compare these variances with that of the sample mean based on an unstratified sample of n units drawn with srs wr.

7.14 Suppose a population of N units is divided into K strata at random such that the s th stratum has a prespecified number of units N_s , $s = 1, 2, \dots, K$ and within each stratum srs wr is adopted using proportional allocation. Show that the variance of the estimator of the overall population mean in this case would be equal to that of the sample mean in the case of unstratified srs wr with the same overall sample size.

7.15 Assuming the population of N units to be drawn from super populations with the following models, compare the expected variances of the estimators of \bar{Y} (a) obtained by stratifying the population of N units into n strata of equal number of

units and selecting one unit from each stratum with srs, and (b) based on a systematic sample of n units assuming N to be a multiple of n :

$$(i) E(Y_i) = \bar{Y}, V(Y_i) = \sigma_i^2, \text{Cov}(Y_i, Y_{i'}) = 0, i' \neq i;$$

$$(ii) E(Y_i) = \alpha + \beta i, V(Y_i) = \sigma^2, \text{Cov}(Y_i, Y_{i'}) = 0, i' \neq i.$$

where $i, i' = 1, 2, \dots, N$,

(Cochran, W. G., *Sampling Techniques* (1963), Ch. 8, 215-217).

7.16 Assuming the finite population of N units to be drawn from a super-population with the model

$$E(Y_i|X_i) = aX_i, V(Y_i|X_i) = \sigma^2 X_i^g \text{ and } \text{Cov}(Y_i, Y_{i'}|X_i, X_{i'}) = 0, i' \neq i,$$

show that in stratified sampling where the probability of inclusion of a unit in the sample is proportional to its size, the optimum allocation of the total sample size to the strata, which minimizes the expected variance of the estimator

$$\hat{Y} = \sum_{s=1}^K \sum_{i=1}^{n_s} \frac{y_{st}}{\pi_{st}},$$

where π_{st} is the probability of inclusion in the sample of the i -th sample unit in the s -th stratum, is given by

$$n_s = n \sqrt{X_s \sum_{i=1}^{N_s} X_{st}^{g-1}} / \sqrt{\sum_{t=1}^K \sum_{i=1}^{N_s} X_{st}^{g-1}},$$

where X_s is the s -th stratum total for the size measure x . It may be noted that when $g = 2$, the optimum allocation becomes proportional to X_s .

(Rao, T. J., (1966) unpublished).

7.17. In forming two strata in an optimum manner such that the second stratum consisting of N_2 large units is completely enumerated and a sample of $n_1 (= n - N_2)$ units is selected with srs wr from the N_1 smaller units in the first stratum, show that the optimum point of stratification is given by

$$y = \bar{Y}_1 + \sigma_1 \sqrt{N_1/n_1},$$

where \bar{Y}_1 and σ_1 are the mean and the standard deviation of the first stratum.

(Dalenius, T., *Skand. Akt.*, 35, (1952), 61-70).

7.18. Suppose a population with a variable y having the probability density function

$$f(y) = e^{-y}, (y > 0),$$

is divided into two strata, stratum 1 defined by $y \leq y_0$ and stratum 2 by $y > y_0$. Derive the variance of the estimator of \bar{Y} assuming proportional allocation and srswr in each stratum. Find the optimum value of y_0 which will minimize the variance and evaluate the optimum variance.

(Dalenius, T., *Skand. Akt.*, 33, (1950), 203-213).

7.19 Derive the results (7.80) and (7.81) given in Sub-section 7.10a regarding ops for optimum and equal allocations.

(Dalenius, T., *Sampling in Sweden*, (1957), Ch. 7, p. 168).

7.20 From the results (7.82) show that if $V(\hat{X}_{st})$ is inversely proportional to K^2 ,

$$V(\hat{\bar{Y}}_{st}) \geq V(\hat{\bar{Y}}_{us}) \left\{ \frac{p^2}{K^2} + (1-p^2) \right\}$$

(Cochran, W G, *Sampling Techniques*, (1963), Ch 5A, p 134)

7.21 In the case of stratified sampling schemes where one unit is selected from each stratum, the sampling variance is usually estimated by adopting the *method of collapsed strata*. This method consists in pairing the strata to form collapsed strata and estimating the sampling variance as if two units had been selected from each collapsed stratum.

Suppose the proportion of units (W_s) is the same for each of the two strata forming the s th pair ($s = 1, 2, \dots, K/2$). Assuming ars within the strata, show that the variance estimator

$$v(\hat{\bar{Y}}_{st}) = \sum_{s=1}^{K/2} W_s^2 (y_{s1} - y_{s2})^2,$$

where y_{s1} and y_{s2} are the values of units selected from the two strata forming the s th collapsed stratum, over estimates $V(\hat{\bar{Y}}_{st})$ and that the bias is small when the two strata forming the s th pair have approximately the same mean ($s = 1, 2, \dots, K/2$).

(Cochran W G, *Sampling Techniques*, (1963), Ch 5A, p 141)

7.22 If the proportions of units (W_{s1} and W_{s2}) for the two strata forming the s th collapsed stratum (Problem 7.21) are not the same for all s , consider the variance estimator

$$v(\hat{\bar{Y}}_{st}) = \sum_{s=1}^{K/2} (W_{s1}^2 y_{s1} - W_{s2}^2 y_{s2})(y_{s1} - y_{s2})$$

and obtain its bias

(Seth, G R, *J Ind Soc Agr Stat*, 18, (1966), 1-3)

7.23 A sample of n units is drawn with ars wot and two post strata are formed at the estimation stage. There are two possibilities

- (i) each post stratum contains at least one sample unit, and
- (ii) one of the post strata is empty, that is, contains no sample unit

Consider the estimators

$$(a) \hat{\bar{Y}} = W_1 \bar{y}_1 + W_2 \bar{y}_2 \quad \text{and} \quad (b) \hat{\bar{Y}}^* = \alpha D_1 y_1 + (1-\alpha) D_2 \bar{y}_2,$$

where W_1 and W_2 ($= 1 - W_1$) are the proportions of units and \bar{y}_1 and \bar{y}_2 are the sample means for the two post strata, α is 1 or 0 according as stratum 2 or 1 is empty and $D_1 = W_1/P_1$ and $D_2 = W_2/P_2 = (1 - W_1)/(1 - P_1)$, P_1 being the conditional probability that stratum 2 is empty given that possibility (i) has occurred. Show that the estimators (a) and (b) are conditionally unbiased given that possibility (i) or (ii) has occurred respectively. Find the bias and the variance of

$$\hat{\bar{Y}} = \lambda \hat{\bar{Y}} + (1-\lambda) \hat{\bar{Y}}^*,$$

where λ is 1 or 0 according as possibility (i) or (ii) has occurred

(Fuller, W A, *J Amer Stat Assn*, 61, (1966), 1172-1183)

Cluster Sampling

8.1 NEED FOR CLUSTER SAMPLING

Cluster sampling consists in forming suitable clusters of units and surveying all the units in a sample of clusters selected according to an appropriate sampling scheme. The advantages of cluster sampling from the point of view of cost arise mainly due to the fact that collection of data for nearby units is easier, faster, cheaper and more convenient than observing units scattered over a region. For instance, in a population survey it may be cheaper to collect data from all persons in a sample of households than from a sample of the same number of persons selected directly from all the persons. Similarly, it would be operationally more convenient to survey all households situated in a sample of areas such as villages than to survey a sample of the same number of households selected at random from a list of all households. Another example of the utility of cluster sampling is provided by crop surveys, where locating a randomly selected farm or plot (a parcel of land) requires a considerable part of the total time taken for the survey, but once the plot is located, the time taken for identifying and surveying a few neighbouring plots will generally be only marginal.

Because of its operational convenience and the possible reduction in cost, cluster sampling is resorted to in many surveys, using *mutually exclusive* or *overlapping* clusters formed by grouping nearby units or units which can be conveniently observed together. In general, for a given total number of sampling units, cluster sampling

is less efficient than sampling of individual units from the view point of sampling variance as the latter is expected to provide a better cross section of the population than the former due to the usual tendency of units in a cluster to be similar. In fact the sampling efficiency of cluster sampling is likely to decrease with increase in cluster size. However cluster sampling is operationally more convenient and less costly than sampling of units directly due to the possible saving in time for journey, identification contact, etc., and hence in many practical situations the loss in sampling efficiency is likely to be offset by the reduction in cost.

In a general sense any system of sampling may be regarded as a kind of *cluster sampling* since in every sampling scheme the units are conceptually grouped to form samples (clusters) and one of them is selected with a certain specified probability. For instance, systematic sampling may be considered a particular case of cluster sampling since in this case the population is divided into a number of clusters each cluster consisting of units distributed at a fixed interval (systematically) over the whole population and one such cluster is selected at random. But by *cluster sampling* is usually meant sampling of clusters of units formed by grouping neighbouring units or units which can be conveniently surveyed together. It may be noted that the various sampling procedures namely srs systematic sampling pps and stratified sampling discussed in the earlier chapters can be applied to sampling of clusters by treating the clusters themselves as sampling units.

8.2 SAMPLING OF EQUAL CLUSTERS

Let us first consider the case of clusters, which are mutually exclusive and have an equal number of units. Though the size of natural clusters such as villages (clusters of households or persons) or branches of trees (clusters of leaves, flowers fruits) usually varies over clusters it is possible to have equal clusters when clusters are artificially formed. For instance in a crop survey, we may consider clusters of two or more plots or other area units of a given size and

shape as clusters and in a household survey two or more neighbouring households may be grouped to form clusters. Similarly, in a production process in an industry, the number of items produced at regular intervals of time may be the same and the production at different intervals of time can be considered to constitute the clusters.

8.2a SAMPLING OF ONE CLUSTER

Suppose a finite population of NM units is divided into N mutually exclusive clusters of M units each and one cluster is selected with srs for estimating the population mean

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i, \quad \left(\bar{Y}_i = \frac{1}{M} \sum_{j=1}^M Y_{ij} \right), \quad \dots \quad (8.1)$$

where Y_{ij} is the value of the j -th unit in the i -th cluster. An unbiased estimator of \bar{Y} is clearly given by the sample cluster mean. Its variance is given by

$$V(\hat{\bar{Y}}_c) = \frac{1}{N} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2 = \sigma_b^2, \quad \dots \quad (8.2)$$

where the subscript c denotes that the estimator is based on a cluster sample and σ_b^2 stands for the between-cluster variance. Sampling of one cluster is being considered here mainly to bring out the implications of using cluster sampling from the view-point of sampling variance. It may be noted that it is not possible to estimate the variance of the estimator unbiasedly on the basis of a sample of one cluster just as it is not possible to estimate unbiasedly the variance of a systematic sample estimator on the basis of a single sample. The question of sampling $n (\geq 2)$ clusters is considered in Sub-section 8.2b.

Comparing (8.2) with the variance of the sample mean \bar{y} based on M units drawn from NM units with srswr, namely,

$$V(\hat{\bar{Y}}_r) = \frac{\sigma^2}{M}, \quad \sigma^2 = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M (\bar{Y}_{ij} - \bar{Y})^2, \quad \dots \quad (8.3)$$

where the subscript r denotes srswr and σ^2 is the total variance, we find that the sampling efficiency of cluster sampling as compared to srswr is

$$E_s = \frac{1}{M} \frac{\sigma^2}{\sigma_b^2} = \frac{1}{M} \left\{ 1 + \frac{\sigma_w^2}{\sigma_b^2} \right\}, \quad (8.4)$$

since $\sigma^2 = \sigma_b^2 + \sigma_w^2$, where σ_w^2 is the within cluster variance given by

$$\sigma_w^2 = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2 \quad \text{From this, it can be seen that cluster}$$

sampling will be more efficient than srswr only when the total variance is greater than M times the between cluster variance σ_b^2 , that is, when the within cluster variance σ_w^2 is greater than $(M-1)$ times σ_b^2 . This is not likely to be the case since σ_b^2 will usually be larger due to the within cluster homogeneity. Hence, purely from the point of view of sampling variance, cluster sampling is generally less efficient than srs, though there may be special situations where the former may be as efficient as or even more efficient than the latter.

The variance in (8.2) and the efficiency E_s can be expressed in terms of the intraclass correlation coefficient ρ_c between pairs of units within clusters. For, σ_b^2 can be written as

$$\begin{aligned} \sigma_b^2 &= \frac{1}{N} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2 = \frac{1}{NM^2} \sum_{i=1}^N \left\{ \sum_{j=1}^M (Y_{ij} - \bar{Y}) \right\}^2 \\ &= \frac{1}{NM^2} \left[\sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y})^2 + \sum_{i=1}^N \sum_{j=1}^M \sum_{j' \neq j}^M (Y_{ij} - \bar{Y})(Y_{ij'} - \bar{Y}) \right]. \end{aligned}$$

This may be written as

$$\sigma_b^2 = \frac{\sigma^2}{M} \{1 + (M-1)\rho_c\}, \quad (8.5)$$

where

$$\rho_c = \frac{\sum_{i=1}^N \sum_{j=1}^M \sum_{j \neq j'}^M (Y_{ij} - \bar{Y})(Y_{ij'} - \bar{Y})}{NM(M-1)\sigma^2}.$$

Hence,

$$E_s = \frac{1}{1+(M-1)\rho_c}. \quad \dots \quad (8.6)$$

Substituting $(\sigma^2 - \sigma_{iv}^2)$ for σ_b^2 in (8.5), we get

$$\rho_c = 1 - \frac{M-1}{M-1} \frac{\sigma_{iv}^2}{\sigma^2}$$

Noting that $0 \leq \sigma_{iv}^2 / \sigma^2 \leq 1$, we find that ρ_c lies in the range $\{-1/(M-1)\}$ to 1.

The expression for E_s shows that cluster sampling will be more efficient than srs wr only if ρ_c is negative. But in practice ρ_c is usually positive when nearby units are grouped to form clusters and hence cluster sampling is generally less efficient than srs. Though usually ρ_c decreases with increase in M , the efficiency of cluster sampling declines, because the factor $(M-1)\rho_c$ generally increases with increasing cluster size. Here it may be mentioned that there may be situations where ρ_c is likely to be negative. Possible examples of such situations are the sex and the age compositions in households and villages (cf. Problem 8.7, p.315).

8.2b SAMPLING OF n CLUSTERS

In sampling n clusters with srs wr, the mean of the sample cluster means is an unbiased estimator of \bar{Y} , that is,

$$\hat{\bar{Y}}_c = \frac{1}{n} \sum_{i=1}^n \bar{y}_i = \frac{1}{nM} \sum_{i=1}^n \sum_{j=1}^M y_{ij}$$

where \bar{y}_i is the i -th sample cluster mean. The variance and an unbiased variance estimator of $\hat{\bar{Y}}_c$ are given by

$$V(\hat{\bar{Y}}_c) = \sigma_b^2/n, \quad \dots \quad (8.7)$$

and

$$v(\hat{\bar{Y}}_c) = \frac{s_b^2}{n}, \quad s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \hat{\bar{Y}}_c)^2. \quad \dots \quad (8.8)$$

Evidently, the efficiency of cluster sampling in this case as compared to a sample of the same size (nM units) selected with srswr is the same as that given in (8.4) and (8.6).

If n clusters are selected with srs wor instead of with srswr, then the sampling variance and the variance estimator of \hat{Y}_c are

$$V(\hat{Y}_c) = \frac{N-n}{N-1} \frac{\sigma_b^2}{n} = \frac{N-n}{N-1} \frac{\sigma^2}{nM} \{1 + (M-1)\rho_b\}, \quad (8.9)$$

and

$$v(\hat{Y}_c) = (1-f)s_b^2/n, \quad (8.10)$$

where f is the sampling fraction and s_b^2 is as defined in (8.8). The efficiency of cluster sampling in this case as compared to sampling of nM units with srs wor is given by

$$E_s = \frac{M(N-1)}{NM-1} \frac{1}{1 + (M-1)\rho_b} \quad (8.11)$$

It may be noted that the efficiency of sampling one cluster as compared to sampling of M units with srs wor is also the same as (8.11).

8.2c ESTIMATION OF EFFICIENCY

Given a sample of n clusters of M units selected with srs wor, it is of interest to estimate from the same sample itself the efficiency of cluster sampling as compared to that of direct sampling of units with srs wor. The variance of the sample mean based on nM units in the latter case is

$$V(\hat{Y}_r) = \frac{NM-nM}{NM-1} \frac{\sigma^2}{nM} \quad . \quad (8.12)$$

Since $V(\hat{Y}_c)$ is unbiasedly estimated by (8.10), it is sufficient if we can obtain an unbiased estimator of (8.12) on the basis of the cluster sample to be able to estimate the relative efficiency.

$$E_s = V(\hat{Y}_r)/V(\hat{Y}_c) \quad (8.13)$$

An unbiased estimator of $\frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M Y_{ij}^2$ is given by $\frac{1}{nM} \sum_{i=1}^n \sum_{j=1}^M y_{ij}^2$

and that of \bar{Y}^2 is $\hat{\bar{Y}}_c^2 - v_c(\hat{\bar{Y}}_c)$, since $E(\hat{\bar{Y}}_c^2) = \bar{Y}^2 + V(\hat{\bar{Y}}_c)$, where the subscript c to the variance estimator denotes that the estimator is based on a cluster sample. Hence, an unbiased estimator of $V(\hat{\bar{Y}}_r)$ based on a cluster sample is

$$v_c(\hat{\bar{Y}}_r) = \frac{NM-nM}{NM-1} \frac{1}{nM} \left[\frac{1}{nM} \sum_{i=1}^n \sum_{j=1}^M y_{ij}^2 - \hat{\bar{Y}}_c^2 + v_c(\hat{\bar{Y}}_c) \right] \dots (8.14)$$

The efficiency is estimated by substituting in (8.13) the estimators of $V(\hat{\bar{Y}}_r)$ and $V(\hat{\bar{Y}}_c)$ based on the cluster sample.

8.3 OPTIMUM CLUSTER SIZE

We have seen that for a given total sample size in terms of units, the sampling variance increases with cluster size and decreases with increasing number of clusters, while the cost decreases with increasing cluster size and increases with number of clusters. Hence, in practice it is necessary to strike a balance between these two opposing points of view by finding the optimum values for the cluster size and the number of sample clusters, which would minimize the sampling variance for a fixed cost or alternatively minimize the cost for a specified sampling variance. This problem is considered in this section.

8.3a VARIANCE FUNCTION

The sampling variance (8.9) of the estimator of \bar{Y} based on a sample of n clusters is a function of (i) N , the number of clusters in the population, (ii) σ^2 , the inherent variability in the population, (iii) ρ_c , the intraclass correlation coefficient, which depends on cluster size M , and (iv) n , the number of sample clusters. The main difficulty that arises in studying the behaviour of the variance and in determining the optimum cluster size is that ρ_c is not an explicit function of M and that the behaviour of ρ_c with increase in M varies from population to population. However, in practice, the behaviour of

the variance function with increase in cluster size can be studied on the basis of extensive empirical studies. The results of some empirical studies are given in this section and in Section 8.6.

Mahalanobis (1940a, 1940b, 1942, 1944) has considered in detail the question of determining the optimum cluster size in case of crop surveys from the points of view of both cost and variance, and on the basis of extensive empirical studies carried out during the period 1937-1941, he has shown that the variance of the estimator of crop proportion P in a region based on a sample of one cluster (a square shaped area unit of x acres, termed *grid*) can be approximated by

$$V(\hat{P}) = PQ/(bx)^g, \quad (8.15)$$

where b and g are constants. The variance relative to the binomial form PQ becomes $1/(bx)^g$ or a/x^g , where a and g are constants. The results of empirical studies, on which the relative variance function a/x^g is based are presented in Table 8.1 (cf Problem 8.1, p. 314). H. F. Smith (1938), Hansen and Hurwitz (1942), Jessen (1942) and Sukhatme (1953) have also studied the question of sampling efficiency of cluster sampling.

TABLE 8.1 OBSERVED AND GRADUATED VALUES
OF THE VARIANCE RELATIVE TO PQ

grid size in acres	values of V/PQ	
	observed	graduated
(1)	(2)	(3)
1	0.2494	0.2562
2	0.1821	0.1761
3	0.1501	0.1414
4	0.1145	0.1210
6	0.1001	0.0972
9	0.0749	0.0780

Source Mahalanobis P. C (1942) *General Report on the Sample Census of Area under Jute in Bengal, 1941*, Indian Central Jute Committee

It may be mentioned that numerous studies have also been undertaken to examine the possibility of bias due to border effects and the behaviour of the sampling variance of the estimator of crop yield rate for different shapes and sizes of *sample cuts* (Sukhatme, 1947; Mahalanobis and Sengupta, 1951).

8.3b COST FUNCTION

Since the cost of survey depends both on the cluster size and number of sample clusters, the simplest type of cost function is of the form

$$C = C_0 + nC_1 + nMC_2, \quad \dots \quad (8.16)$$

where C is the total cost, C_0 the overhead cost, C_1 the cost per cluster for preliminary operations (such as journey, identification, contact, etc.) involved in conducting the survey in a cluster and C_2 is the cost of surveying one unit. In general, C_2 is expected to be considerably less than C_1 . Though the cost function (8.16) is simple and serves as a rough approximation to the pattern of cost structure in cluster sampling, sometimes it would be desirable to examine the components of cost more carefully and build up more realistic cost functions depending on particular situations. For instance, the expected value of the shortest distance between n random points in a region is deduced to be proportional to $\sqrt{n}-1/\sqrt{n}$ and this result has also been empirically verified in case of crop surveys (Mahalanobis, 1940a; Jessen, 1942). Hence, if n is expected to be large, then the cost function may be taken as

$$C = C_0 + A\sqrt{n} C_{11} + nC_{12} + nMC_2, \quad \dots \quad (8.17)$$

where C_{11} is the cost of journey, C_{12} is the cost of other preliminary operations such as identification, contact, etc. involved in surveying a cluster and A is a constant depending on the area of the region.

8.3c DETERMINATION OF CLUSTER SIZE

The optimum values of the cluster size and the number of sample clusters can be determined so as to (i) minimize the sampling variance of the estimator for a fixed cost or (ii) minimize the cost while ensuring a specified value for the sampling variance. When the cost is fixed, the procedure of obtaining the optimum values consists in expressing the value of n in terms of M from the cost function (8.16 or 8.17) and finding that value of M , which minimizes the sampling variance, after substituting this value of n in the variance function. For instance, we find, from (8.16) that when C is fixed at C_0 the value of n is given by

$$n = (C' - C_0) / (C_1 + MC_2)$$

and substituting this in the sampling variance (8.9), we get

$$V(\hat{Y}_c) = \left\{ \frac{N(C_1 + MC_2)}{(C - C_0)} - 1 \right\} \frac{\sigma_b^2}{(N-1)} \quad (8.18)$$

When N is large compared to n or when sampling is done with replacement, (8.18) reduces to

$$V(\hat{Y}_c) = \frac{(C_1 + MC_2)}{(C - C_0)} \sigma_b^2 = \frac{C_1 \sigma^2}{C - C_0} \left\{ \left(1 + M \frac{C_2}{C_1} \right) \left(\frac{1 + (M-1)\rho_c}{M} \right) \right\} \quad (8.19)$$

From (8.19) we see that the optimum value of M which minimizes the variance can be obtained by plotting the values of the expression within curly brackets against M and by determining that value of M for which the graph has the minimum value. The value of ρ_c for different values of M may be obtained by conducting empirical studies on the data collected in some previous census or pilot survey for the characteristic under study or for some other suitable auxiliary characteristic and the values of C_1 and C_2 may be arrived at on the basis of a previous survey or by conducting a pilot study for this purpose. Similarly, when the sampling variance is specified, the values of n required for ensuring the specified sampling variance are to be found

through empirical studies for different cluster sizes and then that pair of values of n and M , which gives rise to the minimum cost, is to be determined. Two examples are given here to illustrate the procedure of determining the optimum cluster size.

Example 1

The first example relates to sampling of clusters of plots for estimating the crop acreage and the problem here is to find that cluster size which minimizes the sampling variance for a given cost. For this purpose the plot-wise data on acreage under paddy obtained in a special harvest survey in 1955-56 for the village Mayurpukur in Bankura district of West Bengal are used. Mutually exclusive clusters were formed by grouping the *survey numbers* of plots. The values of σ_b^2/σ^2 are found for different cluster sizes ($M = 2, 5, 10, 20$) and these are used in determining the optimum cluster size when the cost ratio C_2/C_1 is taken as 0.1, 0.3, 0.5 and 1.0. The values of

$$\lambda(M) = \left(1 + M \frac{C_2}{C_1}\right) \frac{\sigma_b^2}{\sigma^2},$$

which is proportional to the actual sampling variance (S.19), are given in Table 8.2 for different values of M and C_2/C_1 . From this table, it can be seen that the optimum cluster size is about 10 plots when C_2/C_1 is 0.1, 0.3 and 0.5, and it is 2 plots when C_2/C_1 is 1.0. The values of ρ_c given in column (3) show that the decrease in ρ_c is substantial for initial increases in cluster size, but becomes marginal after a certain stage.

TABLE 8.2. DETERMINATION OF OPTIMUM CLUSTER SIZE (NUMBER OF PLOTS) FOR ESTIMATING ACREAGE UNDER PADDY.

cluster size	efficiency ($\sigma^2/M\sigma_b^2$) (%)	intraclass correlation ρ_c	value of $\lambda(M)$ when C_2/C_1 is			
			0.1	0.3	0.5	1.0
(1)	(2)	(3)	(4)	(6)	(6)	(7)
1	100.00	—	1.100	1.300	1.500	2.000
2	84.08	0.177	0.706	0.941	1.177	1.765
5	64.24	0.139	0.467	0.778	1.090	1.868
10	55.63	0.089	0.360	0.719	1.079	1.978
20	39.46	0.081	0.380	0.887	1.394	2.661

total number of plots : 851; geographical area : 900 acres; area under paddy : 299 acres.

Example 2

The second example relates to determination of the optimum cluster size such that the cost is minimum for ensuring a specified sampling variance. For this purpose the results of a study, conducted by Mahalanobis (1940b) to determine the optimum size of a grid (square shaped area unit) for estimating the area under jute in Bengal, are presented in Table 8.3. The sampling procedure here consists in locating at random a specified number of grids of equal size on the map of a region and taking the mean of the proportions of area under the crop in the sample grids as an estimate of the proportion of area under the crop in the region. From the table it is clear that the optimum grid size is about 4 acres for most of the different levels of coefficient of error, which is defined as $\sqrt{V/PQ}$, V being the sampling variance, P the crop proportion and $Q = 1 - P$.

TABLE 8.3 COST PER SQUARE MILE FOR DIFFERENT LEVELS OF CLUSTER SIZE AND COEFFICIENT OF ERROR

coefficient of error $100\sqrt{V/PQ}$	cost in rupees per square mile for grids of size (in acres)				
	1	4	9	16	36
(1)	(2)	(3)	(4)	(5)	(6)
100	1.32	1.22	1.24	1.25	1.36
90	1.38	1.29	1.30	1.33	1.45
80	1.47	1.37	1.40	1.43	1.59
70	1.59	1.49	1.53	1.57	1.78
60	1.77	1.67	1.74	1.80	2.07
50	2.06	1.97	2.06	2.16	2.54
40	2.58	2.49	2.65	2.81	3.41
30	3.66	3.59	3.90	4.19	5.25
20	6.61	6.64	7.37	8.06	10.45

(1 square mile = 640 acres)

Source : Mahalanobis, P. C. (1940b) : Report on the Sample Census of Jute, 1939; Indian Central Jute Committee.

8.3d USE OF OTHER SAMPLING SCHEMES

Though we have considered so far only the simplest scheme for sampling clusters, it may be useful in practice to select the clusters through some other sampling scheme, such as systematic sampling after a suitable arrangement of the clusters or sampling with pro-

bability proportional to the value of an auxiliary variable related to the study variable. The theory developed in the case of srs can be extended to the other sampling schemes. The possibility of using pps sampling scheme for selecting clusters of varying sizes is considered in some detail in Section 8.5.

8.4 ESTIMATION OF A PROPORTION

In this section we consider the question of estimation of P , the proportion of units in the population belonging to a specified category, on the basis of a sample of clusters of units. The expressions relating to the estimator, sampling variance, variance estimator and efficiency of cluster sampling, derived earlier for estimating \bar{Y} in Section 8.2, can be applied directly to the case of estimation of P .

Suppose a sample of n clusters is selected with srs w/o. Then an unbiased estimator of the overall population proportion P is given by

$$\hat{P}_c = \frac{1}{n} \sum_{i=1}^n p_i = \bar{p}, \quad \dots \quad (8.20)$$

where p_i is the proportion of units belonging to the specified category in the i -th sample cluster. The sampling variance of \hat{P}_c is

$$V(\hat{P}_c) = \frac{N-n}{N-1} \frac{\sigma_b^2}{n}, \quad \dots \quad (8.21)$$

where σ_b^2 is the variance between cluster proportions and is given by

$$\sigma_b^2 = \frac{1}{N} \sum_{i=1}^N (P_i - P)^2 = PQ - \frac{1}{N} \sum_{i=1}^N P_i Q_i. \quad \dots \quad (8.22)$$

Since $\sigma^2 = \sigma_b^2 + \sigma_w^2 = PQ$, the within-variance σ_w^2 in this case is given by $\frac{1}{N} \sum_{i=1}^N P_i Q_i$. The sampling variance in (8.21) can also be expressed in terms of the intraclass correlation coefficient ρ_c . That is

$$V(\hat{P}_c) = \frac{N-n}{N-1} \frac{PQ}{nM} \{1 + (M-1)\rho_c\}, \quad \dots \quad (8.23)$$

where

$$\rho_c = 1 - \frac{M}{M-1} \frac{1}{N} \sum_{i=1}^N \frac{P_i Q_i}{PQ} \quad (8.24)$$

It can be shown that an unbiased estimator of $V(\hat{P}_c)$ is given by

$$v(\hat{P}_c) = (1-f) \frac{s_b^2}{n}, \quad s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (p_i - \bar{p})^2 \quad (8.25)$$

The efficiency of cluster sampling as compared to srs w.r.t. given in (8.11) in this case becomes

$$E_s = \frac{\frac{N-1}{NM-1}}{\frac{NPQ}{NPQ - \sum_{i=1}^N P_i Q_i}} \quad (8.26)$$

An estimator of the total number of units belonging to the specified category can be got by simply multiplying \hat{P}_c by NM and the expressions for its sampling variance and variance estimator are $N^2 M^2$ times the corresponding expressions for \hat{P}_c .

8.5 VARYING CLUSTER SIZE

There are a number of situations where it is convenient to take certain naturally formed groups of units as clusters and in such cases the cluster size would in general vary from cluster to cluster. For instance households which are groups of persons and villages or urban blocks which are groups of households and persons, are usually considered as clusters for purposes of sampling because of operational convenience.

8.5a SIMPLE RANDOM SAMPLING

Suppose there are N clusters and let M_i be the number of units in the i th cluster ($i = 1, 2, \dots, N$). The overall population mean \bar{Y} is given by

$$\bar{Y} = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij} = \frac{1}{NM} \sum_{i=1}^N M_i \bar{Y}_i, \quad M = \frac{1}{N} \sum_{i=1}^N M_i \quad (8.27)$$

Suppose n clusters are selected with srs w/o r and all the units in the sample clusters are surveyed. An unbiased estimator of \bar{Y} is given by

$$\hat{\bar{Y}}_c = \frac{1}{NM'} \left\{ \frac{N}{n} \sum_{i=1}^n M_i \bar{y}_i \right\} = \frac{1}{n} \sum_{i=1}^n \left(\frac{M_i}{M'} \right) \bar{y}_i, \quad \dots \quad (8.28)$$

where \bar{y}_i is the mean of the i -th sample cluster, since $M_i \bar{y}_i$ is the total y_i for the i -th sample cluster and $\frac{N}{n} \sum_{i=1}^n y_i$ is an unbiased estimator of the population total \bar{Y} .

When the value of M_i is known only for the sample clusters and not for all the clusters, then \bar{Y} can be unbiasedly estimated by

$$\hat{Y}_c = \frac{N}{n} \sum_{i=1}^n M_i \bar{y}_i. \quad \dots \quad (8.29)$$

An estimator of \bar{Y} in this case may be taken as either

$$\hat{\bar{Y}}' = \frac{1}{n} \sum_{i=1}^n \bar{y}_i, \quad \dots \quad (8.30)$$

which is considered later in this section, or

$$\hat{\bar{Y}}'' = \sum_{i=1}^n M_i \bar{y}_i / \sum_{i=1}^n M_i, \quad \dots \quad (8.31)$$

which is a ratio of two random variables and hence it is, in general, biased for \bar{Y} and such estimators are considered in Chapter 10.

The sampling variance of $\hat{\bar{Y}}_c$ given in (8.28) and its unbiased variance estimator are

$$V(\hat{\bar{Y}}_c) = \frac{N-n}{N-1} \frac{\sigma_b'^2}{n}, \quad \sigma_b'^2 = \frac{1}{N} \sum_{i=1}^N \left(\frac{M_i}{M'} \bar{Y}_i - \bar{Y} \right)^2 \quad \dots \quad (8.32)$$

and

$$v(\hat{\bar{Y}}_c) = \frac{N-n}{Nn} s_b'^2, \quad s_b'^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{M_i}{M'} \bar{y}_i - \hat{\bar{Y}}_c \right)^2. \quad \dots \quad (8.33)$$

Variance as Function of ρ_c

Substituting the value $\frac{1}{M_t} \sum_{j=1}^{M_t} Y_{tj}$ for \bar{Y}_t in $\sigma_b'^2$, we get

$$NM'^2 \sigma_b'^2 = \sum_{t=1}^N \left(\sum_{j=1}^{M_t} Y_{tj} - M' \bar{Y} \right)^2.$$

Adding and subtracting $M_t \bar{Y}$ inside the brackets in $NM'^2 \sigma_b'^2$ and expanding, we get

$$\begin{aligned} NM'^2 \sigma_b'^2 &= \sum_{t=1}^N \sum_{j=1}^{M_t} (Y_{tj} - \bar{Y})^2 + \sum_{t=1}^N \sum_{j=1}^{M_t} \sum_{j' \neq j}^{M_t} (Y_{tj} - \bar{Y})(Y_{tj'} - \bar{Y}) \\ &\quad + \bar{Y}^2 \sum_{t=1}^N (M_t - M')^2 + 2\bar{Y} \sum_{t=1}^N M_t(M_t - M')(\bar{Y}_t - \bar{Y}). \end{aligned}$$

Noting that in this case the overall variance σ^2 and the intraclass correlation coefficient ρ_c are given by

$$\sigma^2 = \frac{1}{NM'} \sum_{t=1}^N \sum_{j=1}^{M_t} (Y_{tj} - \bar{Y})^2$$

and

$$\rho_c = \frac{\sum_{t=1}^N \sum_{j=1}^{M_t} \sum_{j' \neq j}^{M_t} (Y_{tj} - \bar{Y})(Y_{tj'} - \bar{Y})}{\sum_{t=1}^N M_t(M_t - 1) \sigma^2},$$

$V(\hat{\bar{Y}}_c)$ can be written as

$$\begin{aligned} V(\hat{\bar{Y}}_c) &= \frac{N-n}{N-1} \cdot \frac{1}{n} \left[\frac{\sigma^2}{M} \left\{ 1 + \sum_{t=1}^N \frac{M_t(M_t - 1)}{NM'} \rho_c \right\} \right. \\ &\quad \left. + \frac{\bar{Y}^2}{NM'^2} \sum_{t=1}^N (M_t - M)^2 + 2 \frac{\bar{Y}}{NM'^2} \sum_{t=1}^N M_t(M_t - M)(\bar{Y}_t - \bar{Y}) \right]. \dots (8.34) \end{aligned}$$

It may be noted that when the cluster sizes are equal, that is, when $M_t = M' = M$ for all t , the expression (8.34) reduces to the expression (8.9) derived earlier.

The estimator of \bar{Y} given in (8.30), namely,

$$\hat{\bar{Y}}' = \frac{1}{n} \sum_{t=1}^n \bar{y}_t,$$

is biased, since its expected value is given by

$$E(\hat{\bar{Y}}') = \frac{1}{N} \sum_{t=1}^N \bar{Y}_t = \bar{Y},$$

which, in general, is not equal to $\bar{Y} (= \bar{Y}/NM')$ and the bias is

$$B(\hat{\bar{Y}}') = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i - \frac{1}{NM'} \sum_{i=1}^N M_i \bar{Y}_i = -\frac{1}{M'} \text{Cov}(\bar{Y}_i, M_i).$$

This shows that the bias is expected to be small when M_i and \bar{Y}_i are not highly correlated. In such a case, it may be desirable to use this estimator, since, though biased, its mean square error, namely,

$$\begin{aligned} M(\hat{\bar{Y}}'_c) &= V(\hat{\bar{Y}}'_c) + B^2(\hat{\bar{Y}}'_c) \\ &= \frac{N-n}{N-1} \frac{1}{n} \frac{1}{N} \sum_{i=1}^N (\bar{Y}_i - \bar{Y}')^2 + \frac{1}{M'^2} \{\text{Cov}(\bar{Y}_i, M_i)\}^2, \dots \end{aligned} \quad (8.35)$$

is likely to be considerably less than the variance given in (8.32).

8.5b VARYING PROBABILITY SAMPLING

Since in many practical situations the cluster total for the estimation variable is likely to be positively correlated with the number of units in the cluster, it may be profitable to select the clusters with probability proportional to the number of units in the cluster instead of with equal probability, or to stratify first the clusters on the basis of their sizes and then to have srs within each stratum. Suppose n clusters are selected with ppswr, size being the number of units in the cluster, then an unbiased estimator of \bar{Y} is given by

$$\hat{\bar{Y}}_c = \frac{1}{n} \sum_{i=1}^n \bar{y}_i. \quad \dots \quad (8.36)$$

Since \bar{y}_i takes the values $\{Y_i\}$ with probabilities $\{M_i/NM'\}$, $i = 1, 2, \dots, N$, we have

$$E(\hat{\bar{Y}}_c) = \frac{1}{n} \sum_{i=1}^n E(\bar{y}_i) = \sum_{i=1}^N \bar{Y}_i \frac{M_i}{NM'} = \bar{Y},$$

and

$$V(\hat{\bar{Y}}_c) = \frac{1}{n} \frac{1}{NM'} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2 M_i. \quad \dots \quad (8.37)$$

An unbiased variance estimator can be derived by noting that unbiased estimators of $\sum_{i=1}^n \bar{Y}_i^2$, $\frac{M}{NM}$, and \bar{Y}^2 are given by $\frac{1}{n} \sum_{i=1}^n \bar{y}_i^2$ and $\hat{\bar{Y}}_c^2 - v(\hat{\bar{Y}}_c)$ respectively and solving for $v(\hat{\bar{Y}}_c)$ after substituting these estimators in (8.37). That is,

$$v(\hat{\bar{Y}}_c) = \frac{1}{n} \left\{ \frac{1}{n} \sum_{i=1}^n \bar{y}_i^2 - \hat{\bar{Y}}_c^2 + v(\hat{\bar{Y}}_c) \right\}$$

Hence, we get

$$v(\hat{\bar{Y}}_c) = \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\bar{Y}}_c)^2 \quad (8.38)$$

The efficiency of sampling n unequal clusters with ppswr as compared to selection of nM' units with srswr can be obtained by comparing (8.37) with the variance in the latter case, namely,

$$V(\hat{\bar{Y}}_r) = \frac{\sigma^2}{nM}, \quad \sigma^2 = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y})^2 \quad (8.39)$$

Expanding $V(\hat{\bar{Y}}_r)$ after adding and subtracting \bar{Y} , within the brackets in σ^2 , we get

$$\begin{aligned} V(\hat{\bar{Y}}_r) &= \frac{1}{nM} \frac{1}{NM} \left\{ \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2 M_i + \sum_{i=1}^N \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2 \right\} \\ &= \frac{1}{M} \left\{ V(\hat{\bar{Y}}_c) + \frac{\sigma_w^2}{n} \right\}, \end{aligned}$$

where σ_w^2 is the within cluster variance given by

$$\sigma_w^2 = \frac{1}{NM'} \sum_{i=1}^N M_i \sigma_i^2, \quad \sigma_i^2 = \frac{1}{M_i} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2$$

Hence, the efficiency of cluster sampling is given by

$$E_s = \frac{V(\bar{Y}_r)}{V(\hat{\bar{Y}}_c)} = \frac{1}{M'} \frac{1}{1 - (\sigma_w^2 / \sigma^2)} \quad (8.40)$$

To increase the efficiency further in the case of unequal cluster size, other sampling schemes, such as pps wr sampling and pps systematic sampling with a suitable arrangement of the clusters, may be used.

8.6 ILLUSTRATIVE EXAMPLES

In this section the results of an empirical study are given to illustrate the behaviour of the efficiency of cluster sampling with increase in cluster size for different sampling schemes. This study relating to the estimation of the total acreage under autumn paddy in a village is based on the same plot-wise data used in Sub-section 3.3c for determining the optimum cluster size for a fixed cost. The relative variances (V/Y^2) for the estimators of acreage under paddy based on samples of clusters selected through

- (i) srs without replacement,
- (ii) simple systematic sampling,
- (iii) ppswr, size being geographical area, and
- (iv) pps systematic sampling

have been calculated for different numbers of plots per cluster ($M = 1, 2, 5, 10, 20$) keeping the total number of sample plots fixed at 40. The values of the relative variances as well as their efficiencies compared to direct sampling of the individual plots ($M = 1$) for the corresponding sampling schemes are presented in Table 8.4.

TABLE 8.4. BEHAVIOUR OF RELATIVE VARIANCE AND EFFICIENCY FOR DIFFERENT CLUSTER SIZES WHEN TOTAL NUMBER OF SAMPLE PLOTS IS FIXED.

cluster size	no. of sample clusters	srs wor		systematic		ppswr		pps systematic	
		rel. var.	eff. (%)	rel. var.	eff. (%)	rel. var.	eff. (%)	rel. var.	eff. (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	40	.0798	100	.0648	100	.0454	100	.0132	100
2	20	.0940	85	.0721	90	.0821	55	.0262	50
5	8	.1248	64	.0872	74	.1792	25	.1065	12
10	4	.1450	55	.1260	51	.2625	17	.1279	10
20	2	.2069	39	.2457	26	.4095	11	.4510	3

rel. var.—relative variance; eff.—efficiency.

total number of plots : 851; geographical area : 000 acres; area under paddy : 299 acres.

From this table we see that for all the sampling schemes the efficiency of cluster sampling decreases with increase in the cluster size when the total number of sample plots is fixed and that the rates of decrease in the efficiency are different for the different sampling schemes. For instance, the decrease in efficiency with increase in cluster size is much slower in case of srs and systematic sampling than in case of ppswr and pps systematic sampling. This study of the efficiency of cluster sampling is realistic only under the assumption that the cost of survey is proportional to the total number of sample plots. But the cost of survey is usually not just proportional to the number of sample plots in case of cluster sampling. Taking the cost of survey to be proportional to $n + nM(C_2/C_1)$, the relative efficiency per unit of cost namely,

$$E_c = \frac{V(\hat{Y}_r)}{V(\hat{Y}_c)} \cdot \frac{C_r}{C_c} = \frac{V(\hat{Y}_r)}{V(\hat{Y}_c)} \cdot \frac{M(1+C_2/C_1)}{(1+MC_2/C_1)}, \quad (8.41)$$

has been worked out for different values of C_2/C_1 and the results are presented in Table 8.5

TABLE 8.5 RELATIVE EFFICIENCY PER UNIT OF COST FOR DIFFERENT CLUSTER SIZES AND DIFFERENT SAMPLING SCHEMES

cluster size $\frac{C_2}{C_1}$	relative efficiency per unit cost ($V_r/V_c(C_r/C_c)$)											
	srs w/or			systematic			ppswr			pps systematic		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1	100	100	100	100	100	100	100	100	100	100	100	100
2	156	133	127	165	146	135	101	90	83	93	82	76
5	234	166	137	272	193	159	98	66	54	46	32	27
10	303	179	138	293	167	128	95	56	43	50	29	23
20	253	143	105	193	93	72	81	41	30	21	11	8

From Table 8.5 it can be seen that the efficiency of cluster sampling per unit of cost increases with cluster size upto a certain stage, but thereafter decreases with further increase in cluster size for srs and systematic sampling. The cluster size for which the efficiency is the maximum is to be considered as the optimum cluster size. It may be noted that the decrease in efficiency in case of ppswr and pps systematic sampling has become more gradual than in Table 8.4 due to the use of a more realistic cost structure (cf.) Problem 8.2, p.314).

REFERENCES

- HANSEN, M. H. and HURWITZ, W. N. (1942) : Relative efficiencies of various sampling units in population enquiries; *J. Amer. Stat. Assn.*, 37, 89-94.
- JESSEN, R. J. (1942) : Statistical investigation of a sample survey for obtaining farm facts; *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.
- MAHALANOBIS, P. C. (1940a) : A sample survey of acreage under jute in Bengal, *Sankhyā*, 4, 511-530.
- MAHALANOBIS, P. C. (1940b) : *Report on the Sample Census of Jute*, 1939; Indian Central Jute Committee.
- MAHALANOBIS, P. C. (1942) : *General Report on the Sample Census of Area under Jute in Bengal*, 1941; Indian Central Jute Committee.
- MAHALANOBIS, P. C. (1944) : On large-scale sample surveys; *Phil. Trans. Roy. Soc.*, London, 231, (B), 329-451.
- MAHALANOBIS, P. C. and SENGUPTA, J. M. (1951) : On the size of sample cuts in crop cutting experiments in the ISI, 1939-1950; *Bull. Inter. Stat. Inst.*, 33, (2), 359-403.
- SMITH, H. F. (1938) : An empirical law describing heterogeneity in the yields of agricultural crops, *J. Agri. Sci.*, 28, 1-23.
- SUKHATME, P. V. (1947) : The problem of plot size in large-scale yield surveys; *J. Amer. Stat. Assn.*, 42, 297-310, 460.
- SUKHATME, P. V. (1953) : *Sampling Theory of Surveys with Applications*; Chapter VI, Iowa State College Press, Ames, Iowa, Indian Society of Agricultural Statistics, New Delhi.

COMPLEMENTS AND PROBLEMS

8.1 In planning a sample survey for estimating the proportion P of area under jute in a region, a pilot study was undertaken in which independent samples of clusters of different sizes (x) were taken up for estimating the values of σ_b^2 with a view to studying the relationship between x and σ_b^2 . The results obtained in this pilot survey are given in Table 8.6

TABLE 8.6 ESTIMATES OF $\sigma_b^2 / P(1-P)$ FOR DIFFERENT CLUSTER SIZES

cluster size (acres)	$\sigma_b^2 / P(1-P)$	cluster size (acres)	$\sigma_b^2 / P(1-P)$
(1)	(2)	(1)	(2)
1.00	0.1120	12.25	0.0454
2.25	0.0813	16.00	0.0419
4.00	0.0659	25.00	0.0398
6.25	0.0577	36.00	0.0342
9.00	0.0505		

Source Mahalanobis, P C, *Phil Trans Roy Soc, (B)*, 231, (1944), 329-451, Table 9, p 412

(i) Examine whether column (2) in Table 8.6 conforms to the relation a/x^g , where a and g are constants to be determined

(ii) Assuming the cost function $C = 1000 + 2 \ln n + 0.7nx$, determine the optimum values of x and n for estimating P when the cost is fixed at Rs 10 000

8.2 Using the data given in Table 8.4 (p 311) about the relative variances of estimates of acreage under paddy for different cluster sizes in case of various sampling schemes, determine the optimum cluster sizes for these schemes when the cost is assumed to be proportional to the square root of the size of the cluster

8.3 In a forest nursery, there are six rows each of length 434 feet in the bed. To arrive at a suitable sampling unit for estimating the total number of seedlings in the bed, the entire population was studied using four types of sampling unit (a) one foot length of single row, (b) two feet length of single row (c) one foot of the complete width of the bed and (d) two feet of the complete width of the bed. The results of this study are given in Table 8.7. Find out the optimum sampling unit after comparing the relative cost efficiencies of the four types of units considered here.

TABLE 8.7. DATA ON COST AND VARIANCE FOR FOUR TYPES OF SAMPLING UNIT.

type of unit (1)	total number of units N (2)	variance per unit (3)	length of a row (in feet) covered in 15 minutes (4)
one-foot row	2604	2.537	44
two-feet row	1302	6.746	62
one-foot bed	434	23.094	78
two-feet bed	217	68.558	108

8.4 Suppose in a study on cluster sampling, a sample of n clusters of M units each was selected with srsrwr. Let b and w be unbiased estimates of between-cluster and within-cluster variances. Assuming the sample size in terms of the number of units to be fixed, obtain an estimate of the relative efficiency of cluster sampling as compared to that of direct sampling of units by estimating the sampling variances in the two cases unbiasedly.

8.5 Derive the results (8.23) and (8.26) relating to the variance and efficiency of cluster sampling in estimating a population proportion.

8.6 Derive the result (8.34) and show that it reduces to (8.9) when the clusters are of equal size.

8.7 For examining the efficiency of sampling households instead of persons for estimating the proportion of males in a given area, the following simplifying assumptions are made : (i) each household consists of four persons (husband, wife and two children) and (ii) the sex of a child is binomially distributed. Show that the intraclass correlation coefficient in this case is $(-1/6)$ and that the efficiency of sampling households compared to that of sampling persons is 200%.

(Sukhatme, P. V., *Sampling Theory of Surveys with Applications*, (1953), Ch. VI, 248-250).

8.8 Let there be N clusters of M units each. When n clusters are selected systematically for estimating the population mean per unit, derive the sampling variance of the estimator in terms of the intraclass correlation coefficient (ρ_c) between pairs of units in the clusters and that (ρ'_c) between pairs of clusters in the samples, assuming N to be a multiple of n .

(Madow, W. G., *Ann. Math. Stat.*, 20, (1949), 333-354).

8.9 If the NM units in a population are grouped at random to form N clusters of M units each, show that sampling n clusters with srs w/o r would have the same efficiency as sampling nM units with srs w/o r.

8.10 Let a finite population of M_0 units be divided into N clusters with the i th cluster having M_i units. Suppose a sample of m units is selected from the M_0 population units with srs w/o r and then the sample units are grouped according to the clusters to which they belong. Sampling of n clusters from these clusters (including the empty ones) and observing only the originally sampled units in them is termed *post cluster sampling*. If y is the sample total based on the values of the sample units in the n selected clusters, show that the estimator

$$\hat{\bar{Y}} = Ny/mn$$

is unbiased for the overall population mean \bar{Y} and derive its sampling variance.

(Ghosh, S P, *Ann Math Stat*, 34, (1963) 587-597)

Multi-stage Sampling

9.1 SAMPLING PROCEDURE

In Chapter 8, it has been stated that though cluster sampling is economical under certain circumstances, it is generally less efficient than sampling of individual units directly. A compromise between cluster sampling and direct sampling of units can be achieved by selecting a sample of clusters and surveying only a sample of units in each sample cluster instead of completely enumerating all the units in the sample clusters. Such a procedure is known as *two-stage sampling*, since the units are selected in two stages. Here clusters are termed *first stage units* (fsu) or *primary stage units* (psu) and the ultimate observational units are termed *second stage units* (ssu) or *ultimate stage units* (usu). It may be noted that this procedure can be easily generalized to give rise to *multi-stage sampling*, where the sampling units at each stage are clusters of units of the next stage and the ultimate observational units are selected in stages, sampling at each stage being done from each of the sampling units or clusters selected in the previous stage. This procedure, being a compromise between *uni-stage* or direct sampling of units and cluster sampling, can be expected to be (i) more efficient than uni-stage sampling and less efficient than cluster sampling from considerations of operational convenience and cost, and (ii) less efficient than uni-stage sampling and more efficient than cluster sampling from the view-point of sampling variability, when the sample size in terms of number of ultimate units is fixed. It is of interest to note that an *r*-stage

design reduces to a stratified ($r-1$) stage design when all the fsu's are included in the sample

It may be mentioned that multi stage sampling may be the only feasible procedure in a number of practical situations, where a satisfactory sampling frame of ultimate observational units is not readily available and the cost of obtaining such a frame is prohibitive or where the cost of locating and physically identifying the usu's is considerable. For instance, for conducting a socio economic survey in a region, where generally household is taken as the usu, a complete and up to date list of all the households in the region may not be available, whereas a list of villages or parishes and urban blocks which are groups of households may be readily available. In such a case, a sample of villages and urban blocks may be selected first and then a sample of households may be drawn from each selected village and urban block after making a complete list of households in them. It may happen that even a list of villages is not available, but only a list of all tehsils or counties (groups of villages) is available. In this case a sample of households may be selected in three stages by selecting first a sample of tehsils (counties), then a sample of villages (parishes) from each selected tehsil (county) after making a list of all the villages (parishes) in it and finally a sample of households from each selected village (parish) after listing all the households in it. Since the selection is done in three stages, this procedure is termed *three stage sampling*. Here tehsils (counties) are taken as first stage units (fsu), villages (parishes) as second stage units (ssu) and households as third or ultimate stage units (tsu).

In practice, it usually happens that we have more information for groups of sampling units than for individual units. Hence, if these groups are taken as fsu's, the information available for them can be used in effecting good stratification or arrangement and in selecting the sample of fsu's. Further, since the ssu's are selected only from the sample fsu's, it would be practicable to collect some suitable information about the ssu's at the time of listing them and use this information for obtaining a better sample of ssu's. Because of this, it may be possible that a multi stage design, where the information

available at every stage is properly utilized, is more efficient than one-stage sampling even from the point of view of sampling variability. Multi-stage sampling has been found to be very useful in practice and this procedure is being currently used in a number of surveys. Bhattacharjee (1940) used this sampling procedure in crop surveys carried out in Bengal during the period 1937–1941, and he had termed this procedure as *nested sampling* (Ganguli, 1941). Cochran (1939) and Hansen and Hurwitz (1943) have considered the use of this procedure in agricultural and population surveys respectively. Lahiri (1954) has discussed the use of multi-stage sampling in the Indian National Sample Survey, and Roy (1957) and D. Singh (1958) have considered the estimation of variance components for this sampling scheme.

Another type of sampling in stages consists in drawing a large sample of units in the first stage or phase, for which information on some auxiliary variable is collected and then selecting a sub-sample of these units for the main survey using the auxiliary information for stratification, selection and estimation. This method of sampling is termed *two-phase sampling*. If the ultimate sample is selected in two or more phases, the sampling procedure is termed *multi-phase sampling*. This is considered briefly in Section 9.12 and also in Chapters 10 and 11. This procedure also leads to reduction in cost as compared to uni-stage sampling, though not necessarily to the extent achieved in multi-stage sampling.

9.2 ESTIMATION AND SAMPLING VARIANCE

To illustrate the technique of building up estimators of population total and mean in the case of multi-stage sampling, let us consider the application of two-stage sampling to a population, where the units are grouped into N groups or clusters and the i -th cluster contains M_i units, ($i = 1, 2, \dots, N$). Let Y_{ij} denote the value of the characteristic under consideration for the j -th unit in the i -th cluster. Then the population total ΣY is given by

$$Y = \sum_{i=1}^N Y_i, \quad (Y_i = \sum_{j=1}^{M_i} Y_{ij}). \quad \dots \quad (9.1)$$

Taking the N clusters as fsu's and the units themselves as ssu's, we may sample n fsu's with any given probability scheme and from the i th selected fsu, m_i ssu's may be selected with certain specified probabilities

Let y_{ij} be the value of j th selected ssu in the i th selected fsu ($j = 1, 2, \dots, m_i$, $i = 1, 2, \dots, n$). If the total values of the selected fsu's were known it can be seen that it would be possible to get an estimator of Y with the help of the probability scheme at the first stage as in cluster sampling. But in two stage sampling, the actual totals of the selected fsu's are not known and hence they have to be estimated on the basis of the selected ssu's using the probability scheme adopted in selecting them. That is, in cluster sampling the estimator is of the form $\hat{Y}_c = \sum_{i=1}^n a_i y_i$, where y_i is the total of the i th sample cluster and a_i is the corresponding inflation factor. But in two stage sampling the value of y_i itself is to be estimated by

$$\hat{y}_i = \sum_{j=1}^{m_i} a_{ij} y_{ij}$$

where a_{ij} is the inflation factor at the second stage selection and therefore the estimator of Y takes the form

$$\hat{Y} = \sum_{i=1}^n a_i \hat{y}_i = \sum_{i=1}^n a_i \sum_{j=1}^{m_i} a_{ij} y_{ij} \quad (9.2)$$

For instance, if the units at both the stages are selected with equal probability with or without replacement or circular systematically, an unbiased estimator of Y is given by

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^n \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

Similarly, if n fsu's are selected with probabilities $\{P_i\}$ ($i = 1, 2, \dots, N$), with replacement or circular systematically, and m_i ssu's are selected with probabilities $\{P_{ij}\}$, ($j = 1, 2, \dots, M_i$) with replacement or circular systematically, an unbiased estimator of Y is

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n \frac{1}{P_i} \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{y_{ij}}{P_{ij}},$$

where $\{p_i\}$ and $\{p_{ij}\}$ denote the probabilities of the selected first and second stage units. Extending this technique to the selection of the sample in more than two stages, we find that the estimator of the total and mean can be obtained by considering the estimators of the totals of the sample units at different stages built up from the next stage sample units. In the case of an r -stage design, the estimator will be of the form

$$\hat{Y} = \sum_{i_1=1}^n a_{i_1} \sum_{i_2=1}^{m_{i_1}} a_{i_1 i_2} \cdots \sum_{i_r=1}^{m_{i_1 \dots i_{r-1}}} a_{i_1 i_2 \dots i_r} y_{i_1 i_2 \dots i_r}, \dots \quad (9.3)$$

where $a_{i_1 i_2 \dots i_j}$ is the inflation factor at the j -th stage and $m_{i_1 i_2 \dots i_{j-1}}$ is the number of j -th stage units selected from a $(j-1)$ -th stage unit, ($j = 1, 2, \dots, r$), m_{i_0} being n .

Since in multi-stage sampling the units are selected in stages by adopting a random or probability mechanism at each stage, the selection procedures at all the stages are to be considered in deriving the expected value and the variance of an estimator based on the observations made on a sample of usu's. This is usually done in stages starting from the ultimate stage and moving towards the first stage. The conditional expectation at the last stage is taken for a given set of selected penultimate stage units, the conditional expectation of this at the penultimate stage is taken for a given set of units selected in the previous stage, and this procedure is continued till the unconditional expected value of the conditional expectation of the estimator, taken over the second and subsequent stages of selection, is taken at the first or the primary stage. The question of obtaining the expected value and sampling variance of estimators based on units selected through randomization at two or more stages has been considered briefly in Section 2.8 of Chapter 2 (p.41), where it has been shown that the expected value and the sampling variance of the estimator \hat{Y} based on a two-stage sample are symbolically given by

$$E_{12}(\hat{Y}) = E_1 E_2(\hat{Y}), \dots \quad (9.4)$$

and

$$V_{12}(\hat{Y}) = V_1 E_2(\hat{Y}) + E_1 V_2(\hat{Y}), \dots \quad (9.5)$$

where E_{12} and V_{12} denote the expectation and the variance over the two stages, E_1 and V_1 the unconditional expectation and variance over the first stage and E_2 and V_2 the conditional expectation and variance over the second stage for a given sample of fsu's. Similarly proceeding for a r stage design, we get

$$E(\hat{Y}) = E_1 E_2 \dots E_r(\hat{Y}), \quad (9.6)$$

and

$$\begin{aligned} V(\hat{Y}) = & V_1 E_2 \dots E_{r-1} E_r(\hat{Y}) + E_1 V_2 \dots E_{r-1} E_r(\hat{Y}) + \\ & + E_1 E_2 \dots E_{r-1} V_r(\hat{Y}) \end{aligned} \quad (9.7)$$

It is to be noted that in (9.5) the expression $V_1 E_2(\hat{Y})$ is a measure of the variation between fsu's and the other expression $E_1 V_2(\hat{Y})$ is a measure of the variation between ssu's within fsu's. In other words these expressions are measures of the contribution to the total variance from the two stages of sampling. Similarly, in an r stage design, the total variance consists of r parts, each part giving the variation between units of a particular stage within the units of the previous stage.

9.3 TWO-STAGE SAMPLING WITH SRS

Suppose a sample of n fsu's is selected with srs and from the i th selected fsu a sample of m_i ssu's is selected again with srs. An unbiased estimator of Y is given by

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^n \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}, \quad (9.8)$$

since

$$E(\hat{Y}) = E_1 E_2(\hat{Y}) = E_1 \left(\frac{N}{n} \sum_{i=1}^n y_i \right) = Y,$$

where y_i is the total of the i th sample fsu

9.3a SRS WOR AT BOTH THE STAGES

Suppose the sampling of units at both the stages is done with srs wor. The expression for $V(\hat{Y})$ can be derived by applying (9.5) and we get

$$\begin{aligned} V(\hat{Y}) &= V_1 E_2(\hat{Y}) + E_1 V_2(\hat{Y}) \\ &= V_1 \left(\frac{N}{n} \sum_{i=1}^n y_i \right) + E_1 \left[\frac{N^2}{n^2} \sum_{i=1}^n M_i^2 (1-f_i) \frac{\sigma'_{wi}^2}{m_i} \right] \\ &= N^2 M'^2 (1-f) \frac{\sigma'_b^2}{n} + \frac{N}{n} \sum_{i=1}^N M_i^2 (1-f_i) \frac{\sigma'_{wi}^2}{m_i}, \quad \dots (9.9) \end{aligned}$$

where

$$\sigma'_b{}^2 = \frac{1}{N-1} \sum_{i=1}^N \left(\frac{M_i}{M'} \bar{Y}_i - \bar{Y} \right)^2, \quad \sigma'_{wi}^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2,$$

$f = n/N$, $f_i = m_i/M_i$ and M' is the average number of ssu's per fsu.

An unbiased estimator of $V(\hat{Y})$ can be obtained if it is possible to estimate $\sigma'_b{}^2$ and σ'_{wi}^2 unbiasedly. Since the ssu's within the fsu's are selected with srs wor, an unbiased estimator of σ'_{wi}^2 is given by

$$s_{wi}^2 = \frac{1}{m_i-1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2$$

and hence the second term in (9.9) is unbiasedly estimated by

$$\frac{N^2}{n^2} \sum_{i=1}^n M_i^2 (1-f_i) \frac{s_{wi}^2}{m_i}. \quad \dots (9.10)$$

An unbiased estimator of $\sigma'_b{}^2$ can be obtained by estimating $\sum_{i=1}^N M_i^2 \bar{Y}_i^2$ and \bar{Y}^2 unbiasedly, for $\sigma'_b{}^2$ can be written as

$$\sigma'_b{}^2 = \frac{1}{(N-1)M'^2} \left(\sum_{i=1}^N M_i^2 \bar{Y}_i^2 - \frac{\bar{Y}^2}{N} \right).$$

Since $V_2(\bar{y}_i) = E_2(\bar{y}_i^2) - \bar{Y}_i^2$, an unbiased estimator of \bar{Y}_i^2 is given by

$$\bar{y}_i^2 - v_2(\bar{y}_i) = \bar{y}_i^2 - (1-f_i) \frac{s_{wi}^2}{m_i}$$

and hence that of $\sum_{i=1}^N M_i^2 \bar{Y}_i^2$ is

$$\frac{N}{n} \sum_{i=1}^n M_i^2 \left[\bar{y}_i^2 - (1-f_i) \frac{s_{wi}^2}{m_i} \right].$$

An unbiased estimator of \bar{Y}^2 is given by $\hat{Y}^2 - v(\hat{Y})$. Substituting these estimators in σ_b^2 and simplifying, we get an unbiased estimator of $\sigma_b'^2$ as

$$\frac{N}{n(N-1)} \left[(n-1)s_b^2 - \sum_{i=1}^n \frac{M_i^2}{M'^2} (1-f_i) \frac{s_{wi}^2}{m_i} \right] + \frac{v(\hat{Y})}{N(N-1)M'^2},$$

where $s_b^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{M_i}{M'} \bar{y}_i - \hat{Y} \right)^2$, $\hat{Y} = \hat{Y}/NM'$. Substituting in $V(\hat{Y})$ this estimator and that obtained in (9.10), we get after simplification

$$v(\hat{Y}) = N^2 M'^2 (1-f) \frac{s_b^2}{n} + \frac{N}{n} \sum_{i=1}^n M_i^2 (1-f_i) \frac{s_{wi}^2}{m_i}. \quad \dots \quad (9.11)$$

It may be noted that for calculating the first term in (9.11) the value of M' is not required, since it can be rewritten as

$$\frac{(1-f)}{n(n-1)} \sum_{i=1}^n (NM_i \bar{y}_i - \hat{Y})^2$$

From (9.10) and (9.11) it can be seen that an unbiased estimator of $\sigma_b'^2$ is given by

$$(\hat{\sigma}'^2) = s_b^2 - \frac{1}{n} \sum_{i=1}^n \left(\frac{M_i}{M'} \right)^2 (1-f_i) \frac{s_{wi}^2}{m_i} \quad \dots \quad (9.12)$$

and that an unbiased estimator of the first term of $V(\hat{Y})$ given in (9.9) is got by multiplying (9.12) by $N^2 M'^2 (1-f)/n$.

9.3b SRSWR AND SRS WOR AT THE TWO STAGES

In large-scale surveys, it is desirable to select the fsu's with replacement, since it enables us to get an estimate of the sampling variance of the estimator without having to calculate separately the estimates of the within and between components. Suppose the sampling is done with srswr at the first stage and with srs wor at the second stage, the estimator of \bar{Y} given in (9.8) remains unbiased and its sampling variance can be verified to be

$$V(\hat{Y}) = N^2 M'^2 \frac{\sigma_b^2}{n} + \frac{N}{n} \sum_{i=1}^N M_i^2 (1-f_i) \frac{\sigma'_{wi}^2}{m_i}, \quad \dots \quad (9.13)$$

where $\sigma_b^2 = (N-1)\sigma'_b^2/N$. Noting that the n unbiased estimates of \bar{Y} obtained from n sample fsu's, namely,

$$t_i = NM_i\bar{y}_i, \quad i = 1, 2, \dots, n, \quad \dots \quad (9.14)$$

are statistically independent and have the same sampling variance, we get an unbiased variance estimator of $\frac{1}{n} \sum_{i=1}^n t_i = \hat{Y}$ as

$$v(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^n (NM_i\bar{y}_i - \hat{Y})^2 = N^2 M'^2 \frac{s_b^2}{n}, \quad \dots \quad (9.15)$$

where s_b^2 is as defined earlier.

9.3c SRSWR AT BOTH THE STAGES

Suppose the sampling at both the stages is done with srswt, then also the estimator given in (9.8) remains unbiased for \bar{Y} and its sampling variance is given by

$$V(\hat{Y}) = N^2 M'^2 \left[\frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nm} \right], \quad \dots \quad (9.16)$$

where $\sigma_w^2 = \frac{1}{N} \sum_{i=1}^N \frac{M_i}{M'} \sigma'_{wi}^2$, $\sigma'_{wi}^2 = \frac{M_i-1}{M_i} \sigma_{wi}^2$, when $f_i = \frac{m}{M}$.

Since here too the fsu's are selected with replacement, an unbiased variance estimator is given by (9.15).

9.3d ESTIMATION OF POPULATION MEAN

In all the cases considered in this section an unbiased estimator of the population mean $\bar{Y} = \bar{Y}/NM'$ can be easily obtained by dividing \hat{Y} by NM' , if the value of M' is known in advance. In that case, the expressions for the variance and the variance estimator can be obtained by dividing the corresponding expressions for \hat{Y} by $N^2M'^2$. If the value of M' is not known in advance, then it has to be estimated by the mean of the number of ssus in the n sample fsu's and in this case the estimator of \bar{Y} is given by

$$\hat{\bar{Y}} = \frac{N}{n} \sum_{t=1}^n \frac{M_t}{m_t} \sum_{f=1}^{m_t} y_{tf} / \sqrt{\frac{N}{n} \sum_{t=1}^n M_t} \quad (9.17)$$

This, being a ratio of two unbiased estimators, is generally biased and such estimators are considered in detail in Chapter 10.

9.3e UNI-STAGE AND CLUSTER SAMPLING

It is of interest to note that two stage sampling reduces to cluster sampling if $m_t = M_t$ in the sampling schemes considered in Sub sections 9.3a and 9.3b and hence the variances of estimators for cluster sampling can be obtained as special cases of the variances in the case of two stage sampling by substituting $f_t = 1$ in (9.9) and (9.13). For comparing the efficiency of two stage sampling with that of uni stage sampling and cluster sampling for a given total sample size in terms of number of ultimate units, let us consider the simplified case where $M_t = M$ and $m_t = m$ for $t = 1, 2, \dots, N$. In this case, the estimator of \bar{Y} is given by

$$\hat{\bar{Y}} = \frac{1}{nm} \sum_{t=1}^n \sum_{f=1}^m y_{tf} \quad (9.18)$$

and its variance in two stage sampling with srswr at the first stage and srs wr at the second stage becomes

$$V(\hat{\bar{Y}}) = \frac{\sigma_b^2}{n} + \frac{M-m}{M-1} \frac{\sigma_w^2}{nm} \quad (9.19)$$

Noting that σ_b^2 and σ_u^2 can be expressed in terms of the population variance σ^2 and the intraclass correlation coefficient ρ_c . (cf. Subsection 8.2a of Chapter 8, p. 275), namely,

$$\sigma_b^2 = \frac{\sigma^2}{M} \{1 + (M-1)\rho_c\} \text{ and } \sigma_u^2 = \frac{M-1}{M} \sigma^2(1-\rho_c),$$

we get after simplification

$$V(\hat{\bar{Y}}_t) = \frac{\sigma^2}{nm} \{1 + (m-1)\rho_c\}, \quad \dots \quad (9.20)$$

where the subscript t is used to denote two-stage sampling.

If the total number of units is fixed at nm , then the variances of estimators of \bar{Y} in cluster sampling and uni-stage sampling can be shown to be given by

$$V(\hat{\bar{Y}}_c) = \frac{\sigma^2}{nm} \{1 + (M-1)\rho_c\} \quad \dots \quad (9.21)$$

and

$$V(\hat{\bar{Y}}_r) = \frac{\sigma^2}{nm}, \quad \dots \quad (9.22)$$

where the subscripts c and r denote cluster sampling and uni-stage srs respectively. Comparing (9.21) and (9.22) with (9.20), we find that

$$V(\hat{\bar{Y}}_r) \leq V(\hat{\bar{Y}}_t) \leq V(\hat{\bar{Y}}_c),$$

if $\rho_c \geq 0$ which is likely to be the case in practice when nearby units are grouped to form the clusters or fsu's. This shows that the sampling efficiency of two-stage sampling design is expected to be between those of the uni-stage srs and cluster sampling for fixed total sample size.

If in the above case, we adopt sampling with replacement at the second stage also, the sampling variance given in (9.19) reduces to

$$V(\hat{\bar{Y}}_t) = \frac{\sigma_b^2}{n} + \frac{\sigma_u^2}{nm} \quad \dots \quad (9.23)$$

and when expressed in terms of σ^2 and ρ_c it becomes

$$V(\bar{Y}_t) = \frac{\sigma^2}{nm} \left[\frac{m}{M} + \frac{M-1}{M} \{1 + (m-1)\rho_c\} \right] \quad (9.24)$$

Comparing this with $V(\hat{\bar{Y}}_t) = \sigma^2/nm$ we verify that $V(\hat{\bar{Y}}_t) \geq V(\bar{Y}_t)$ since $\rho_c \geq -1/(M-1)$. The behaviour of the sampling variance and the determination of optimum values of n and m for a suitable cost function are considered in Section 9.7.

9.4 SAMPLING WITH PPSWR AND SRS WOR

As mentioned earlier it is possible to increase the efficiency of multi stage sampling by utilizing the auxiliary information that may be available for the first and subsequent stage units in stratification and selection at the different stages. For instance in a socio-economic survey where the ultimate unit is a household the villages may be treated as fsu's and selected with ppswr size being the number of households or population. Similarly in a level of living survey of industrial workers the factories may be treated as fsu's and selected with ppswr size being the number of workers. Then from the sample factories the workers (ssu's) may be selected with srs wor. Further if the number of units in the fsu's differ considerably it is desirable to select the fsu's with pps size being M_i 's instead of with srs as this is expected to reduce the between fsu variance substantially or to stratify the fsu's on the basis of their size and adopt two stage sampling within each stratum. Here again one may adopt a more efficient design for selecting ssu's by collecting some supplementary information for all the ssu's in the selected fsu's and utilizing them for stratification arrangement and selection at the second stage.

Suppose a sample of n fsu's is selected with ppswr and from the i th selected fsu a sample of m_i ssu's is drawn with srs wor. An unbiased estimator of \bar{Y} is given by

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n \frac{M_i y_i}{p_i} \quad y_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \quad (9.25)$$

where p_i is the probability of selecting the i -th fsu at each draw. Noting that

$$E_2(\hat{Y}) = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} \text{ and } V_2(\hat{Y}) = \frac{1}{n^2} \sum_{i=1}^n \frac{M_i^2}{p_i^2} (1-f_i) \frac{\sigma'_{ic_i}^2}{m_i},$$

where y_i is the total value of the i -th fsu, we get the variance of \hat{Y} as

$$V(\hat{Y}) = \frac{1}{n} \sum_{i=1}^N \left(\frac{Y_i}{P_i} - Y \right)^2 P_i + \frac{1}{n} \sum_{i=1}^N \frac{M_i^2}{P_i} (1-f_i) \frac{\sigma'_{ic_i}^2}{m_i}. \quad \dots \quad (9.26)$$

Since the fsu's are selected with replacement, each of the n estimates $\{M_i \bar{y}_i / p_i\}$ is unbiased for Y and they are independently distributed with the same variance. Hence, an unbiased variance estimator of the combined estimator \hat{Y} is given by

$$v(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{M_i \bar{y}_i}{p_i} - \hat{Y} \right)^2. \quad \dots \quad (9.27)$$

It is to be noted that this procedure of estimation holds good irrespective of the method of selection adopted at the second stage provided the fsu's are selected with replacement. But if the between and within components of the variance are to be estimated, then the within-fsu variance can be estimated unbiasedly by

$$w(\hat{Y}) = \frac{1}{n^2} \sum_{i=1}^n \frac{M_i^2}{p_i^2} (1-f_i) \frac{s_{ic_i}^2}{m_i} \quad \dots \quad (9.28)$$

and the between-fsu variance can be obtained by subtracting (9.28) from (9.27).

9.5 VARIANCE FUNCTION AND ITS BEHAVIOUR

From the expressions for sampling variances of the estimators considered in the last two sections, we find that the total variance in two-stage sampling consists of two parts—*between-fsu* variance and *within-fsu* variance—and that the former decreases with increase in n , and the latter decreases with increases in both n and m . In fact,

the sampling variance of an estimator of Y in a two stage design can be written in the following form

$$\frac{1}{n} \left\{ A_1 + \sum_{i=1}^N \frac{A_{2i}}{m_i} \right\} + A_3 \quad (9.29)$$

where A_1 and A_{2i} are respectively the coefficients of $1/n$ and $1/nm_i$ and A_3 is a constant term. If $m_i = r_i m$ where r_i depends on the method of determining the values of $\{m_i\}$, (9.29) becomes

$$\frac{1}{n} \left(A_1 + \frac{A_2}{m} \right) + A_3 \quad (9.30)$$

where A_1 , A_2 and A_3 are functions of population parameters and are independent of the sample sizes n and m at the two stages of sampling. It may be noted that if sampling is done with replacement at the first stage, A_3 would be zero.

From (9.30) it is clear that the behaviour of the variance as n and m increase would depend only on the terms involving A_1 and A_2 . It can be easily seen that increasing of n plays a more important part than increasing of m since in the former case both the components of variance A_1/n and A_2/nm get reduced, whereas in the latter case only the component A_2/nm gets reduced. The ratio of the variance in two-stage sampling to that of sampling one ssu from the selected fsu's has been tabulated in Table 9.1 for different values of m and of the ratio of A_1 to A_2 assuming A_3 to be zero and the number of sample fsu's to be fixed.

It is of interest to note from this table that the gain in efficiency in increasing m is substantial for initial increases in m , but becomes quite marginal after a certain stage. From this, it is clear that it is desirable to increase n for effecting substantial reduction in sampling variance, provided the cost of increasing n is commensurate with the gain achieved. For instance if the cost is taken as proportional to the total sample size nm , then the behaviour of efficiency of two stage sampling for different values of m as compared to that in the case of $m = 1$ can be studied for different values of A_1/A_2 assuming A_3 to be zero and keeping nm a constant. The results of such a study are

TABLE 9.1. BEHAVIOUR OF SAMPLING VARIANCE IN TWO-STAGE SAMPLING WHEN NUMBER OF SAMPLE FSU'S IS FIXED.

number of ssu's per fsu	variance ratio* (%) when $A_1 : A_2$ is						
	4 : 1	3 : 1	2 : 1	1 : 1	1 : 2	1 : 3	1 : 4
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	100	100	100	100	100	100	100
2	90	88	83	75	67	62	60
3	87	83	78	67	56	50	47
4	85	81	75	62	50	44	40
5	84	80	73	60	47	40	36
6	83	79	72	58	44	38	33
7	83	79	71	57	43	36	31
8	82	78	71	56	42	34	30
9	82	78	70	56	41	34	29
10	82	77	70	55	40	32	28
15	81	77	69	53	38	30	25
20	81	76	68	52	37	29	24
25	81	76	68	52	36	28	23
30	81	76	68	52	36	27	23
40	80	76	68	51	35	27	22
50	80	76	67	51	35	26	22
60	80	75	67	51	34	26	21
70	80	75	67	51	34	26	21
80	80	75	67	51	34	26	21
90	80	75	67	51	34	26	21
100	80	75	67	50	34	26	21

$$* \text{ variance ratio} = \left\{ \frac{1}{n} \left(A_1 + \frac{A_2}{m} \right) \right\} / \left\{ \frac{1}{n} (A_1 + A_2) \right\}.$$

given in Table 9.2. From the table, we find that the relative efficiency is not low for the initial values of m , but it becomes quite low after a certain stage, though the decrease in efficiency is less sharp when A_1/A_2 is less than unity. The question of the behaviour of efficiency for a more realistic cost function is examined in Section 9.8.

TABLE 9.2 BEHAVIOUR OF EFFICIENCY OF TWO STAGE SAMPLING
WHEN TOTAL NUMBER OF SAMPLE SSUs IS FIXED

number of ssus per fsu	relative efficiency* (%) when A_1 , A_2 is							
	4 1	3 1	2 1	1 1	1 2	1 3	1 4	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
1	100	100	100	100	100	100	100	100
2	56	57	60	67	75	80	83	
3	38	40	43	50	60	67	71	
4	29	31	33	40	50	57	62	
5	24	25	27	33	43	50	56	
6	20	21	23	29	38	44	50	
7	17	18	20	25	33	40	46	
8	15	16	18	22	30	36	42	
9	14	14	16	20	27	33	38	
10	12	13	14	18	25	31	36	
15	8	9	10	12	18	22	27	
20	6	7	7	10	14	17	21	
25	5	6	6	8	11	14	17	
30	4	4	5	6	9	12	15	
40	3	3	4	5	7	9	11	
50	2	3	3	4	6	8	9	
60	2	2	2	3	5	6	8	
70	2	2	2	3	4	6	7	
80	2	2	2	2	4	5	6	
90	1	1	2	2	3	4	5	
100	1	1	1	2	3	4	5	

$$* \text{relative efficiency} = \left\{ \frac{1}{nm} (A_1 + A_2) \right\} / \left\{ \frac{1}{n} \left(A_1 + \frac{A_2}{m} \right) \right\}$$

The expressions for A_1 , A_2 and A_3 for the sampling schemes so far considered are given in Table 9.3 together with their unbiased estimators,

TABLE 9.3. VARIANCE COMPONENTS OF THE ESTIMATOR OF \bar{Y} AND THEIR ESTIMATORS IN TWO-STAGE SAMPLING.

term	expression	estimator
(1)	(2)	(3)
1. srs wor at both the stages		
A_1	$\sigma_b'^2 = \frac{1}{N} \sum_{t=1}^N (\alpha_t/M_t)$	$s_b^2 = \frac{1}{n} \sum_{t=1}^n (\hat{\alpha}_t/m_t)$
A_2	$\frac{1}{N} \sum_{t=1}^N (\alpha_t/r_t)$	$\frac{1}{n} \sum_{t=1}^n (\hat{\alpha}_t/r_t)$
A_3	$-\frac{1}{N} \sigma_b'^2$	$-\frac{1}{N} \left\{ s_b^2 - \frac{1}{n} \sum_{t=1}^n (1-f_t) \frac{\hat{\alpha}_t}{m_t} \right\}$
2. srswr and srs wor at first and second stages		
A_1	$\sigma_b^2 = \frac{1}{N} \sum_{t=1}^N (\alpha_t/M_t)$	$s_b^2 = \frac{1}{n} \sum_{t=1}^n (\hat{\alpha}_t/m_t)$
A_2	$\frac{1}{N} \sum_{t=1}^N (\alpha_t/r_t)$	$\frac{1}{n} \sum_{t=1}^n (\hat{\alpha}_t/r_t)$
3. srswr at both the stages		
A_1	σ_b^2	$s_b^2 = \frac{1}{n} \sum_{t=1}^n (\hat{\alpha}_t/m_t)$
A_2	$\frac{1}{N} \sum_{t=1}^N \alpha_t(M_t-1)/M_t r_t$	$\frac{1}{n} \sum_{t=1}^n (\hat{\alpha}_t/r_t)$
4. ppsswr and srs wor at first and second stages		
A_1	$\frac{1}{N^2 M'^2} \sum_{t=1}^N \left(\frac{Y_t}{P_t} - \bar{Y} \right)^2 P_t$	$\frac{1}{n-1} \sum_{t=1}^n \left(\frac{M_t \bar{y}_t}{P_t} - \hat{Y} \right)^2 / N^2 M'^2$
	$- \frac{1}{N^2} \sum_{t=1}^N (\alpha_t/M_t P_t)$	$- \frac{1}{N^2} \frac{1}{n} \sum_{t=1}^n (\hat{\alpha}_t/m_t P_t^2)$
A_2	$\frac{1}{N^2} \sum_{t=1}^N (\alpha_t/r_t P_t)$	$\frac{1}{N^2} \frac{1}{n} \sum_{t=1}^n (\hat{\alpha}_t/r_t P_t^2)$

$$\sigma_b'^2 = \frac{1}{N-1} \sum_{t=1}^N \left(\frac{M_t \bar{Y}_t}{M'} - \bar{Y} \right)^2; \quad \sigma_b^2 = \frac{N-1}{N} \sigma_b'^2; \quad \alpha_t = \left(\frac{M_t}{M'} \right)^2 \sigma_{wt}^2; \quad f_t = \frac{m_t}{M_t};$$

$$\sigma_{wt}^2 = \frac{1}{M_t-1} \sum_{i=1}^{M_t} (Y_{ti} - \bar{Y}_t)^2; \quad \hat{\alpha}_t = \left(\frac{M_t}{M'} \right)^2 s_{wt}^2; \quad s_{wt}^2 = \frac{1}{m_t-1} \sum_{j=1}^{m_t} (y_{tj} - \bar{y}_t)^2;$$

$$r_t = (m_t/m); \quad s_b^2 = \frac{1}{n-1} \sum_{t=1}^n \left(\frac{M_t}{M'} \bar{y}_t - \hat{Y} \right)^2.$$

9.6 COST FUNCTION

In a two stage sampling design, the cost of survey may be considered to consist of two parts one depending mainly on the number of sample fsu's and the other on the number of sample ssu's. A simple type of cost function is of the form

$$C = C_0 + nC_1 + nmC_2, \quad (9.31)$$

where C_0 is the overhead cost, C_1 is the cost per sample fsu for preliminary field operations such as location and identification of the unit including journey and preparation of the sampling frame for selection of ssu's and preliminary tabulation operations such as computation of inflation factors. C_2 is the cost per sample ssu for selection, identification, data collection and processing of data and m is the average number of sample ssu's per fsu, which depends on the procedure used for allocating the total sample size of ssu's to the selected fsu's.

The cost C_1 relating to fsu's may be considered to consist of two components C_{11} and C_{12} , where C_{11} is a constant and C_{12} depends on the average number of ssu's in a selected fsu. That is, $C_1 = C_{11} + C_{12}$ where $C_{12} = MC_{12}$, M being the average number of ssu's in a sample fsu and this would depend on the procedure adopted for selecting the sample fsu's. For instance, in the case of using srs at the first stage $M = \frac{1}{N} \sum_{i=1}^N M_i$ whereas if pps is used at the first stage, $M = \sum_{i=1}^N M_i P_i$. This division of the cost component C_1 is necessary only when comparison is being made between different procedures of selection of the fsu's and when the optimum size of fsu's is to be determined. Further, if there is evidence to believe that the expected distance between the n sample fsu's is proportional to \sqrt{n} as mentioned in Sub section 8.3b of Chapter 8 (p 301), then the cost component nC_{11} can be replaced by $\sqrt{n} \alpha C'_{11} + nC''_{11}$, where α is a constant, and C'_{11} is the cost of journey per unit distance. Hence the cost function in its more detailed form can be written as

$$C = C_0 + \sqrt{n} \alpha C'_{11} + nC''_{11} + nMC_{12} + nmC_2 \quad . \quad (9.32)$$

In practice, C_1 is likely to be larger than C_2 . Hence, a unit increase in n increases the cost more than a unit increase in m , and this is just the reverse of the situation in respect of the variance, which decreases at a more rapid rate for increases in n than for increases in m .

9.7 OPTIMUM VALUES OF n AND m

Because of the behaviour of the variance and cost functions in opposite directions for increases in n and m , it is necessary in practice to evolve an optimum (or near optimum) solution for the problem of determination of the values of n and m such that the efficiency per unit of cost is the maximum.

9.7a EXPECTED NUMBER OF SAMPLE SSU'S FIXED

Before proceeding to the determination of optimum values of n and m taking into consideration both sampling variance and cost, let us briefly examine the question of optimum allocation of the total sample size in terms of ssu's to the selected first stage units such that the average number of sample ssu's per selected fsu is m . Noting that the term involving $\{m_i\}$ in the variance of \hat{Y} is of the form $\sum_{i=1}^N (A_{2i}/m_i)$ and that $E(m_i) = \sum_{i=1}^N m_i P_i$, we can find the set of values of $\{m_i\}$ which minimizes this component of the variance subject to the condition that $E(m_i) = m$. By equating to zero the partial derivative of the following expression

$$\sum_{i=1}^N (A_{2i}/m_i) + \lambda \left(\sum_{i=1}^N m_i P_i - m \right)$$

with respect to m_i and solving for m_i and λ , the Lagrangian multiplier, we get

$$m_i = \sqrt{A_{2i}/\lambda P_i} \text{ and } \sqrt{\lambda} = \frac{1}{m} \sum_{i=1}^N \sqrt{A_{2i} P_i}.$$

Hence the optimum value of m_i is given by

$$m \sqrt{A_{2i}/P_i} / \sum_{i=1}^N \sqrt{A_{2i} P_i}, \quad (9.33)$$

which reduces to

$$m_i = N m \sqrt{A_{2i}} / \sum_{i=1}^N \sqrt{A_{2i}}, \quad (9.34)$$

if srs is used at the first stage

Substituting these values of $\{m_i\}$ in $\sum_{i=1}^N A_{2i}/m_i$ we get A_2/m where A_2 is independent of m . Since in practice the value of A_2 may not be available for the estimation variable it has to be calculated for a suitable auxiliary variable on the basis of data collected in a previous complete census or sample survey or on the basis of a pilot study to be conducted for this purpose. The problem of determination of the values of $\{m_i\}$ along similar lines has been considered by Ranganathan (1957) and J N K Rao (1961) (cf Problems 9.11 and 9.12 pp 358-9). Another method of determining the values of $\{m_i\}$ which equalizes the inflation factors is considered in Chapter 12.

We have shown in this section that the expression for the variance given in (9.29) can be reduced to that given in (9.30) by adopting a suitable method of determining $\{m_i\}$ such that the average number of sample ssus per selected fsu is m . Using this expression for the variance and the cost function given in (9.31) it is possible to find the optimum values of n and m which minimize (i) the sampling variance for a fixed cost or (ii) the cost for ensuring a specified sampling variance.

9.7b COST FIXED

Suppose the cost is fixed at C . Then the optimum values of n and m which minimize the variance can be obtained by noting that the value of n for any given m is

$$n = (C - C_0)/(C_1 + mC_2) \quad (9.35)$$

and minimizing the variance with respect to m after substituting for n from (9.35). That is,

$$\begin{aligned} V(\hat{Y}) &= \frac{C_1 + mC_2}{C' - C_0} \left(A_1 + \frac{A_2}{m} \right) + A_3 \\ &= \frac{C_1 A_1}{C' - C_0} \left\{ \left(m \frac{C_2}{C_1} + \frac{1}{m} \frac{A_2}{A_1} \right) + \left(1 + \frac{C_2}{C_1} \frac{A_2}{A_1} \right) \right\} + A_3. \quad \dots \quad (9.36) \end{aligned}$$

The optimum value m_0 of m which minimizes the variance can be found by equating to zero the partial derivative of the terms involving m , namely,

$$m \frac{C_2}{C_1} + \frac{1}{m} \frac{A_2}{A_1} = 0 \quad \dots \quad (9.37)$$

with respect to m and solving for m . Thus we get

$$m_0 = \sqrt{A_2 C_1 / A_1 C_2}. \quad \dots \quad (9.38)$$

Substituting this value of m in (9.35), we get the optimum value n_0 of n as

$$n_0 = (C' - C_0) \frac{\sqrt{A_1/C_1}}{\sqrt{A_1 C_1} + \sqrt{A_2 C_2}}. \quad \dots \quad (9.39)$$

Substituting the values of n and m given in (9.38) and (9.39) in the expression for variance, we get the minimum variance as

$$V(\hat{Y}) = \frac{1}{C' - C_0} (\sqrt{A_1 C_1} + \sqrt{A_2 C_2})^2 + A_3. \quad \dots \quad (9.40)$$

9.7c VARIANCE FIXED

Suppose the variance of the estimator in two-stage sampling is fixed at a given value V_0 . Then the values of n and m , which minimize the cost ensuring at the same time the required value for the variance, are obtained by noting that

$$n = \frac{A_1 + (A_2/m)}{V_0 - A_3} \quad \dots \quad (9.41)$$

and

$$C = C_0 + \frac{A_1 C_1}{V_0 - A_3} \left\{ \left(m \frac{C_2}{C_1} + \frac{1}{m} \frac{A_2}{A_1} \right) + \left(1 + \frac{C_2}{C_1} \frac{A_2}{A_1} \right) \right\} \quad (9.42)$$

Since the terms involving m in (9.42) are exactly the same as in the previous case here also the value of m_0 is given by (9.38). Substituting this value of m in (9.41) we get the optimum value n_0 of n

$$n_0 = \frac{\sqrt{A_1 C_1} + \sqrt{A_2 C_2}}{V_0 - A_3} \sqrt{\frac{A_1}{C_1}} \quad (9.43)$$

Substituting the optimum values of n and m in the cost function we get the minimum cost as

$$C = C_0 + \frac{(\sqrt{A_1 C_1} + \sqrt{A_2 C_2})^2}{V_0 - A_3} \quad (9.44)$$

9.7d GRAPHICAL METHOD

From the previous two sub sections, we find that the optimum value of m is the same whether cost is considered as fixed or variance is taken as prespecified and that the optimum value of n is determined using the cost or the variance restriction. The optimum value of m can also be obtained graphically by plotting the values of the expression in (9.37) for different values of m , when the values of C_1/C_2 and A_1/A_2 are known for a particular situation on the basis of a previous survey or a pilot study and by locating that value of m where the graph has the minimum value. Besides providing the optimum value of m the graph brings out the behaviour of the variance or cost with increase in m and in the neighbourhood of the optimum value, which is of considerable interest in practice for studying the effect of departures from the optimum value. The graphs of the expression in (9.37) for some values of C_1/C_2 and A_1/A_2 are shown in figure 9.1. The graphs in this figure are rather flat at the optimum

values showing that slight departures from the optimum would not appreciably affect the sampling variance or the cost.

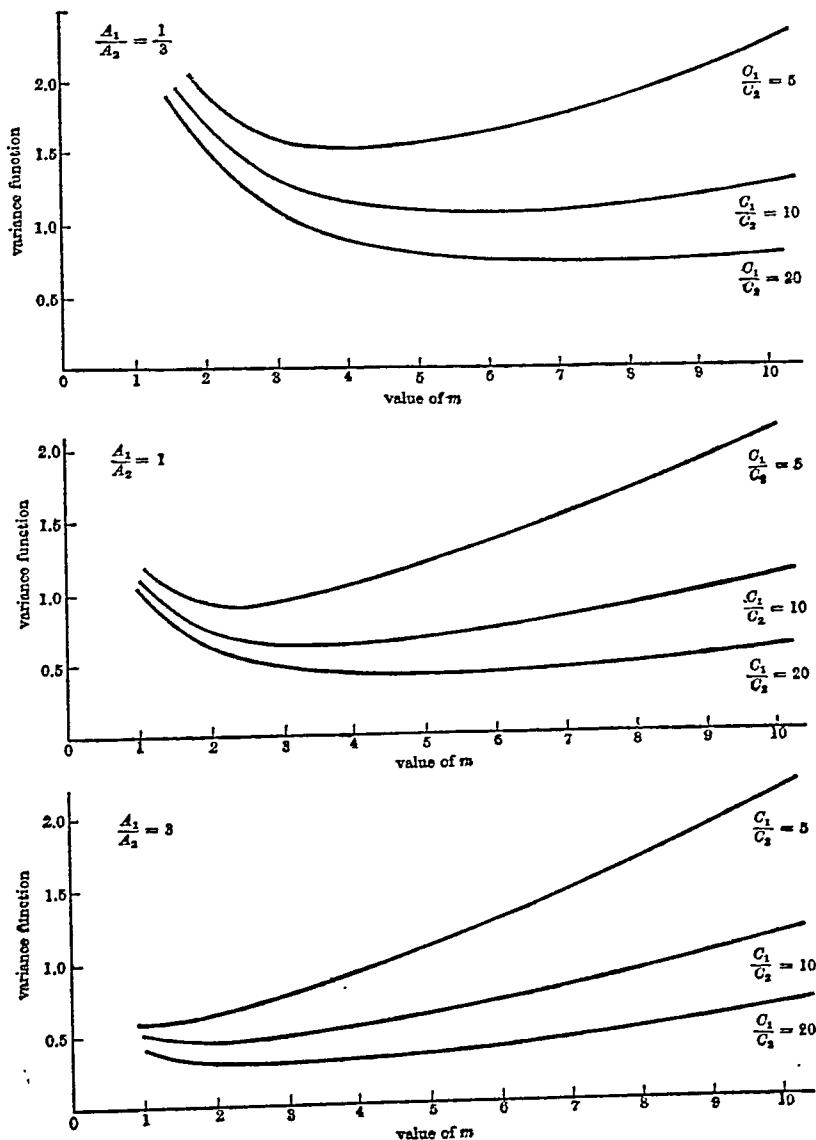


Figure 9.1. Behaviour of variance function in two-stage sampling when cost is fixed.

9.7c OPTIMUM SIZE OF FSU

When the cost is fixed at C , the minimum variance obtained by substituting optimum values of n and m is given in (9.40). Noting that $C_1 = C_{11} + MC_{12}$ and that A_1 and A_2 are functions of M and ρ , it is possible to find the optimum size of the fsu's by plotting the values of the minimum variance for different sizes of the fsu's and by locating that value of M which leads to the least minimum variance. In practice, one has of course to take into account other factors such as administrative and operational convenience in deciding about the size of the fsu's in addition to the purely variance cost approach.

9.8 EFFICIENCY OF TWO-STAGE SAMPLING

In Sub section 9.3e, the sampling efficiency of two stage sampling has been compared with that of cluster sampling and uni stage sampling. In this section we shall study the behaviour of the efficiency of two stage sampling per unit of cost and compare it with those of cluster sampling and uni stage sampling and also obtain estimates of the relative efficiencies based on a two stage sample.

9.8a BEHAVIOUR OF EFFICIENCY

Considering the inverse of the sampling variance as a measure of sampling efficiency the relative sampling efficiency of two stage sampling with n sample fsu's and m sample ssu's per sample fsu as compared to sampling of nm fsu's and one ssu per sample fsu is

$V(\hat{Y}_t)/V(\hat{Y}_s)$. Since the cost of survey would be different for the two schemes being considered the sampling efficiency is to be multiplied by the ratio of costs (C_t/C_s) to yield the cost efficiency (cf Section 2.10 of Chapter 2, p 44). Noting that for a two stage sampling design where the fsu's are selected with replacement, the variance of the estimator of \hat{Y} and the cost of survey are given by

$$V(\hat{Y}_t) = \frac{1}{n} \left(A_1 + \frac{A_2}{m} \right) \quad \text{and} \quad C_t = n(C_1 + C_2 m)$$

and that in the case of sampling nm fsu's and one ssu per sample fsu the variance and cost functions reduce to

$$V(\hat{Y}_r) = \frac{1}{nm} (A_1 + A_2) \quad \text{and} \quad C_r = nm(C_1 + C_2),$$

the expression for the cost-efficiency is of the form

$$\left(1 + \frac{A_2}{A_1}\right) \left(1 + \frac{C_2}{C_1}\right) / \left(1 + \frac{1}{m} \frac{A_2}{A_1}\right) \left(1 + m \frac{C_2}{C_1}\right). \quad \dots \quad (9.45)$$

The values of (9.45) for different values of the ratio of the components of the variance (A_1/A_2) and of the cost ratio (C_1/C_2) are presented in Table 9.4. From this table, we find that the cost-efficiency increases with m upto a certain stage and then it decreases, showing that there is an optimum value for the number of ssu's to be selected per sample fsu.

TABLE 9.4. COST-EFFICIENCIES OF TWO-STAGE SAMPLING AS COMPARED TO SAMPLING OF ONE SSU PER SAMPLE FSU WHEN TOTAL NUMBER OF SAMPLE SSU'S IS FIXED.

number of ssu's per fsu	A_1/A_2	1/4				1				4				
		20	10	5	1	20	10	5	1	20	10	5	1	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1		100	100	100	100	100	100	100	100	100	100	100	100	100
2		159	153	143	111	127	122	114	89	106	102	95	74	
3		196	181	160	107	137	127	113	75	105	98	87	58	
4		218	196	167	100	165	126	107	64	103	92	78	47	
5		233	204	167	93	160	122	100	56	100	87	71	40	
6		242	206	164	86	139	118	93	49	97	82	65	34	
7		247	206	159	80	136	113	87	44	94	78	60	30	
8		250	204	154	74	133	108	82	39	91	74	56	27	
9		250	200	148	69	130	104	77	36	88	70	52	24	
10		250	196	143	65	127	100	73	33	85	67	49	22	
15		237	174	118	49	112	83	56	23	74	54	37	18	
20		219	153	100	40	100	70	46	18	65	45	30	12	
25		201	135	86	33	90	60	38	14	58	38	25	9	
50		139	85	51	18	59	46	21	8	37	23	14	5	
100		84	48	29	9	35	20	11	4	22	13	7	2	

9.8b SRS WOR AT BOTH THE STAGES

Suppose n fsu's are selected with srs wor and then m_i ssu's are selected with srs wor from M_i ssu's in the i th selected fsu. In this case we note from (9.9) and (9.11) of Sub section 9.3a that the variance and the variance estimator are given by

$$V_t(\hat{Y}_t) = (1-f) \frac{\sigma_b^2}{n} + \frac{1}{Nn} \sum_{i=1}^N \left(\frac{M_i}{M} \right)^2 (1-f_i) \frac{\sigma_{ui}^2}{m_i} \quad (9.46)$$

and

$$v_t(\hat{Y}_t) = (1-f) \frac{s_b^2}{n} + \frac{1}{Nn} \sum_{i=1}^N \left(\frac{M_i}{M} \right)^2 (1-f_i) \frac{s_{ui}^2}{m_i}. \quad (9.47)$$

The cost of survey in this case is given by

$$C_t = C_0 + nC_1 + nmC_2 \quad (9.48)$$

and the efficiency per unit of cost is $(1/V(\hat{Y}_t))/C_t$

If n fsu's are selected with srs wor and are completely enumerated, we get cluster sampling and the sampling variance in that case is of the form

$$V(\hat{Y}_c) = (1-f) \frac{\sigma_b^2}{n}, \quad (9.49)$$

since the within fsu component of the variance becomes zero. It is possible to estimate this variance unbiasedly on the basis of a two stage sample by noting that an unbiased estimator of σ_b^2 is given in (9.12). Thus we have

$$v_t(\hat{Y}_c) = \frac{1-f}{n} \left\{ s_b^2 - \frac{1}{n} \sum_{i=1}^N \left(\frac{M_i}{M} \right)^2 (1-f_i) \frac{s_{ui}^2}{m_i} \right\} \quad (9.50)$$

The expected cost of survey for cluster sampling becomes

$$C_c = C_0 + nC_1 + nM C_2 \quad (9.51)$$

and the efficiency per unit of cost is $(1/V(\hat{Y}_c))/C_c$

In unit stage sampling of NM units with srs wor, the variance of the estimator of \bar{Y} is given by

$$V(\hat{Y}_r) = \left(\frac{1}{nm} - \frac{1}{NM} \right) \sigma^2, \quad \sigma^2 = \frac{1}{NM-1} \left(\sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij}^2 - NM \bar{Y}^2 \right) \quad (9.52)$$

This variance can be estimated unbiasedly on the basis of a two stage sample by noting that an unbiased estimator of the sum of squares of y 's can be obtained in the same manner as the estimator of \bar{Y} and that an unbiased estimator of \bar{Y}^2 is given by $\hat{Y}_t^2 - v_t(\hat{Y}_t)$. Hence, we get

$$v_t(\hat{Y}_r) = \frac{1}{nm} \frac{(NM-nm)}{NM(NM-1)} \left[\frac{N}{n} \sum_{i=1}^N \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}^2 - NM (\hat{Y}_t^2 - v_t(\hat{Y}_t)) \right] \quad (9.53)$$

The cost of survey for uni-stage sampling is given by

$$C_r = C_0 + nm(C_1 + C_2). \quad \dots \quad (9.54)$$

The efficiency per unit of cost is $\{1/V(\hat{Y}_r)\}/C_r$.

We see from (9.50) and (9.53) that it is possible to estimate the efficiency of two-stage sampling per unit of cost relative to those in cluster sampling and uni-stage sampling on the basis of a two-stage sample. It may be mentioned that though the cost structure is taken to be of the same form in these three situations, it is possible that C_0 , C_1 and C_2 may be different for the three schemes considered here.

9.8c PPSWR AND SRS WOR AT THE TWO STAGES

Suppose a sample of n fsu's is drawn with ppswr and in the i -th selected fsu m_i ssu's are selected from the M_i ssu's with srs wor ($i = 1, 2, \dots, n$). If y_{ij} is the value of the j -th selected ssu in the i -th selected fsu, an unbiased estimator of Y , its variance and the variance estimator are respectively given by (9.25), (9.26) and (9.27) in Section 9.4.

Using this sample itself it is possible to estimate the variance of the estimator based on (i) a cluster sample of n fsu's selected with ppswr and (ii) a sample of nm units selected directly with srswr and hence the efficiency of two-stage sampling as compared to cluster sampling and uni-stage sampling can be estimated. The variance of the estimator of the population total in the case of cluster sampling, which amounts to uni-stage sampling of fsu's, is given by

$$V(\hat{Y}_c) = \frac{1}{n} \sum_{i=1}^N \left(\frac{M_i \bar{Y}_i}{P_i} - Y \right)^2 P_i. \quad \dots \quad (9.55)$$

This is the same as the between-fsu component of the variance in two-stage sampling with ppswr at the first stage and hence it can be estimated unbiasedly by subtracting an unbiased estimator of the within-fsu component from $v_t(\hat{Y}_t)$ given in (9.27). Thus we get,

$$v_t(\hat{Y}_c) = v_t(\hat{Y}_t) - \frac{1}{n^2} \sum_{i=1}^n \frac{M_i^2}{P_i^2} (1-f_i) \frac{s_{wi}^2}{m_i}. \quad \dots \quad (9.56)$$

An unbiased estimator of Y in the case of uni-stage sampling of nm units with srswr is $NM'\bar{y}$, where \bar{y} is the sample mean, and the variance of this estimator is

$$V(\hat{Y}_r) = N^2 M'^2 \frac{\sigma^2}{nm} = \frac{1}{nm} \left[NM' \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij}^2 - Y^2 \right]. \quad \dots \quad (9.57)$$

Proceeding as for $v(\hat{Y}_r)$ given in (9.53) in Sub-section 9.8b, we get an unbiased estimator of $V(\hat{Y}_r)$ based on a two-stage sample as

$$v_t(\hat{Y}_r) = \frac{1}{nm} \left[\frac{NM'}{nm} \sum_{i=1}^n \frac{M_i}{P_i} \sum_{j=1}^{m_i} y_{ij}^2 - \hat{Y}_t^2 + v_t(\hat{Y}_t) \right]. \quad \dots \quad (9.58)$$

The efficiency per unit cost and the relative efficiencies can be compared using these estimators and the corresponding values of the cost of survey.

9.9 AN ILLUSTRATIVE EXAMPLE

To study the efficiency of two stage sampling as compared to uni stage sampling the plot wise and village wise figures for the area under autumn paddy available for Nadia district of West Bengal have been utilized. The sampling designs compared are

- (i) selection of n villages with ppawr, size being geographical area (ppa) and selection of m plots from each sample village with ppawr and

- (ii) selection of nm plots directly from the population with ppawr.

It can be easily verified that unbiased estimators of the total area under autumn paddy in these two cases are given by

$$\hat{Y}_1 = \frac{G}{nm} \sum_{i=1}^n \sum_{j=1}^m p_{ij} \quad \text{and} \quad \hat{Y}_2 = \frac{G}{nm} \sum_{i,j}^{nm} p_{ij},$$

where p_{ij} is the ratio of area under paddy to the geographical area for the sample plot and G is the total geographical area. If the proportion of the area under autumn paddy in a plot is assumed to be either 0 or 1 the variances of \hat{Y}_1 and \hat{Y}_2 can be shown to be

$$V(\hat{Y}_1) = \frac{G}{n} \left[GPQ - \frac{m-1}{m} \sum_{i=1}^N G_i P_i Q_i \right] \quad (9.59)$$

and

$$V(\hat{Y}_2) = G^2 PQ/nm, \quad (9.60)$$

where P_i is the proportion of the crop area in the i th village, P the overall crop proportion, $Q_i = 1 - P_i$, and $Q = 1 - P$ (cf. Section 6.8 of Chapter 6 p. 197)

The values of the sampling efficiency, cost ratio and the cost efficiency are given in Table 9.5 assuming $C_0 = 500$, $C_1 = 15$ and $C_2 = 1$. It is found from this table that though the sampling efficiency decreases with increase in m for a given total number of sample ssu's, the cost efficiency increases upto a certain stage showing

that there is an optimum division of the total sample size between the two stages. In this case $V(\hat{Y}_1)$ is of the form

$$V(\hat{Y}_1) = \frac{A_1}{n} + \frac{A_2}{nm},$$

and the ratio of A_2 to A_1 turned out to be about 6. Taking the total cost C as fixed at 1000, the optimum values of m and n are given by 10 and 20 respectively.

TABLE 9.5. EFFICIENCY OF TWO-STAGE SAMPLING IN ESTIMATING CROP PROPORTION.

no. of sample villages	no. of plots per village	sampling efficiency (%)	cost ratio C_r/C_t	cost efficiency (%) (3) \times (4)
(1)	(2)	(3)	(4)	(5)
300	1	100	1.00	100
30	10	47	4.24	199
20	15	39	4.82	188
15	20	31	5.17	160

(total number of villages : 1408; area under autumn paddy : 276471 acres.)

Source : Crop data for Nadia District in West Bengal.

9.10 THREE-STAGE SAMPLING DESIGN

The treatment of two-stage sampling considered in the previous sections can easily be extended to the case of sampling with more than two stages. Suppose a sample of nml units are selected in three stages adopting ppswr at each stage. Let n be the number of sample fsu's, m the number of ssu's selected from each sample fsu and l the number of third stage units (tsu's) drawn from each sample ssu and let the probabilities of selection at each draw for the three stages be $\{P_i\}$, $\{P_{ij}\}$ and $\{P_{ijk}\}$, ($i = 1, 2, \dots, N$), ($j = 1, 2, \dots, M_i$), ($k = 1, 2, \dots, L_{ij}$). If y_{ijk} denotes the value of the k -th tsu, ($k = 1, 2, \dots, l$), in the j -th ssu, ($j = 1, 2, \dots, m$), of the i -th fsu, ($i = 1, 2, \dots, n$), in the sample, an estimator of Y is given by

$$\hat{Y} = \frac{1}{nml} \sum_{i=1}^n \frac{1}{p_i} \sum_{j=1}^m \frac{1}{p_{ij}} \sum_{k=1}^l \frac{y_{ijk}}{p_{ijk}}. \quad \dots \quad (9.61)$$

Noting from Section 2.8 of Chapter 2 (p. 43) that the expected value and the variance of this estimator can be written as

$$E(\hat{Y}) = E_1 E_2 E_3(\hat{Y})$$

and

$$V(\hat{Y}) = V_1 E_2 E_3(\hat{Y}) + E_1 V_2 E_3(\hat{Y}) + E_1 E_2 V_3(\hat{Y}),$$

it can be verified that \hat{Y} given in (9.61) is unbiased for Y , and that the variance of this estimator is given by

$$\begin{aligned} V(\hat{Y}) &= \frac{1}{n} \left(\sum_{i=1}^N \frac{Y_i^2}{P_i} - Y^2 \right) + \frac{1}{nm} \sum_{i=1}^N \frac{1}{P_i} \left(\sum_{j=1}^{M_i} \frac{Y_{ij}^2}{P_{ij}} - Y_i^2 \right) \\ &\quad + \frac{1}{nml} \sum_{i=1}^N \frac{1}{P_i} \sum_{j=1}^{M_i} \frac{1}{P_{ij}} \left(\sum_{k=1}^{L_{ij}} \frac{Y_{ijk}^2}{P_{ijk}} - Y_{ij}^2 \right) \quad \dots \quad (9.62) \end{aligned}$$

Since the selection at the first stage is done with replacement, the n unbiased estimates

$$t_i = \frac{1}{p_i} \frac{1}{m} \sum_{j=1}^m \frac{1}{p_{ij}} \left(\frac{1}{l} \sum_{k=1}^l \frac{y_{ijk}}{p_{ijk}} \right)$$

are independent and have the same variance. Hence, an unbiased variance estimator of $\hat{Y} = t$ is given by

$$v(\hat{Y}) = \frac{1}{n(n-1)} \left(\sum_{i=1}^n t_i^2 - n\bar{t}^2 \right) \quad \dots \quad (9.63)$$

We may note that in this case the variance function is of the form

$$V(\hat{Y}) = \frac{A_1}{n} + \frac{A_2}{nm} + \frac{A_3}{nml} \quad \dots \quad (9.64)$$

and a simple type of cost function is given by

$$C = C_0 + nC_1 + nmC_2 + nmlC_3, \quad \dots \quad (9.65)$$

where C_0 is the overhead cost, C_1 and C_2 are the costs of preliminary operations involved in a sample fsu and in a sample ssu respectively and C_3 is the cost of enumeration and analysis per tsu. If the cost

is fixed at C' , the values of n , m and l , which minimize the variance, are given by

$$n = \frac{(C' - C_0) \sqrt{A_1/C_1}}{\sqrt{A_1/C_1} + \sqrt{A_2/C_2} + \sqrt{A_3/C_3}}, \quad m = \sqrt{\frac{A_2/C_1}{A_1/C_2}}, \quad l = \sqrt{\frac{A_3/C_2}{A_2/C_3}}. \quad \dots \quad (9.66)$$

Substituting these values of n , m and l in (9.64), we get the minimum variance for a fixed cost C' as

$$V(\hat{Y}) = \frac{1}{C' - C_0} (\sqrt{A_1/C_1} + \sqrt{A_2/C_2} + \sqrt{A_3/C_3})^2. \quad \dots \quad (9.67)$$

If the variance is fixed at V_0 instead of the cost being fixed, then the values of n , m and l , which minimize the cost, are given by

$$n = \frac{\sqrt{A_1/C_1} + \sqrt{A_2/C_2} + \sqrt{A_3/C_3}}{V_0 \sqrt{C_1/A_1}}, \quad m = \sqrt{\frac{A_2/C_1}{A_1/C_2}}, \quad l = \sqrt{\frac{A_3/C_2}{A_2/C_3}}. \quad \dots \quad (9.68)$$

Substituting these values of n , m and l in (9.65), we get the minimum cost for ensuring a specified variance V_0 as

$$C = C_0 + \frac{1}{V_0} (\sqrt{A_1/C_1} + \sqrt{A_2/C_2} + \sqrt{A_3/C_3})^2. \quad \dots \quad (9.69)$$

9.11 MULTI-SUBJECT SURVEYS

In multi-stage sampling, when the number of sample fsu's is increased, the sampling variance decreases considerably, but the cost increases. Hence, to attain the highest possible precision per unit of cost for estimates of population parameters, one has to evolve methods (i) to decrease the variability at all the stages of sampling, and (ii) to reduce the cost of survey operations at the different stages. As regards (i), we have a number of sampling schemes, namely, srs, systematic sampling, pps sampling and stratified sampling, at our disposal for using them at the different stages of sampling to reduce sampling variability. For instance, one can collect some auxiliary information about all the ssu's in the sample fsu's at the time of listing of such units for being used in arrangement,

stratification and selection of the sample of ssu's thereby reducing the within fsu sampling variance. We shall consider here briefly point (ii) relating to the possibility of reducing the cost of survey or of fully utilizing the available resources through *integration* of two or more surveys.

Suppose we are required to carry out sample surveys to provide estimates on the demographic characteristics and on the level of living of the rural population. Noting that household is a convenient ultimate sampling unit in these two cases we may adopt a two-stage sampling design with villages as fsu's and households as ssu's. If the two surveys are carried out independently with n_1 and n_2 , ($n_1 > n_2$) sample fsu's and $n_1 m_1$ and $n_2 m_2$ ($m_1 > m_2$) sample ssu's for attaining the specified precisions in the two cases, then the total cost of the survey would be of the form

$$C = 2C_0 + (n_1 + n_2)C_1 + n_1 m_1 C_{21} + n_2 m_2 C_{22}, \quad (9.70)$$

where C_{21} and C_{22} denote the cost of survey per ssu for the two surveys. If we can integrate the two surveys in the sense of having a common set of sample fsu's for them, that is if we have the sample of n_2 fsu's required for the second survey as a sub sample of the n_1 sample fsu's selected for the first survey, then the cost reduces to

$$C = C_0 + n_1 C_1 + n_1 m_1 C_{21} + n_2 m_2 C_{22}, \quad (9.71)$$

the reduction in cost being $C_0 + n_2 C_1$ which will generally be substantial since C_1 is likely to be considerably larger than C_{11} and C_{12} and C_0 being the overhead cost will also be large. It may be mentioned that due to this integration the overhead cost C_0 and the cost per fsu C_1 may get increased but this increase would be quite small compared to the original values of C_0 and C_1 . There is also the possibility of C_{21} and C_{22} getting reduced due to integration on account of a possible reduction in idle time in the fsu's.

The reduction in cost achieved by integrating the two surveys can be utilized (i) to increase the sample fsu's for the second survey from n_2 to n_1 with a view to increasing the precision of the estimates, since

it could be done at a marginal cost as the preliminary operations in the $(n_1 - n_2)$ sample fsu's are being done for the first survey, the marginal cost requirement being only $(n_1 - n_2)m_2C_{22}$, (ii) to increase the number of sample fsu's for both the surveys, thereby either reducing the sampling variability in both the surveys or reducing the number of sample ssu's required to give the same precision, and (iii) to carry out a survey on a third subject in some or all of the fsu's selected for the first survey.

It may be beneficial to integrate two or more surveys even when the sampling units at the ultimate stage are different. For instance, we may integrate a family budget enquiry with household as the ssu with a crop survey with plot or cluster of plots as the ssu by having a common set of sample fsu's, which may be village or some other area units. It is possible that in some cases the sampling design might become less efficient for one or more of the integrated surveys from the point of view of sampling variance because of the need to have a common design and a common sample up to at least the first stage, but this loss in efficiency is likely to be compensated by the increase in the sample fsu's made possible through integration. In such situations one may also explore the possibility of having a partial integration in the sense of having a common sample supplemented by special samples for those subjects, for which the sample design has become inefficient through integration.

9.12 MULTI-PHASE SAMPLING

Another method of selection, which is similar to sampling in stages consists in selecting a sample of units in the first phase for collecting data on some suitable auxiliary variables, and then selecting a sub-sample of these units for the main survey, by utilizing the auxiliary information obtained in the first phase for arrangement, stratification and selection or for estimation. This procedure is *termed two-phase sampling* and when the sample for the main survey is selected in two or more phases, the sampling procedure is termed *multi-phase sampling*.

The main difference between multi stage and multi phase sampling is that in the former, clusters of next stage units are selected at each draw, whereas in the latter, at each phase (excepting the first) a sub sample of the sampling units selected in the previous phase is drawn. That is in the case of multi phase sampling it is necessary to have a complete frame of ultimate units, as at each phase a sample of ultimate units is selected whereas in multi stage sampling a frame of the next stage units is required only for the sample units selected at any stage. This design is resorted to mainly to reduce the cost of survey by collecting data on some suitable auxiliary variables, which are easy and cheaper to observe, for the units in an initial sample and selecting a sub sample of the initial sample in as efficient manner as possible for the main survey.

An illustration of the use of multi phase sampling may be given by considering the question of estimating the total consumer expenditure in a town through a sample survey, when only just a list of all households in the town is available, without any other particulars about the households. One procedure is to select a sample of households and collect data on consumer expenditure, but such a procedure may require a rather large sample and hence the cost involved may be considerable if there is large variation among the households. An alternative procedure in such a case, which is likely to be more economical would be to collect data on some simple characteristics related to consumer expenditure such as household size, means of livelihood etc for a sample of households selected in the first phase and to use this information for arrangement, stratification and selection of the second phase sample of households for the collection of the data on consumer expenditure. The former is *uni phase sampling* whereas the latter method is *two phase sampling*. It may be noted that different sampling procedures may be used at the different phases depending on the information available for the sample units.

As regards the question of estimation of population total in multi phase sampling, the procedure consists in estimating the total of the sample units at any phase on the basis of the sub sample selected in the next phase.

For instance, suppose a two-phase sample is selected by first selecting a sample of n_1 units with srs w/o r and then a sub-sample of n_2 units from the initial sample is drawn with ppswr using a suitable measure of size collected in the first phase, then an unbiased estimator of the population total Y is given by

$$\hat{Y} = \frac{N}{n_1} \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{y_j}{p_j}, \quad p_j = \frac{x_j}{x'}, \quad \dots \quad (9.72)$$

where y_j is the value of y for the j -th unit in the sub-sample and x' is the total of x for the first phase sample. For

$$E(\hat{Y}) = E_1 E_2(\hat{Y}) = E_1 \left\{ \frac{N}{n_1} \sum_{i=1}^{n_1} y_i \right\} = Y,$$

where y_i denotes the value of the i -th unit in the first phase sample.

The variance of \hat{Y} can be obtained by noting that

$$V(\hat{Y}) = V_1 E_2(\hat{Y}) + E_1 V_2(\hat{Y}),$$

and we have

$$\begin{aligned} V(\hat{Y}) &= V_1 \left\{ \frac{N}{n_1} \sum_{i=1}^{n_1} y_i \right\} + E_1 \left[\frac{N^2}{n_1^2} - \frac{1}{n_2} \left\{ x' \sum_{i=1}^{n_1} \frac{y_i^2}{x_i} - \left(\sum_{i=1}^{n_1} y_i \right)^2 \right\} \right] \\ &= N^2 \frac{N-n_1}{N-1} \frac{\sigma_y^2}{n_1} + \frac{N}{N-1} \frac{n_1-1}{n_1} \frac{V_p}{n_2}, \quad \dots \quad (9.73) \end{aligned}$$

where V_p stands for the variance of the estimator of Y based on one sample unit selected with pps from the whole population (Des Raj, 1964). If n_1 is large, $(n_1-1)/n_1$ may be taken to be approximately equal to unity and hence the variance may be written as

$$V(\hat{Y}) = \frac{A_1}{n_1} + \frac{A_2}{n_2} + A_3, \quad \dots \quad (9.74)$$

where A_1 , A_2 and A_3 are constants.

The cost function in this case is of the form

$$C = C_0 + C_1 n_1 + C_2 n_2, \quad \dots \quad (9.75)$$

where C_1 is the cost per unit for observing and using the value of the auxiliary variable and $C_2 (> C_1)$ is the cost per unit for obtaining and tabulating the value of the study variable. It can be easily shown that the optimum values of n_1 and n_2 , which minimize the sampling variance for a given cost C , are

$$n_1 = \frac{(C - C_0) \sqrt{A_1/C_1}}{\sqrt{A_1 C_1} + \sqrt{A_2 C_2}} \quad (9.76)$$

and

$$n_2 = \frac{(C - C_0) \sqrt{A_2/C_2}}{\sqrt{A_1 C_1} + \sqrt{A_2 C_2}}, \quad (9.77)$$

and that those which minimize the cost for ensuring a given value V_0 for the variance, are given by

$$n_1 = \frac{\sqrt{A_1 C_1} + \sqrt{A_2 C_2} \sqrt{\frac{A_1}{C_1}}}{V_0 - A_3} \quad (9.78)$$

and

$$n_2 = \frac{\sqrt{A_1 C_1} + \sqrt{A_2 C_2} \sqrt{\frac{A_2}{C_2}}}{V_0 - A_3} \quad (9.79)$$

The variance of \hat{Y} given in (9.73) can be unbiasedly estimated by considering suitable statistics based on the sample observations (cf 10.3 of Appendix 2). The use of multi phase sampling for applying ratio and regression methods of estimation is discussed in Chapters 10 and 11.

9.13 COMPOSITE SAMPLING DESIGNS

A sampling design where two or more of the basic sampling schemes discussed in the previous chapters are simultaneously used may be termed a *composite sampling design*. For instance, multi stage sampling is a composite design since we select clusters of units at the first stage and subsequent stages and then select the units at the ultimate stage. Further, the selection procedures used at the different stages may also be different. Multi phase sampling is also a composite design, since here also usually the method of

sampling varies over the phases depending on the amount and the nature of available supplementary information.

Multi-stage and multi-phase sampling designs may also be used after suitable stratification of the population. Suppose we adopt a stratified two-stage sampling design, then an unbiased estimator of the population total can be obtained by just addition of the unbiased estimators of stratum totals over the strata. In that case, the variance function will be of the form

$$V(\hat{Y}) = \sum_{s=1}^K \left\{ \frac{A_{1s}}{n_s} + \frac{A_{2s}}{n_s m_s} + A_{3s} \right\} \quad \dots \quad (9.80)$$

and the cost function may be taken as

$$C = C_0 + \sum_{s=1}^K n_s C_{1s} + \sum_{s=1}^K n_s m_s C_{2s}, \quad \dots \quad (9.81)$$

where subscript s denotes the s -th stratum. The optimum values of n_s and m_s can be obtained by minimizing the variance for a fixed cost or by minimizing the cost for a specified variance applying the usual methods to (9.80) and (9.81).

REFERENCES

- COCHRAN, W. G. (1939) : The use of analysis of variance in enumeration by sampling; *J. Amer. Stat. Assn.*, 34, 492-510.
- DES RAJ (1964) : On double sampling for pps estimation; *Ann. Math. Stat.*, 35, 900-902.
- GANGULI, M. (1941) : A note on nested sampling; *Sankhyā*, 5, 449-452.
- HANSEN, M. H. and HURWITZ, W. N. (1943) : On the theory of sampling from finite populations; *Ann. Math. Stat.*, 14, 333-362.
- LAHIRI, D. B. (1954) : Technical paper on some aspects of the development of the sample design; *Sankhyā*, 14, 264-316.
- MAHALANOBIS, P. C. (1940) : *Report on the Sample Census of Jute in Bengal*, 1939; Indian Central Jute Committee.
- RANGARAJAN, R. (1957) : A note on two-stage sampling; *Sankhyā*, 17, 373-376.
- RAO, J. N. K. (1961) : On sampling with varying probabilities in sub-sampling designs; *J. Ind. Soc. Agr. Stat.*, 13, 211-217.

- ROY, J (1957) A note on estimation of variance components in multi stage sampling with varying probabilities, *Sankhya*, 17, 367-372
- SINGH, D (1958) Estimates of variance components in finite population, *J Ind Soc Agr Stat*, 10, 1-15

COMPLEMENTS AND PROBLEMS

9.1 To estimate the total number of words (Y) in an English dictionary, 10 out of 26 alphabets were selected with ppswr, size being the number of pages devoted to an alphabet and for each selected alphabet two pages were selected with srs wr. The relevant sample data are given in Table 9.6

TABLE 9.6 NUMBER OF WORDS IN A SAMPLE OF PAGES

sr no	sample alphabet	no of pages devoted	no of words in sample page	
			1	2
(1)	(2)	(3)	(4)	(5)
1	S	131	34	27
2	C	97	27	26
3	N	21	44	38
4	S	131	24	29
5	F	43	25	32
6	J	7	42	48
7	U	18	24	21
8	P	85	53	24
9	A	49	47	55
10	D	54	38	57

(Total number of pages in the dictionary is 980)

- (i) Estimate unbiasedly Y and obtain an estimate of its rse
- (ii) Estimate also the efficiency of the above method of sampling compared to that of drawing 20 pages from the dictionary with srs wr

9.2 In a sample survey for estimating the number of standards of pepper in a tehsil having 72 villages, a sample of 12 villages was selected with srs wr and from each sample village 5 clusters of 20 fields each were drawn with srs wr. Data on number of clusters in the sample villages and on the number of standards in the sample clusters are given in Table 9.7

TABLE 9.7. NUMBER OF STANDARDS OF PEPPER IN SAMPLE CLUSTERS.

sample village	no. of clusters	number of standards in sample clusters				
		1	2	3	4	5
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	27	430	402	363	975	389
2	24	586	1234	100	368	344
3	14	1164	546	3060	1724	1274
4	116	693	218	836	1218	575
5	25	191	270	4502	4184	243
6	118	1036	1333	1179	728	1957
7	147	1555	254	950	382	355
8	26	910	452	129	122	243
9	91	340	0	92	28	340
10	171	57	59	0	0	21
11	86	159	45	242	1075	539
12	88	84	462	147	16	10

Estimate unbiasedly the total number of standards in the tehsil and obtain its rse by estimating unbiasedly its variance.

9.3 For estimating the average household expenditure \bar{Y} in a region, it is proposed to adopt a two-stage sampling design, where villages in the first stage and households in the second stage would be selected with srswr. To help in planning the survey, a pilot survey was conducted and it was found that (a) estimate of \bar{Y} is 50, (b) estimate of between-village variation $\sigma_b^2 = 85.5$, (c) estimate of between-household variation within villages $\sigma^2_{hv} = 36.5$, (d) cost of travel, etc., per village $C_1 = \text{Rs. } 9$ and (e) cost of survey per household $C_2 = \text{Re. } 1$. Using this information and assuming the overhead cost to be Rs. 1000, determine the optimum number of sample villages and number of households to be sampled per sample village, when the total cost is fixed at (i) Rs. 5000, (ii) Rs. 10000, and (iii) Rs. 50000. Also calculate the minimum rse's attained in the three cases.

9.4. Raw wool contains varying amounts of grease, dirt and other impurities and its quality is measured by the percentage of the weight of clean wool to that of raw wool, termed *clean content*. To estimate the clean content an electrical core-boring machine is used, which takes cores of about 1/4 lb. from a bale, which are then subjected to laboratory analysis. In an experiment 6 bales were drawn from a large lot with equal probability and from each bale 4 cores were taken at random and clean content was determined. The results of this experiment are given in Table 9.8.

TABLE 9.8 THE CLEAN CONTENT OF WOOL FOR 24 CORES

core	sample bales					
	1	2	3	4	5	6
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	54.3	57.0	54.6	54.9	59.9	57.8
2	56.2	58.7	57.5	60.1	57.8	59.7
3	58.9	58.2	59.3	58.7	60.3	59.6
4	55.5	57.4	57.5	55.6	57.5	58.1

(i) Estimate the average clean content of wool (\bar{x}) for the lot and also obtain an estimate of its rse

(ii) Obtain the efficiency of sampling 12 bales and 2 cores from each bale as compared to that of the above scheme

9.5 For estimating the average catch of fish landed per operating fishing unit (\bar{Y}) all the fish landing centres in the coast were grouped into two strata having equal number of centres, and in each stratum 5 centres were selected with srswt and from each sample centre 3 operating units were similarly selected. The survey was carried out in two seasons using the same design but independently drawn samples and the results are given in Table 9.9

TABLE 9.9 TOTAL CATCH OF FISH FOR SAMPLE OPERATING UNITS

stratum number	sample centre	season 1			season 2		
		1	2	3	1	2	3
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1	610	454	618	112	124	200
	2	297	400	515	56	75	58
	3	1180	112	480	187	160	363
	4	1297	533	1130	226	175	520
	5	860	357	657	67	137	616
2	1	1085	754	980	434	204	185
	2	817	736	926	160	101	290
	3	900	616	109	391	565	150
	4	510	320	412	420	186	617
	5	906	900	735	144	36	21

(Catch of fish is in kilogrammes)

(i) Assuming that the total number of operating units is the same for each centre in a given season and that this number is the same in season 1 and in season 2, estimate unbiasedly the average catch per operating unit (\bar{Y}) for the entire coast covering both the fishing seasons and also estimate its rse.

(ii) Calculate the number of centres required to estimate the average catch with an rse of 3% when optimum allocation to the four sub-strata (2 strata \times 2 seasons) is used and when the number of operating units to be selected at a centre is 3.

9.6. It is proposed to draw a sample of n clusters of M units each from a population of N clusters and a sub-sample of m units from each sample cluster using srswr at both the stages for estimating the mean per unit of a specified characteristic.

(i) Assuming the cost function to be of the form

$$C = C_0 + C_1 n + C_2 nm,$$

determine the optimum values of m and n when C is fixed at C' .

(ii) Given that $C' = 1000$, $C_0 = 300$, $C_1 = 9$ and $C_2 = 1$ (in rupees), find the optimum values of m and n using the following analysis of variance table.

TABLE 9.10. ANALYSIS OF VARIANCE FOR THE STUDY VARIABLE.

source of variation	degrees of freedom	sum of squares	mean square (3)/(2)
(1)	(2)	(3)	(4)
between clusters	89	$20 \sum_{t=1}^{90} (\bar{Y}_t - \bar{\bar{Y}})^2$	180.9
within clusters	1710	$\sum_{t=1}^{90} \sum_{j=1}^{20} (Y_{tj} - \bar{Y}_t)^2$	49.5
total	1799	$\sum_{t=1}^{90} \sum_{j=1}^{20} (Y_{tj} - \bar{\bar{Y}})^2$	56.0

(total number of clusters : 90; number of units in a cluster : 20).

9.7. For estimating the total yield of paddy (Y) in a district, a stratified two-stage sampling design was adopted, where 4 villages were selected from each stratum, with ppswr, size being geographical area, and 4 plots were drawn from each sample village circular systematically for ascertaining the yield of paddy. Using the information given in Table 9.11, estimate unbiasedly \bar{Y} and obtain an estimate of its rse.

TABLE 9.11 YIELD OF PADDY FOR THE SAMPLE PLOTS

stratum	sample village	inverse of probability	total no. of plots	yield of paddy (in kilogrammes)			
				1	2	3	4
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1	410.21	25	104	182	148	87
	2	660.43	14	108	64	132	156
	3	31.50	240	100	115	50	172
	4	113.38	76	346	350	157	119
2	1	21.00	256	124	111	135	216
	2	16.80	258	123	177	106	133
	3	24.76	212	264	78	144	55
	4	49.99	69	300	114	68	111
3	1	67.68	159	110	281	120	114
	2	339.14	42	80	61	118	124
	3	100.00	134	121	212	174	106
	4	68.07	161	243	116	314	129

9.8 For estimating the total number of cultivators (I) a sample of n villages is selected from the N villages in the population with ppswr size being the current number of households (M_i) and from each sample village m households are selected circular systematically. The number of cultivators in each of the nm households is determined. Suggest an unbiased estimator of I and obtain an unbiased variance estimator for it.

9.9 Derive the results given in (9.6) and (9.7) regarding the expected value and the variance of an estimator in the case of an r -stage design.

9.10 Obtain the results given in (9.13) and (9.15) in connection with the variance and the variance estimator in the case of two stage sampling with srswr in the first stage and str wr in the second stage.

9.11 In two stage sampling where n fsu's are selected with ppswr and from the i th sample fsu (containing M_i units) m_i ssu's are selected with str wr for estimating the population total ΣY , the sub sampling numbers (m_i) may be so fixed that

(i) expected value of m_i is fixed at m or

(ii) total number of sample ssu's is fixed at m_0 .

For these two cases obtain the optimum values of $\{m_i\}$ that would minimize the variance of the estimator and compare their minimum variances.

9.12 Suppose n fsu's are selected with ppswr and if the i -th unit occurs r_i times in the sample, then any of the following three procedures may be adopted in sub-sampling :

- (i) $r_i m_i$ ssu's are selected with srs wor;
- (ii) r_i independent samples of m_i units each (selected with srs wor) are drawn; and
- (iii) m_i units are selected with srs wor and the observations are weighted by r_i .

Obtain the variances of unbiased estimators of population totals for these three cases and compare their efficiencies.

(Rao, J. N. K., *J. Ind. Soc. Agr. Stat.*, 13, (1961), 211–217).

9.13 Derive the results given in (9.38) to (9.40), (9.43) and (9.44) relating to optimum values of n , m , variance and cost in the case of two-stage sampling.

9.14 Obtain the expression (9.62) for the variance of the estimator of Y in the case of three-stage sampling with ppswr design.

9.15 Derive the results given in (9.66) to (9.69) relating to the optimum values of n , m , l , variance and cost in the case of three-stage sampling.

9.16 If in a stratified two-stage sampling design, the samples at the first and the second stages are drawn in the form of n and m independent sub-samples of equal size, show how the components of the variance of the estimator of the population total can be unbiasedly estimated using the nm possible estimates.

(Murthy, M. N., *Metrika*, 13).

9.17 Suppose n fsu's are selected with pps wor and from each sample fsu, m ssu's are selected with srs wor. Suggest an unbiased estimator of the population total on the lines proposed by Des Raj in the case of a uni-stage design (cf. Section 6.11a of Chapter 6, p.212), and obtain its unbiased variance estimator.

(Des Raj, *J. Amer. Stat. Assn.*, 51, (1956), 269–284).

9.18 A sample of N balls is drawn with srs from a large supply of MP red and MQ white balls ($P+Q=1$). From these N balls, n balls are drawn with srs and r of these n balls are found to be red. Find the expected value and the variance of r , when the sample of n balls is drawn (a) with replacement and (b) without replacement. Obtain unbiased estimators of the variances in the two cases.

9.19 Consider the following two sampling schemes for estimating the population mean of a characteristic: (a) the population is divided into N clusters of M units each and two-stage sampling is adopted where n clusters and m units from each sample cluster are selected with srswr; and (b) the population is divided into clusters of m' units each and a sample of n' such clusters is selected with srswr. Show that in both

the cases the sample mean is unbiased for the population mean and derive the variances in the two cases. Determine the condition for the efficiencies of these two schemes to be the same, when $n_m = n \cdot m$.

(Singh, D., *J. Ind. Soc. Agr. Stat.*, 8, (1956), 45-55)

9.20 Suppose a population of M units has been grouped into K unequal strata. To determine the strata sizes, a sample of n units is drawn from the whole population with srswr. Let n_s of the n sample units fall in the s th stratum, $s = 1, 2, \dots, K$. Using allocation proportional to n_s' , n_s units are selected from the s th stratum with srswr. If E_{ds} denotes the relative efficiency of the above double sampling stratified procedure for estimating the population mean as compared to unstratified srs and if E_s denotes the relative efficiency of stratified sampling with proportional allocation when the strata sizes are known, show that

$$E_{ds} = \frac{E_s}{1 + (n/n) (E_s - 1)}$$

(Cochran, W. G., *Sampling Techniques*, (1963), Ch. 12, p. 332)

Method of Ratio Estimation

10.1 NEED FOR RATIO ESTIMATION

So far, we have considered a number of sampling procedures for the estimation of population totals, means and proportions. But in practice, knowledge of the ratio of population totals of two characteristics is as important as, and sometimes more important than, that of population totals and means. For instance, in socio-economic surveys one may be interested in such ratios as per household and per capita income or expenditure, proportion of expenditure on different items, proportion of unemployed persons, sex-ratio, birth-rate, death-rate, etc. Similarly, estimation of yield rates in a crop survey and input-output ratios in an industrial survey are of considerable importance. In estimating such ratios, the commonly used procedure has been to take the ratio of unbiased estimators of the numerator and the denominator of the population ratio as an estimator and such an estimator is termed *ratio estimator*. It may be mentioned that it may be necessary to use a ratio estimator even for estimating the population or sub-population mean and proportion, if the total number of units is not known, (Sub-section 3.7b of Chapter 3, p. 72 and Sub-section 9.3d of Chapter 9, p. 326).

In situations where the actual value of the denominator of the ratio is known, one may feel that it is sufficient to estimate only the numerator and that the population ratio could be estimated by dividing this estimator by the known value of the denominator. Such an estimator may not necessarily be very efficient as compared to

the ratio of the estimators of the numerator and the denominator. In fact if the estimators of the numerator and the denominator are approximately proportional, that is, if there is a linear regression between them and the regression line passes through the origin, the latter procedure of estimation becomes more efficient than the former. This is due to the fact that in this case the ratio of the estimators would be more stable than the ratio of the estimator of the numerator to the actual value of the denominator. In such a case it would be profitable to use the product of the ratio of the estimators and the actual value of the denominator as an estimator of the parameter appearing in the numerator, and this procedure of estimation is termed *ratio method of estimation*.

Suppose the ratio $R = Y/X$, where Y and X are the population totals for the variables y and x is to be estimated on the basis of a sample selected through any given sampling scheme. Let \hat{Y} and \hat{X} be unbiased estimators of Y and X respectively. Then an estimator of the ratio R is given by

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \quad (10.1)$$

Similarly, a ratio estimator of Y is given by

$$\hat{Y}_R = \hat{R}X = \frac{\hat{Y}}{\hat{X}} X, \quad (10.2)$$

when information on the total X of a related auxiliary variable x is available. It is to be noted that for estimating a total by the method of ratio estimation, we should know the value of X from some other source. Cochran (1940) has considered the problem of ratio method of estimation based on some suitable auxiliary variable.

In what follows the expected value and the mean square error of different types of ratio estimators are considered. The discussion applies equally well to the case of estimating Y by the ratio method of estimation and the expression for the expected value and the mse in the latter case can be obtained by multiplying the corresponding

expressions for the former by X and X^2 respectively. The treatment given in this chapter is quite general and may be applied to any sampling design.

10.2 BIAS OF RATIO ESTIMATOR

In the earlier chapters, we had confined ourselves to unbiased estimators. It is to be noted that a biased estimator may be preferred to an unbiased estimator, if the mse of the former is less than the variance of the latter. For the commonly used selection procedures, the ratio estimator given above is, in general, biased for the corresponding population ratio. For instance, if \bar{y} and \bar{x} are the sample means of the characteristics y and x respectively based on a sample of n units selected from a population of N units with srs w/o, the estimator $\hat{R} = \bar{y}/\bar{x}$ is biased for the ratio $R = Y/X$, since

$$E(\hat{R}) = \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \frac{\bar{y}_s}{\bar{x}_s}$$

which is not generally equal to R . It may be noted that an alternative estimator in this case would be $\hat{R}' = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}$ and that this estimator is also biased for R , since

$$E(\hat{R}') = \frac{1}{N} \sum_{i=1}^N \frac{Y_i}{X_i} \neq R.$$

For deriving the expected value and hence the expression for the bias of the ratio estimator (\hat{Y}/\hat{X}) in the general case, we may proceed in the following way. Let $e = (\hat{Y} - Y)/Y$ and $e' = (\hat{X} - X)/X$. Since \hat{Y} and \hat{X} are unbiased for Y and X , $E(e) = E(e') = 0$, $V(e) = V(\hat{Y})/Y^2$, $V(e') = V(\hat{X})/X^2$ and $\text{Cov}(e, e') = \text{Cov}(\hat{X}, \hat{Y})/XY$. Substituting $Y(1+e)$ and $X(1+e')$ for \hat{Y} and \hat{X} in (10.1), we get

$$\hat{R} = R \frac{1+e}{1+e'} \quad \dots \quad (10.3)$$

The expected value of \hat{R} is given by $E(\hat{R}) = RE\{(1+e)(1+e')\}$. If it is assumed that $|e'| < 1$, the second term may be expanded as an infinite series which would be convergent. This assumption would mean that all the possible estimates \hat{X} of X lie between 0 and $2X$, which is likely to be so if the variation in \hat{X} is not large. For the variation in \hat{X} to be small, the sample size has to be fairly large. With this assumption we get

$$\begin{aligned} E(\hat{R}) &= RE[(1+e)(1-e'+e^2 - \dots)] \\ &= RE[1+(e-e')+(e'^2-ee')+\dots] \end{aligned} \quad (10.4)$$

If we further assume that terms involving second and higher powers of (e, e') in the above expression would be negligible due to the fact that e and e' are likely to be small quantities when the sample size is large we have $E(\hat{R}) = R$ since $E(e-e') = 0$. This would mean that to the first order of approximation, the ratio estimator is unbiased. But if we assume that only terms involving powers of (e, e') greater than 2 could be considered to be negligibly small since this would be more realistic than the former assumption, we get the bias of \hat{R} as

$$\begin{aligned} B(\hat{R}) &= E(\hat{R}-R) = R \left[\frac{V(\hat{X})}{X^2} - \frac{\text{Cov}(\hat{X}, \hat{Y})}{XY} \right] \\ &= \frac{1}{X^2} [RV(\hat{X}) - \text{Cov}(\hat{X}, \hat{Y})] \end{aligned} \quad (10.5)$$

In most of the sampling designs ordinarily used, $B(\hat{R})$ decreases with increase in sample size. Equating (10.5) to zero, we find that the approximate expression for the bias in the ratio estimator becomes zero if $R = \text{Cov}(\hat{X}, \hat{Y})/V(\hat{X})$, which condition is satisfied when the regression line of \hat{Y} on \hat{X} is a straight line passing through the origin (cf Problem 10.11, p. 401).

It is to be noted that in arriving at the expression (10.5) for the bias, two assumptions have been made :

- (i) $\left| \frac{\hat{X} - X}{X} \right| < 1$, that is, \hat{X} lies between 0 and $2X$ and
- (ii) terms of degree greater than 2 in (e, e') in the expansion for $(1+e)(1+e')^{-1}$ can be neglected.

Both these assumptions are likely to be valid only for large samples, unless the population is fairly homogeneous in which case even with a smaller sample size the two assumptions may be valid. It may also happen that even if in some samples $|e'| \geq 1$, the expression (10.5) may still be a good approximation to the bias of the ratio estimator provided such samples are negligibly few. It has been found empirically by some authors that the above assumptions hold for most of the populations usually met with in practice, if the sample size is appreciably large. This result is to be taken with reservation because the validity or otherwise of the above assumptions depends much on the actual population under consideration and the sampling design adopted. The behaviour of the bias and the closeness or otherwise of the approximate expression for the bias have been empirically studied in Section 10.7.

The bias of the ratio estimator given in (10.5) can be estimated by substituting in it the estimators of X , R , $V(\hat{X})$ and $\text{Cov}(\hat{X}, \hat{Y})$. Thus, an estimator of the bias is given by

$$b(\hat{R}) = -\frac{1}{\hat{X}^2} [\hat{R} v(\hat{X}) - \text{cov}(\hat{X}, \hat{Y})]. \quad \dots \quad (10.6)$$

For instance, if the sample is drawn in the form of m independent interpenetrating sub-samples and \hat{X} and \hat{Y} are the means of the m sub-sample estimates \hat{X}_i and \hat{Y}_i , then unbiased estimators of $V(\hat{X})$ and $\text{Cov}(\hat{X}, \hat{Y})$ are simply obtained as

$$v(\hat{X}) = \frac{1}{m(m-1)} \sum_{i=1}^m (\hat{X}_i - \hat{X})^2$$

and

$$\text{cov}(\hat{X}, \hat{Y}) = \frac{1}{m(m-1)} \sum_{i=1}^m (\hat{X}_i - \bar{\hat{X}})(\hat{Y}_i - \bar{\hat{Y}})$$

Substituting these estimators in (10.6), we get an estimator of bias as

$$b(\hat{R}) = \frac{1}{\hat{X}^2} \frac{1}{m(m-1)} \sum_{i=1}^m \hat{X}_i (\hat{R}\hat{X}_i - \hat{Y}_i) \quad (10.7)$$

It may be mentioned that $b(\hat{R})$ given in (10.6) and (10.7) are not unbiased estimators of $B(\hat{R})$ since they themselves are ratios of estimators. If the values of X and $V(\hat{X})$ are known, then these values may be substituted in the expression for bias instead of their estimates. Some other procedures of estimating the bias and correcting the ratio estimator for its bias are considered in Sections 10.9 and 10.10.

Without any assumptions, it can be shown that for sufficiently large sample size, the bias of the ratio estimator is likely to be negligible. For, the exact bias may be obtained as follows by noting that

$$\text{Cov}(\hat{R}, \hat{X}) = E(\hat{Y}) - E(\hat{R})E(\hat{X})$$

That is,

$$B(\hat{R}) = -\frac{\text{Cov}(\hat{R}, \hat{X})}{\bar{X}} = -\frac{1}{\bar{X}} \rho(\hat{R}, \hat{X}) \sigma(\hat{R}) \sigma(\hat{X}),$$

where $\rho(\hat{R}, \hat{X})$ is the correlation coefficient between \hat{R} and \hat{X} and $\sigma(\hat{R})$ and $\sigma(\hat{X})$ are their standard errors. Hence, the relative bias is given by

$$\left| \frac{B(\hat{R})}{\hat{R}} \right| = |\rho(\hat{R}, \hat{X})| C(\hat{R}) C(\hat{X}), \quad (10.8)$$

where $C(\hat{R})$ and $C(\hat{X})$ are the rse's of \hat{R} and \hat{X} respectively. From (10.8), we find that the bias of the ratio estimator would be small if the sample size is large, since in that case the rse's of \hat{R} and \hat{X} are likely to be small. It may be noted that $\rho(\hat{R}, \hat{X})$ is likely to be

nearly equal to zero, if \hat{Y} and \hat{X} are approximately proportional. Hence, the relative bias of the ratio estimator would be quite small, since it is actually the product of three quantities, one of which can be expected to be nearly equal to zero and the other two are likely to be less than one and very small for sufficiently large sample size. Further, it is possible to have an upper bound for the magnitude of the relative bias and it is given by

$$\left| \frac{B(\hat{R})}{R} \right| \leq C(\hat{R})C(\hat{X}), \quad \dots \quad (10.9)$$

(cf. Problem 10.8, p.401).

General Expression for Bias

Under the assumption that $|e'| < 1$, when the expression for \hat{R} can be represented by a convergent infinite series, the expression for the bias of \hat{R} can be written down to any order of approximation by taking the expected value of the successive terms in (10.4). Thus we get

$$\begin{aligned} B(\hat{R}) &= RE[(e - e') + (e'^2 - ee') + (ee'^2 - e'^3) + \dots] \\ &= R[(v_{20} - v_{11}) + (v_{21} - v_{30}) + \dots], \end{aligned} \quad \dots \quad (10.10)$$

where

$$v_{ij} = \frac{\mu_{ij}}{X^i Y^j} = \frac{E(\hat{X}^i - X^i)(\hat{Y}^j - Y^j)}{X^i Y^j}.$$

It is of interest to note that for a number of sampling schemes, v_{ij} is of the form θ_{ij}/n^{i+j-1} , where n is the sample size and θ_{ij} is a population parameter independent of n , showing that the higher order terms are expected to be negligible if n is sufficiently large.

10.3 MEAN SQUARE ERROR

Since the ratio estimator is biased, we have to consider its mean square error for the purpose of comparing its efficiency with that of any other estimator. The mse is, by definition,

$$M(\hat{R}) = E(\hat{R} - R)^2$$

and substituting for \hat{R} from (10.3), and simplifying, we have

$$M(\hat{R}) = R^2 E[(e - e')^2(1 + e')^{-2}].$$

Again assuming that $|e'| < 1$ and that the terms involving powers of (e, e') greater than 2 may be neglected, we get

$$\begin{aligned} M(\hat{R}) &= R^2 [E(e-e')^2] \\ &= R^2 \left[\frac{V(\hat{Y})}{\hat{X}^2} - 2 \frac{\text{Cov}(\hat{X}, \hat{Y})}{\hat{X}\hat{Y}} + \frac{V(\hat{X})}{\hat{X}^2} \right] \\ &= \frac{1}{\hat{X}^2} [V(\hat{Y}) - 2\hat{R} \text{Cov}(\hat{X}, \hat{Y}) + R^2 V(\hat{X})] \end{aligned} \quad (10.11)$$

The above assumption of neglecting terms of degree greater than 2 in (e, e') amounts to assuming that the bias is negligibly small since we have already seen that the bias obtained by neglecting terms involving second and higher powers of (e, e') is zero. In such a case the variance and the mse are identical.

As in the case of bias, it is possible to estimate the variance by substituting the estimators of X , R , $V(\hat{X})$, $V(\hat{Y})$ and $\text{Cov}(\hat{X}, \hat{Y})$ in (10.11). Thus we have

$$v(\hat{R}) = \frac{1}{\hat{X}^2} [v(\hat{Y}) - 2\hat{R} \text{cov}(\hat{X}, \hat{Y}) + \hat{R}^2 v(\hat{X})] \quad (10.12)$$

(cf Problem 10.15 p 402) When \hat{X} and \hat{Y} are the means of estimates based on m independent interpenetrating sub samples, we get

$$v(\hat{R}) = \frac{1}{\hat{X}^2} \frac{1}{m(m-1)} \sum_{i=1}^m (\hat{Y}_i - \hat{R}\hat{X}_i)^2 \quad (10.13)$$

As in case of the estimators of bias, these estimators are also not unbiased. It is of interest to note that if in this case the estimator is taken as the mean of the ratio estimates based on m sub samples, that is, if

$$\hat{R} = \frac{1}{m} \sum_{i=1}^m \hat{R}_i, \quad \hat{R}_i = \hat{Y}_i / \hat{X}_i, \quad (10.14)$$

then an unbiased estimator of the exact variance of \hat{R}' is given by

$$v(\hat{R}') = \frac{1}{m(m-1)} \sum_{i=1}^m (\hat{R}_i - \hat{R}')^2. \quad \dots \quad (10.15)$$

The efficiencies of the two estimators \hat{R} and \hat{R}' are compared in Section 10.9.

General Expression for MSE

In (10.11), we have given the expression for $M(\hat{R})$ correct to the first order of approximation. It is possible to get the expression for the mse correct to any order of approximation by considering the expected value of the terms in the expansion of $(e-e')^2(1+e')^{-2}$. Thus if $|e'| < 1$, we have

$$\begin{aligned} M(\hat{R}) &= R^2 E[(e-e')^2 - 2e'(e-e')^2 + 3e'^2(e-e')^2 - \dots] \\ &= R^2 [(\nu_{02} - 2\nu_{11} + \nu_{20}) - 2(\nu_{12} - 2\nu_{21} + \nu_{30}) + 3(\nu_{22} - 2\nu_{31} + \nu_{40}) - \dots], \quad \dots \end{aligned} \quad (10.16)$$

where ν_{ij} is as defined in (10.10). Here also it may be noted that the terms within the brackets are of the form θ_2/n , θ_3/n^2 , θ_4/n^3 , etc., for many sampling schemes and hence for large sample sizes the terms involving higher order moments may be neglected. (cf. Problem 10.10 p. 401).

10.4 RATIO METHOD OF ESTIMATION

If \hat{Y} and \hat{X} are unbiased estimators of the population totals of the study and auxiliary variables respectively based on any given sampling scheme and if X is known, then we have seen in (10.2) that an alternative estimator for Y is $\hat{Y}_R = (\hat{Y}/\hat{X})X$. The bias of \hat{Y}_R is given by X times that of \hat{R} given in (10.5), that is,

$$B(\hat{Y}_R) = \frac{1}{X} [RV(\hat{X}) - \text{Cov}(\hat{X}, \hat{Y})]. \quad \dots \quad (10.17)$$

The variance of \hat{Y}_R is $X^2V(\hat{R})$, namely,

$$V(\hat{Y}_R) = V(\hat{Y}) - 2R \text{Cov}(\hat{X}, \hat{Y}) + R^2V(\hat{X}). \quad \dots \quad (10.18)$$

Here the term variance is used, since the variance and the mse are the same for the order of approximation considered. Comparing (10.18) with $V(\hat{Y})$, we find that the estimator \hat{Y}_R is more efficient than \hat{Y} if

$$2R \text{Cov}(\hat{X}, \hat{Y}) > R^2V(\hat{X}),$$

which, when R is positive, becomes

$$\rho(\hat{X}, \hat{Y}) > \frac{1}{2} \frac{C(\hat{X})}{C(\hat{Y})}, \quad (10.19)$$

where $\rho(\hat{X}, \hat{Y})$ is the correlation coefficient between \hat{X} and \hat{Y} and $C(\hat{X})$ and $C(\hat{Y})$ are the rse's of \hat{X} and \hat{Y} respectively. In practice usually the variability of \hat{X} is likely to be less than that of \hat{Y} since the sample design is mainly based on the information available for the auxiliary variable. Hence the inequality (10.19) is likely to be satisfied especially when the auxiliary variable has been well chosen in which case $\rho(\hat{X}, \hat{Y})$ would not only be positive but also near about unity. The possibility of using an alternative estimator when \hat{X} and \hat{Y} are negatively correlated, is considered in Section 10.8.

When R is negative which can happen when either X or Y is positive and the other is negative the condition (10.19) for the ratio estimator \hat{Y}_R to be more efficient than the conventional estimator \hat{Y} becomes

$$\rho(\hat{X}, \hat{Y}) < -\frac{1}{2} \frac{C(\hat{X})}{C(\hat{Y})} \quad (10.20)$$

It may be mentioned that the question of the choice of the ratio estimator also arises in estimating a population ratio Y/X , when the actual value of X is known for in this case one may either use \hat{Y}/\hat{X} or \hat{Y}/X as the estimator of the population ratio. The former though biased is more efficient than the latter if the condition in (10.19) is satisfied. Thus we see that it may be profitable to use a ratio of two estimators even in the case where the actual value of the denominator is known provided the correlation coefficient between the estimators of the numerator and the denominator satisfy the prescribed condition. If $\{\hat{X}_i, \hat{Y}_i\}$, $i = 1, 2, \dots, m$, are unbiased estimates of X and Y based on m independent interpenetrating subsamples then estimators of the bias and the variance of \hat{Y}_R can be obtained by multiplying (10.7) by \hat{X} and (10.13) by \hat{X}^2 respectively and they are given by

$$b(\hat{Y}_R) = \frac{1}{\hat{X}} \frac{1}{m(m-1)} \sum_{i=1}^m \hat{X}_i (\hat{Y}_i - R \hat{X}_i) \quad (10.21)$$

and

$$v(\hat{Y}_R) = \frac{1}{m(m-1)} \sum_{i=1}^m (\hat{Y}_i - \hat{R}\hat{X}_i)^2, \quad \dots \quad (10.22)$$

where \hat{Y} and \hat{X} are the means of the m sub-sample estimates.

10.5 BASIC SAMPLING SCHEMES

In the last three sections, we have derived the general expressions for the bias and the variance of a ratio estimator for any sample design and the expressions for the bias and the variance in the case of any particular sampling scheme can be obtained by substituting in them the expressions for $V(\hat{X})$, $V(\hat{Y})$ and $\text{Cov}(\hat{X}, \hat{Y})$ for that scheme. In this section the ratio estimator is examined for the basic sampling schemes, namely, simple random sampling, systematic sampling and varying probability sampling.

10.5a SIMPLE RANDOM SAMPLING

Suppose a sample of n units is selected with srswr. Then an estimator of the ratio Y/X is \bar{y}/\bar{x} , and the expressions for the bias and the variance of this ratio estimator can be obtained by noting that

$$V(\hat{X}) = N^2\sigma_x^2/n, \quad V(\hat{Y}) = N^2\sigma_y^2/n \quad \text{and} \quad \text{Cov}(\hat{X}, \hat{Y}) = N^2/\sigma_x\sigma_y/n,$$

where σ_x^2 , σ_y^2 and ρ are respectively the population variances of the variables x and y and the correlation coefficient between them. Substituting these in (10.5) and (10.11), we get,

$$B(\hat{R}) = \frac{1}{\bar{X}^2} \frac{1}{n} \left(R\sigma_x^2 - \rho\sigma_x\sigma_y \right) = \frac{R}{n} \left(C_x^2 - \rho C_x C_y \right), \quad \dots \quad (10.23)$$

and

$$V(\hat{R}) = \frac{1}{\bar{X}^2} \frac{1}{n} \left(\sigma_y^2 - 2R\rho\sigma_x\sigma_y + R^2\sigma_x^2 \right) = \frac{R^2}{n} \left(C_y^2 - 2\rho C_x C_y + C_x^2 \right), \quad \dots \quad (10.24).$$

where C_x and C_y are the coefficients of variation of the variables x and y respectively

From the expressions (10.23) and (10.24), it can be seen that the approximate bias and the variance decrease with increase in n . The fact that the actual bias itself decreases with increase in n can be noted from result (10.8), since in this case $C(\hat{R})$ and $C(\hat{X})$ decrease with increase in n . It may also be observed that the ratio of the square of the bias to the variance in this case is of the form

$$\frac{B^2(\hat{R})}{V(\hat{R})} = \frac{1}{n} \cdot \frac{(C_x^2 - \rho C_x C_y)^2}{(C_x^2 - 2\rho C_x C_y + C_y^2)}, \quad (10.25)$$

which decreases with increase in n and hence for large sample sizes the bias might become negligible compared to the variance.

Comparing the variance of the ratio estimator given in (10.24) with the variance of the unbiased estimator y/\bar{X} , namely $\sigma_y^2/n\bar{X}^2$, we find that the former is more efficient than the latter if

$$\text{and } \left. \begin{array}{l} \rho > +\frac{1}{2} (C_x/C_y) \quad \text{for } R > 0 \\ \rho < -\frac{1}{2} (C_x/C_y) \quad \text{for } R < 0 \end{array} \right\} \quad (10.26)$$

It is to be noted that in sampling with srswr and srs wr, it is not desirable to use the estimator

$$\hat{R} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}, \quad (10.27)$$

since the bias of this estimator

$$B(\hat{R}) = \frac{1}{N} \sum_{i=1}^N \frac{Y_i}{X_i} - R = -\frac{1}{\bar{X}} \text{Cov} \left(\frac{y}{x}, x \right), \quad (10.28)$$

which does not depend on the sample size and hence does not decrease with increase in n unlike the bias of the estimator (y/\bar{x}) .

By expanding and simplifying the variance and the covariance terms in (10.24), we get

$$V(\hat{R}) = \frac{1}{nN\bar{X}^2} \sum_{i=1}^N (Y_i - R\bar{X}_i)^2.$$

Since unbiased estimators of σ_x^2 , σ_y^2 and $\text{Cov}(x, y)$ are given by

$$s_x^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}, \quad s_y^2 = \sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n-1} \quad \text{and} \quad s_{xy} = \sum_{i=1}^n \frac{(y_i - \bar{y})(x_i - \bar{x})}{n-1},$$

we get the estimators of the bias and the variance of \hat{R} as

$$b(\hat{R}) = \frac{1}{\bar{x}^2} \frac{1}{n} (\hat{R} s_x^2 - s_{xy}) = \frac{1}{\bar{x}^2} \sum_{i=1}^n \frac{x_i (\hat{R}x_i - y_i)}{n(n-1)} \quad \dots \quad (10.29)$$

and

$$v(\hat{R}) = \frac{1}{\bar{x}^2} \frac{1}{n} (s_y^2 - 2\hat{R} s_{xy} + \hat{R}^2 s_x^2) = \frac{1}{\bar{x}^2} \sum_{i=1}^n \frac{(y_i - \hat{R}x_i)^2}{n(n-1)}. \quad \dots \quad (10.30)$$

It may be noted that the expressions for the bias and the variance of the ratio estimator of the total \hat{Y}_R can be obtained by multiplying (10.23) and (10.24) by X and X^2 respectively. Similarly, the estimators of the bias and the variance can be obtained by multiplying (10.29) and (10.30) by X and X^2 respectively. Further, the expressions for the bias and the variance for the ratio estimator in the case of sampling with srs wor can be obtained by multiplying (10.23) and (10.24) by the finite population factor $(N-n)/(N-1)$. Similarly, the expressions for the estimators of bias and variance in this case can be got by multiplying (10.29) and (10.30) by the factor $(1-f)$.

10.5b SYSTEMATIC SAMPLING

In the case of systematic sampling, the bias and the variance of the ratio estimators are obtained as before by substituting the expressions for the variances and the covariance of the estimators \bar{x} and \bar{y} in the general expressions given in (10.5) and (10.11). Noting that $V(\bar{x})$,

$V(\bar{y})$ and $\text{Cov}(\bar{x}, \bar{y})$ in the case of systematic sampling are of the form

$$V(\bar{x}) = \frac{1}{k} \sum_{s=1}^k (\bar{x}_s - \bar{X})^2, \quad V(y) = \frac{1}{k} \sum_{s=1}^k (y_s - \bar{Y})^2$$

and

$$\text{Cov}(\bar{x}, \bar{y}) = \frac{1}{k} \sum_{s=1}^k (\bar{x}_s - \bar{X})(\bar{y}_s - \bar{Y}),$$

where \bar{x}_s and \bar{y}_s denote the sample means of x and y for the s th possible sample and k is the sampling interval such that $nl = N$, we have

$$B(\hat{R}) = \frac{1}{\bar{X}^2} \cdot \frac{1}{k} \sum_{s=1}^k \bar{x}_s (R\bar{x}_s - \bar{y}_s) \quad . \quad (10.31)$$

and

$$V(\hat{R}) = \frac{1}{\bar{X}^2} \cdot \frac{1}{k} \sum_{s=1}^k (\bar{y}_s - R\bar{x}_s)^2 \quad (10.32)$$

From (10.32) we see that for ratio estimation to be efficient, it is necessary to arrange the units such that the variation between the sample ratios \bar{y}_s/\bar{x}_s in the minimum

It is of interest to note that the variance in (10.32) can be expressed in terms of the intraclass correlation as in the case of the usual estimator of population total or mean (Swain, 1964). For, we have seen in Section 5.6 of Chapter 5, (p 146), that

$$V(\bar{y}) = \frac{\sigma_y^2}{n} \{1 + (n-1)\rho_{cy}\},$$

where ρ_{cy} is the intraclass correlation for y and substituting this and similar expressions for $V(\bar{x})$ and $\text{Cov}(\bar{x}, \bar{y})$ in (10.11), we get

$$V(\hat{R}) = \frac{1}{\bar{X}^2} \cdot \frac{1}{n} \left[\sigma_y^2 \{1 + (n-1)\rho_{cy}\} + R^2 \sigma_x^2 \{1 + (n-1)\rho_{cx}\} - 2R\rho\sigma_x\sigma_y \sqrt{\{1 + (n-1)\rho_{cx}\}\{1 + (n-1)\rho_{cy}\}} \right], \quad (10.33)$$

which, when $\rho_{cx} = \rho_{cy} = \rho_c$, becomes

$$\begin{aligned} V(\hat{R}) &= \frac{1}{\bar{X}^2} \frac{1}{n} \{ \sigma_y^2 - 2R\rho\sigma_x\sigma_y + R^2\sigma_x^2 \} \{ 1 + (n-1)\rho_c \} \\ &= \frac{R^2}{n} \{ C_x^2 - 2\rho C_x C_y + C_y^2 \} \{ 1 + (n-1)\rho_c \}, \end{aligned}$$

where ρ is the correlation coefficient between x and y . Comparing this expression with (10.24), we find that

$$V(\hat{R}_{sys}) = V(\hat{R}_{srs}) \{ 1 + (n-1)\rho_c \}. \quad \dots \quad (10.34)$$

This shows that even in the case of ratio estimation systematic sampling can be made efficient by arranging the units such that the intra-class correlation coefficient ρ_c becomes negative.

10.5c VARYING PROBABILITY SAMPLING

Noting that in sampling with ppswr the expressions for $V(\hat{X})$, $V(\hat{Y})$ and $\text{Cov}(\hat{X}, \hat{Y})$ are given by

$$V(\hat{X}) = \frac{1}{n} \left(\sum_{i=1}^N \frac{X_i^2}{P_i} - \bar{X}^2 \right), \quad V(\hat{Y}) = \frac{1}{n} \left(\sum_{i=1}^N \frac{Y_i^2}{P_i} - \bar{Y}^2 \right),$$

$$\text{Cov}(\hat{X}, \hat{Y}) = \frac{1}{n} \left(\sum_{i=1}^N \frac{X_i Y_i}{P_i} - \bar{X} \bar{Y} \right)$$

and substituting these expressions in (10.5) and (10.11), we get

$$B(\hat{R}) = \frac{1}{n} \frac{1}{\bar{X}^2} \sum_{i=1}^N \frac{X_i}{P_i} (R X_i - Y_i) \quad \dots \quad (10.35)$$

and

$$V(\hat{R}) = \frac{1}{n} \frac{1}{\bar{X}^2} \sum_{i=1}^N \frac{1}{P_i} (Y_i - R X_i)^2. \quad \dots \quad (10.36)$$

It may be mentioned that the ratio estimator would be efficient if the set of probabilities used for selection is appropriate for both the numerator and the denominator variables and if y/p and x/p are positively related.

10.6 STRATIFIED SAMPLING

Suppose \hat{Y}_s and \hat{X}_s are unbiased estimators of the s th stratum totals of the variables y and x , ($s = 1, 2, \dots, K$), based on any sample design. Then an estimator of the ratio ($R = Y/X$) is given by

$$\hat{R} = \frac{\hat{Y}}{\hat{X}}, \quad (\hat{Y} = \sum_{s=1}^K \hat{Y}_s, \quad \hat{X} = \sum_{s=1}^K \hat{X}_s) \quad (10.37)$$

If the value of X is available, a ratio estimator of the population total Y can be obtained by multiplying \hat{R} given in (10.37) by X , that is,

$$\hat{Y}_R = \left\{ \sum_{s=1}^K \hat{Y}_s / \sum_{s=1}^K \hat{X}_s \right\} X \quad (10.38)$$

Since

$$V(\hat{Y}) = \sum_{s=1}^K V(\hat{Y}_s), \quad V(\hat{X}) = \sum_{s=1}^K V(\hat{X}_s), \quad \text{Cov}(\hat{X}, \hat{Y}) = \sum_{s=1}^K \text{Cov}(\hat{X}_s, \hat{Y}_s),$$

$$B(\hat{Y}_R) = \frac{1}{X} \sum_{s=1}^K \{RV(\hat{X}_s) - \text{Cov}(\hat{X}_s, \hat{Y}_s)\} \quad (10.39)$$

and

$$V(\hat{Y}_R) = \sum_{s=1}^K \{V(\hat{Y}_s) - 2R \text{Cov}(\hat{X}_s, \hat{Y}_s) + R^2 V(\hat{X}_s)\} \quad (10.40)$$

If, in addition to the value of X , the values of X_s are also known, then an alternative ratio estimator of Y can be taken as

$$\hat{Y}'_R = \sum_{s=1}^K \hat{R}_s X_s = \sum_{s=1}^K \frac{\hat{Y}_s}{\hat{X}_s} X_s \quad (10.41)$$

It may be seen that this estimator, known as *separate ratio estimator*, is built up from stratum level ratio estimators, whereas the estimator \hat{Y}_R is based on the pooled or combined estimators \hat{Y} and \hat{X} . The

estimator \hat{Y}_R is termed *combined ratio estimator*. The bias and the variance of the estimator \hat{Y}'_R , given in (10.41), are given by

$$B(\hat{Y}'_R) = \sum_{s=1}^K \frac{1}{X_s} \{R_s V(\hat{X}_s) - \text{Cov}(\hat{X}_s, \hat{Y}_s)\} \quad \dots \quad (10.42)$$

and

$$V(\hat{Y}'_R) = \sum_{s=1}^K \{V(\hat{Y}_s) - 2R_s \text{Cov}(\hat{X}_s, \hat{Y}_s) + R_s^2 V(\hat{X}_s)\}. \dots \quad (10.43)$$

Comparing (10.39) and (10.42), we find that the bias of \hat{Y}_R is likely to be much less than that of \hat{Y}'_R , since $\{X_s\}$ would be considerably smaller than X , whereas the ratios $\{R_s\}$ may not be much different from R . Comparing (10.40) and (10.43), we note that the variances of the two estimators are not likely to differ substantially, provided the stratum ratios $\{R_s\}$ are of the same order. This shows that if the stratum ratios do not vary much among themselves, it is desirable to use the combined ratio estimator \hat{Y}_R instead of the separate ratio estimator \hat{Y}'_R obtained at stratum level. However, if R_s 's vary much among themselves, a closer examination is required to choose between the two estimators and it may happen that \hat{Y}'_R is more efficient than \hat{Y}_R in the sense of its having lesser mse, since in this case the variance of \hat{Y}'_R may be expected to be less than that of \hat{Y}_R .

Further, it is to be noted that for obtaining the expressions for the bias and the variance of \hat{Y}_R , it is sufficient to assume that $\left| \frac{\hat{X} - X}{X} \right| < 1$, whereas in the case of \hat{Y}'_R , it is necessary to assume that in each stratum $\left| \frac{\hat{X}_s - X_s}{X_s} \right| < 1$. Thus we see that while we can expect the expressions of the bias and the variance derived for \hat{Y}_R to be valid if the total sample size is large enough, the corresponding expressions derived for \hat{Y}'_R can be expected to be valid only when the sample size in each stratum is sufficiently large.

The variances of \hat{Y}_R and \hat{Y}'_R given in (10.40) and (10.43) can be estimated by substituting in them the estimators of $V(\hat{X}_s)$, $V(\hat{Y}_s)$ and $\text{Cov}(\hat{X}_s, \hat{Y}_s)$. For instance, if the sample in each stratum is

selected in the form of m independent interpenetrating sub samples, then estimators of $V(\hat{Y}_R)$ and $V(\hat{Y}'_R)$ are given by

$$v(\hat{Y}_R) = \frac{1}{m(m-1)} \sum_{s=1}^K \sum_{i=1}^m (\hat{Y}_{si} - \hat{R}\hat{X}_{si})^2 - m(\hat{Y}_s - \hat{R}\hat{X}_s)^2 \quad (10.44)$$

and

$$v(\hat{Y}'_R) = \frac{1}{m(m-1)} \sum_{s=1}^K \sum_{i=1}^m (\hat{Y}_{si} - \hat{R}_s\hat{X}_{si})^2 \quad (10.45)$$

By pooling the strata estimates for each sub sample we get m independent sub sample estimates $\{\hat{X}_i\}$ and $\{\hat{Y}_i\}$ for X and Y and in that case an estimator of the variance of \hat{Y}_R is given by (10.22). The variance estimator given in (10.44) is not as simple to calculate as (10.22) though it is expected to be more efficient than the latter.

At this stage, it may be mentioned that all the estimators of bias and the variance so far considered are generally biased but consistent in the sense that when the sample size is increased, the estimators tend to the population values.

10.7 AN EMPIRICAL STUDY

To study the behaviour of the bias and the sampling variance of the ratio estimator and of the closeness of their approximations to the actual values for different sample sizes the 1951 census data on cultivated area and geographical area for the villages in 10 tehsils of Madras State have been used. For the sake of convenience in selecting systematic samples of size 2, 4, 8, 16 and 32, only 128 of the villages in each tehsil have been considered for this study. For each sample size, all the systematic samples are first formed and the estimates of the ratio of cultivated area (y) to geographical area (x) are worked out. Using these estimates, the actual bias and the variance are calculated. Using the variances and covariance of \hat{Y} and \hat{X} , the values of the approximate expressions for the bias and the variance of the ratio estimator are determined, and these are compared with the corresponding actual

values. The results of this study are presented in Table 10.1. From this, it can be seen that the bias and the variance generally decrease with increase in sample size and that though there is some amount of agreement between the approximate and the actual values in some cases, the differences between them are generally more for smaller sample sizes than for larger sample sizes, which shows that the approximations become increasingly valid with increase in sample size.

TABLE 10.1. APPROXIMATE AND ACTUAL VALUES OF BIAS AND VARIANCE FOR A RATIO ESTIMATOR IN SYSTEMATIC SAMPLING.

sample size	relative bias		relative variance		sample size	relative bias		relative variance	
	approx.	actual	approx.	actual		approx.	actual	approx.	actual
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
tehsil	1.		$R = 0.1754$		tehsil	6.		$R = 0.5174$	
2	.314	.215	.675	.445	2	.053	.025	.158	.090
4	.134	.102	.218	.208	4	.025	.026	.057	.038
8	.047	.032	.066	.053	8	.002	.008	.017	.010
16	.025	.023	.034	.036	16	.002	—	.003	.004
32	—.002	—.004	.005	.006	32	.002	.001	.001	.001
tehsil	2.		$R = 0.6094$		tehsil	7.		$R = 0.5458$	
2	.133	.195	.167	.169	2	.117	.071	.152	.132
4	.074	.080	.096	.087	3	.039	.029	.056	.055
8	.035	.028	.043	.040	8	.015	.014	.015	.018
16	.032	.032	.024	.022	16	.006	.007	.006	.005
32	.005	.004	.005	.005	32	.032	.033	.001	.001
tehsil	3.		$R = 0.4770$		tehsil	8.		$R = 0.6762$	
2	.126	.545	.129	.369	2	.051	.048	.133	.074
4	.069	.271	.063	.129	4	.061	.031	.012	.058
8	.066	.208	.044	.088	8	.014	.005	.008	.011
16	.039	.123	.022	.040	16	.004	—	.004	.005
32	.011	.029	.008	.011	32	—.035	.002	.001	.001
tehsil	4.		$R = 0.6256$		tehsil	9.		$R = 0.4704$	
2	.174	.790	.136	.681	2	.185	.206	.234	.242
4	.145	.551	.079	.509	4	.091	.136	.100	.172
8	.115	.284	.067	.276	8	.057	.065	.130	.068
16	.099	.169	.067	.182	16	.022	.022	.047	.018
32	.058	.067	.051	.074	32	.008	.007	.010	.009
tehsil	5.		$R = 0.3969$		tehsil	10.		$R = 0.2476$	
2	.044	.030	.301	.106	2	.129	.436	.131	.509
4	.050	.056	.319	.195	4	.093	.198	.098	.230
8	.040	.011	.092	.095	2	.065	.089	.066	.082
16	.018	.033	.077	.075	16	.044	.044	.033	.033
32	.001	.001	.018	.020	32	.044	.040	.016	.016

(R : proportion of cultivated area).

10.8 PRODUCT METHOD OF ESTIMATION

In Section 10.4, we have seen that the ratio method of estimation provides a more efficient estimator $\hat{Y}_R = (\hat{Y}/\hat{X})X$ of the population total Y than the estimator \hat{Y} for any sample design, provided the correlation coefficient between \hat{Y} and \hat{X} is greater than $C(\hat{X})/2C(\hat{Y})$, when $R (= Y/X)$ is positive. This shows that if we happen to have an auxiliary variable x such that $\rho(\hat{X}, \hat{Y})$ is negative, we cannot make use of the ratio method of estimation to improve upon the conventional estimator \hat{Y} . In such a situation, we may consider another method of estimation, which is expected to be more efficient than \hat{Y} in situations where $(\hat{Y}/\hat{X})X$ turns out to be less efficient than \hat{Y} and hence which can be taken as complementary to the ratio method of estimation. This method of estimation termed the *product method of estimation* consists in considering $(\hat{Y}\hat{X})/X$ as an estimator of the population total Y , the estimator itself being termed the *product estimator*.

Writing $\hat{Y} = Y(1+\epsilon)$ and $\hat{X} = X(1+\epsilon')$ and substituting these in the product estimator

$$\hat{Y}_P = \hat{Y} \hat{X}/X, \quad (10.46)$$

where the subscript P stands for the product estimator, we get

$$\hat{Y}_P = Y(1+\epsilon)(1+\epsilon') = Y\{1+(\epsilon+\epsilon')+ee'\} \quad \dots \quad (10.47)$$

Taking the expected value of \hat{Y}_P , we get

$$E(\hat{Y}_P) = Y\{1 + \text{Cov}(\hat{X}, \hat{Y})/XY\}, \quad \dots \quad (10.48)$$

which shows that the exact bias in the product estimator is

$$B(\hat{Y}_P) = \frac{1}{X} \text{Cov}(\hat{X}, \hat{Y}) = \rho(\hat{X}, \hat{Y})C(\hat{X})C(\hat{Y})Y. \quad (10.49)$$

For many commonly used sampling schemes the covariance of \hat{X} and \hat{Y} is of the form θ/n , where n is the sample size and θ is a population parameter independent of n and hence the bias of a product estimator, in general, decreases with increase in sample size. It may be noted that if the term of degree 2 in (ϵ, ϵ') in $E(\hat{Y}_P)$ is neglected, then the bias of the product estimator becomes zero.

The exact mse of the product estimator can be obtained by noting that

That is,

$$M(\hat{Y}_P) = E(\hat{Y}_P - Y)^2 = Y^2 E((\epsilon+\epsilon')+ee')^2$$

$$M(\hat{Y}_P) = Y^2(v_{00}+2v_{11}+v_{20})+2(v_{12}+v_{21})+v_{22}, \quad (10.50)$$

where v_{ij} is as defined in (10.10). Since v_{ij} is likely to be of the form θ_{ij}/n^{i+j-1} for many sampling schemes, the terms involving moments of order greater than 2 may

be neglected if the sample size is at least moderately large and thus we have the mse to a first order of approximation

$$\begin{aligned} M(\hat{Y}_P) &= Y^2 \left\{ \frac{V(\hat{Y})}{Y^2} + 2 \frac{\text{Cov}(\hat{X}, \hat{Y})}{XY} + \frac{V(\hat{Y})}{X^2} \right\} \\ &= V(\hat{Y}) + 2R \text{Cov}(\hat{X}, \hat{Y}) + R^2 V(\hat{X}). \end{aligned} \quad \dots \quad (10.51)$$

To this order of approximation, the estimator \hat{Y}_P is unbiased for Y and hence the mse and the variance are the same. Comparing (10.51) with $V(\hat{Y})$, we find that the former is less than the latter if

$$\rho(\hat{X}, \hat{Y}) < -\frac{1}{2} \frac{C(\hat{X})}{C(\hat{Y})}, \quad \dots \quad (10.52)$$

when $P (= YX)$ is positive. The condition becomes $\rho > C(\hat{X})/2C(\hat{Y})$, when P is negative. Thus we see that, given any supplementary variable, we may decide whether to use this variable for obtaining a ratio estimator or a product estimator according as $\rho(\hat{X}, \hat{Y})$ is greater than $C(\hat{X})/2C(\hat{Y})$ or less than $-C(\hat{X})/2C(\hat{Y})$, when P is positive and vice-versa when P is negative. If $\rho(\hat{X}, \hat{Y})$ is between these two values, then it is better to use the conventional estimator \hat{Y} , since in this case it is more efficient than the ratio and the product estimators.

Estimators of bias and variance of the product estimator can be obtained by substituting the estimators of R , $V(\hat{X})$, $V(\hat{Y})$ and $\text{Cov}(\hat{X}, \hat{Y})$ in (10.49) and (10.51). The expressions for the bias and the variance of $\hat{X}\hat{Y}$ can be obtained by multiplying the corresponding expressions for \hat{Y}_P by X and X^2 respectively. L.A. Goodman (1960) has considered the question of obtaining the variance and variance estimators of product of estimators. Murthy (1964) has discussed the uses of the product method of estimation and has given a technique of obtaining unbiased product estimators based on interpenetrating sub-samples.

10.9 ALMOST UNBIASED RATIO ESTIMATORS

We shall now consider a technique of estimating the bias of a ratio estimator unbiasedly to the second order of approximation based on interpenetrating sub-samples. This estimator of the bias may be used to correct the ratio estimator for its bias, thereby getting a ratio estimator, which is unbiased upto the second order of approximation. Suppose the sample is drawn in the form of m independent interpenetrating sub-samples of the same size and selected according to the same sample design and let $\{\hat{Y}_i\}$ and $\{\hat{X}_i\}$, ($i = 1, 2, \dots, m$), be unbiased estimates of the population totals Y and X based on

the m sub samples. In this case, the following two ratio estimators can be considered for $R (= Y/X)$

$$\hat{R}_1 = \frac{\hat{Y}}{\hat{X}}, \quad \left(\hat{Y} = \frac{1}{m} \sum_{t=1}^m \hat{Y}_t, \quad \hat{X} = \frac{1}{m} \sum_{t=1}^m \hat{X}_t \right), \quad (10.53)$$

and

$$\hat{R}_m = \frac{1}{m} \sum_{t=1}^m \frac{\hat{Y}_t}{\hat{X}_t} = \frac{1}{m} \sum_{t=1}^m r_t, \quad \left(r_t = \frac{\hat{Y}_t}{\hat{X}_t} \right). \quad (10.54)$$

It may be noted that \hat{R}_1 is based on a single ratio, whereas \hat{R}_m is the mean of m ratios.

If the sample size in each of the sub samples is large enough for the two assumptions mentioned in Section 10.2 to be satisfied, the formula for the bias given in (10.5) may be applied to the estimators given in (10.53) and (10.54) and we get

$$\begin{aligned} B_1 &= B(\hat{R}_1) = \frac{1}{\bar{X}^2} [RV(\hat{X}) - \text{Cov}(\hat{X}, \hat{Y})] \\ &= \frac{1}{m^2} \frac{1}{\bar{X}^2} \sum_{t=1}^m \{RV(\hat{X}_t) - \text{Cov}(\hat{X}_t, \hat{Y}_t)\} = \frac{1}{m^2} \sum_{t=1}^m B(r_t), \end{aligned} \quad (10.55)$$

and

$$B_m = B(\hat{R}_m) = \frac{1}{m} \sum_{t=1}^m B(r_t). \quad (10.56)$$

Comparing (10.55) and (10.56), we get

$$mB_1 = B_m, \quad (10.57)$$

which shows that the bias of the estimator \hat{R}_m is m times the bias of the estimator \hat{R}_1 . Incidentally, it may be seen that their variances correct to the first order of approximation are given by

$$\begin{aligned} V_1 &= V(\hat{R}_1) = \frac{1}{m^2} \frac{1}{\bar{X}^2} \sum_{t=1}^m \{V(\hat{Y}_t) - 2R \text{Cov}(\hat{X}_t, \hat{Y}_t) + R^2 V(\hat{X}_t)\} \\ &= \frac{1}{m^2} \sum_{t=1}^m V(r_t) = V(\hat{R}_m) = V_m \end{aligned}, \quad (10.58)$$

This shows that the variances of \hat{R}_1 and \hat{R}_m correct to the first order of approximation are the same and that in view of this \hat{R}_1 is to be preferred to \hat{R}_m , since the former has less bias than the latter.

From (10.57), we get

$$E(\hat{R}_m - \hat{R}_1) = B_m - B_1 = (m-1)B_1, \quad \dots \quad (10.59)$$

since $B_m = mB_1$. Hence, an unbiased estimator of B_1 is given by

$$\hat{B}_1 := \frac{1}{m-1} (\hat{R}_m - \hat{R}_1). \quad \dots \quad (10.60)$$

Once the approximate bias is estimated unbiasedly, the ratio estimator \hat{R}_1 can be corrected for its bias. The corrected estimator \hat{R}_c is

$$\hat{R}_c = \hat{R}_1 - \hat{B}_1 = \frac{1}{m-1} (m\hat{R}_1 - \hat{R}_m). \quad \dots \quad (10.61)$$

This corrected estimator may be termed *almost unbiased ratio estimator*, since it is only unbiased upto the second order of approximation.

The general technique of obtaining almost unbiased ratio estimators for any order of approximation has been considered by Murthy and Nanjamma (1959) and this problem has been examined on similar lines for srs by Quenouille (1956), (cf. Problem 10.17, p.403), and Durbin (1959) has shown that for certain populations the corrected ratio estimator is more efficient than the biased ratio estimator (cf. Problem 10.18, p. 403).

10.10 UNBIASED RATIO TYPE ESTIMATORS

In this section, we describe briefly a procedure of making the ratio estimator of the population total completely unbiased. Suppose \hat{Y}_i and \hat{X}_i , ($i = 1, 2, \dots, m$), are unbiased estimators of Y and X based on m independent interpenetrating sub-samples. Then it can be easily seen that the exact bias of the ratio estimator

$$\hat{Y}_{Rm} = \frac{1}{m} \sum_{i=1}^m r_i X = \hat{R}_m X, \quad \dots \quad (10.62)$$

where $r_t = \hat{Y}_t/\hat{X}_t$, is given by

$$B(\hat{Y}_{Rm}) = B(r_t X) = -\text{Cov}(r_t, \hat{X}_t) \quad \dots \quad (10.63)$$

An unbiased estimator of (10.63) is provided by

$$b(\hat{Y}_{Rm}) = -\sum_{i=1}^m \frac{(r_i - \hat{R}_m)(\hat{X}_i - \hat{X})}{m-1} = -\frac{(\hat{Y} - \hat{R}_m \hat{X})m}{m-1}, \quad \dots \quad (10.64)$$

where $\hat{Y} = \frac{1}{m} \sum_{i=1}^m \hat{Y}_i$ and $\hat{X} = \frac{1}{m} \sum_{i=1}^m \hat{X}_i$. By subtracting this estimator of bias from \hat{Y}_{Rm} , we get a completely unbiased estimator

$$\hat{Y}_{UR} = \hat{R}_m \hat{X} + \frac{m}{m-1} (\hat{Y} - \hat{R}_m \hat{X}) \quad \dots \quad (10.65)$$

Since this estimator is not strictly a ratio estimator in the sense of involving only ratio of estimators, this may be termed a *ratio-type estimator*.

It may be mentioned that this method of getting an unbiased ratio estimator is applicable to the case of estimating a population ratio only if the value of the denominator is known, whereas the estimator discussed in Section 10.9 is applicable even when X is not known in such a case.

The unbiased ratio type estimator was proposed by Hartley and Ross (1954) in the case of srs. Suppose a sample of n units is selected with srs wr, then an unbiased ratio type estimator for estimating the population total \bar{Y} is given by

$$\hat{Y}_{UR} = \bar{r} \bar{X} + \frac{n(N-1)}{n-1} (\bar{y} - \bar{r} \bar{x}), \quad \bar{r} = \frac{1}{n} \sum_{i=1}^n r_i, \quad (10.66)$$

where $r_i = y_i/x_i$. It may be noted that if the sample had been selected with srswr, the term $(N-1)$ in (10.66) is to be replaced by N . As usual the estimators for \bar{Y} can be obtained by dividing the estimators for \bar{Y} by N . Robson (1957) and L A Goodman and Hartley (1958) have considered the variance of the ratio type estimator given in (10.66). Mickey (1959) and Williams (1961) have given generalized ratio type estimators, of which the estimator (10.66) is a particular case.

L. A. Goodman and Hartley (1958) have shown that this estimator is more efficient than the usual biased ratio estimator, namely, $(\hat{Y}/\hat{X})X$, if and only if the slope of the regression line of y on x is closer to $\frac{1}{N} \sum_{i=1}^N (Y_i/X_i)$ than to (Y/X) . Nieto de Pascual (1961) has suggested the correction of the estimator $\hat{Y}_{R_1} = (\hat{Y}/\hat{X})X$ for its bias using the unbiased estimator of bias obtained for \hat{Y}_{Rm} and making use of the result derived earlier (10.57). Thus we get

$$\hat{Y}'_{UR} = \hat{R}_1 X + \frac{1}{m-1} (\hat{Y} - \hat{R}_m \hat{X}) \quad \dots \quad (10.67)$$

for any given sample design, where the sample is selected in the form independent interpenetrating sub-samples, and the estimator reduces to

$$\hat{Y}'_{UR} = \frac{\bar{y}}{\bar{x}} X + \frac{N-1}{n-1} (\bar{y} - \bar{r}\bar{x}), \quad \dots \quad (10.68)$$

when a sample of n -units is selected with srs wor. T. J. Rao (1966) has given combinations of \hat{R}_1 and \hat{R}_m which give rise to the estimators (10.66) and (10.68), (cf. Problem 10.21, p.403).

An Empirical Example

The mean square errors of the different estimators in sampling two units with srs wor from a hypothetical population of four units having the values of (y, x) as $(2, 2)$, $(6, 2)$, $(6, 4)$ and $(10, 6)$ considered by Pascual (1961) are given in Table 10.2 to illustrate their relative efficiencies. From this table, it is of interest to note that

TABLE 10.2. EFFICIENCIES OF DIFFERENT TYPES OF RATIO ESTIMATORS

sr. no.	estimator	relative mse	efficiency(%)
(1)	(2)	(3)	(4)
1.	\hat{Y}	0.074	15
2.	\hat{Y}_{R_1}	0.025	44
3.	\hat{Y}_{Rm}	0.067	16
4.	\hat{Y}_{UR}	0.016	69
5.	\hat{Y}'_{UR}	0.012	92
6.	\hat{Y}_e	0.011	100

the almost unbiased ratio estimator obtained by correcting \hat{Y}_{P_1} for its bias turns out to be the most efficient. However it is necessary to conduct large scale model sampling studies for making a reliable comparison of the efficiencies of the different estimators discussed in this section. Pascual has also considered the application of the unbiased ratio type estimators to the case of stratified sampling.

10.11 UNBIASED RATIO ESTIMATORS

In this section a slight modification of the commonly used selection procedures has been given, which makes the ratio estimator unbiased. The ratio estimator $\hat{R} (= \hat{Y}/\hat{X})$ will be unbiased for the ratio $R(-Y/X)$ provided the design is changed such that P_s , the probability of selecting the s th sample, becomes proportional to $\hat{X}_s P'_s$, where P'_s is the probability of selecting the s th sample in the original sample design, that is, if

$$P_s = \frac{\hat{X}_s P'_s}{\sum_s \hat{X}_s P'_s} \quad (10.69)$$

For,

$$E(\hat{R}) = \sum_s \frac{\hat{Y}_s}{\hat{X}_s} P_s = \frac{\sum_s \hat{Y}_s P'_s}{\sum_s \hat{X}_s P'_s} = \frac{Y}{X} = R,$$

since \hat{Y} and \hat{X} are unbiased for Y and X respectively in the original design.

If P_s is the same for all samples, then P_s should be made proportional to \hat{X}_s to make the ratio estimator unbiased. This technique of changing the selection procedure for obtaining unbiased ratio estimators has been considered in a generalized form by Nanjamma, Murthy and Sethi (1959), who have also given the actual modification needed in commonly used sampling schemes. For many of the procedures usually adopted in practice, the modification essentially consists in first selecting one unit with probability proportional to its value of the characteristic occurring in the denominator of the ratio and then selecting the remaining units according to the original scheme of sampling.

10.11a SRS WITHOUT REPLACEMENT

In srs wor, the ratio of the sample means, namely (\bar{y}/\bar{x}) , would be unbiased for R if the original sampling design is modified so as to make the probability of selecting the s -th sample proportional to \bar{x}_s , the sample mean for that sample. For, in this case $P'_s = 1/\binom{N}{n}$ for all s and hence

$$P_s = \frac{\hat{X}_s P'_s}{\bar{X}} = \frac{1}{\binom{N}{n}} \cdot \frac{\bar{x}_s}{\bar{X}}.$$

In other words, the ratio estimator (\bar{y}/\bar{x}) would become an unbiased estimator of the population ratio if the probability of selecting the sample is made proportional to its mean or its total size, (Lahiri, 1951). This can be achieved by selecting one unit with probability proportional to x (ppx) and the rest with srs wor from the remaining units of the population (Midzuno, 1952; Sen, 1952). An unbiased variance estimator of the unbiased ratio estimator based on a sample selected by this method is given by

$$v(\hat{R}) = \hat{R}^2 - \frac{1}{Nn \bar{x} \bar{X}} \left\{ \sum_{i=1}^n y_i^2 + \frac{N-1}{n-1} \sum_{i=1}^n \sum_{i' \neq i} y_i y_{i'} \right\}. \quad \dots \quad (10.70)$$

If \bar{X} is not known, the \bar{X} in (10.70) has to be replaced by its estimator \bar{x} , in which case the above variance estimator ceases to be unbiased.

10.11b SYSTEMATIC SAMPLING

In sampling with equal probability systematically, an unbiased ratio estimator can be obtained by first selecting one unit with ppx and then selecting the other $(n-1)$ units circular systematically with the unit selected first as the random start and with a suitable sampling interval. For this selection procedure, the estimator (\bar{y}/\bar{x}) is unbiased for R . The variance of this estimator in this case would be different from those of the estimators based on srs wr and srs wor. Since the selection is done systematically, it is not possible to estimate the variance unbiasedly on the basis of one sample.

10.11c STRATIFIED SAMPLING

The method of changing the selection procedure for providing unbiased ratio estimators can be applied to other sampling designs such as pps sampling, stratified sampling, etc. Suppose the population is divided into K strata and let N_s and n_s be the number of units in the population and in the sample respectively for the s -th stratum. For stratified srs wor, the modification of the selection procedure consists in selecting one unit (say the i -th unit in the s -th stratum) from the whole population

with ppx, $(n_s - 1)$ units from the remaining $(N_s - 1)$ units in the s th stratum and $n_{s'}$ units from $N_{s'}$ units of the s' th stratum ($s' \neq s$) with srs w/o r. With this procedure, an unbiased estimator of the ratio R is given by

$$\hat{R} = \frac{\sum_{s=1}^K N_s \bar{y}_s}{\sum_{s=1}^K N_s \bar{x}_s}, \quad (10.71)$$

where \bar{y}_s and \bar{x}_s are the sample means in the s th stratum. Des Raj (1954) has shown that an unbiased estimator of the variance of \hat{R} is given by

$$\begin{aligned} \text{v}(\hat{R}) = \hat{R}^2 - \frac{1}{X \sum_{s=1}^K N_s \bar{x}_s} & \left[\sum_{s=1}^K \frac{N_s}{n_s} \sum_{i=1}^{n_s} y_{si}^2 + \sum_{s=1}^K \frac{N_s(N_s-1)}{n_s(n_s-1)} \sum_{i=1}^{n_s} \sum_{i' \neq i} y_{si} y_{s'i} \right. \\ & \left. + \sum_{s=1}^K \sum_{s' \neq s} \frac{N_s N_{s'}}{n_s n_{s'}} \sum_{i=1}^{n_s} \sum_{i' \neq i} y_{si} y_{s'i} \right]. \end{aligned} \quad \dots (10.72)$$

For the original selection procedures corresponding to those considered here, it may be expected that the bias of the conventional ratio estimator is likely to be small in large samples, since the forms of the biased and unbiased estimators are the same and the sample based on the original sampling scheme and that on the modified scheme could be made the same but for a difference of one unit at the most.

10.12 DIFFERENT TYPES OF RATIO ESTIMATORS

So far we have discussed only ratio estimators of the form $(\hat{Y}/\hat{X})X$ in estimating Y except in the case of stratified sampling where we considered $\sum_{s=1}^K (\hat{Y}_s/\hat{X}_s)X$ as an alternative estimator. In this section some other types of ratio estimators are given. As mentioned before, the efficiencies of ratio estimators can be studied only through extensive model sampling experiments, since the mathematical expressions for the bias and the variance are usually derived with certain assumptions and approximations and further, these expressions themselves are quite complicated. Since the expressions for the variances and the conventional variance estimators in cases of various estimators considered in this section are likely to be complicated, it is desirable to select the sample in the form of two or more independent interpenetrating sub-samples for providing a simple variance estimator.

10.12a POST-STRATIFIED RATIO ESTIMATOR

Post-stratified ratio estimator may be taken as an estimator built up from group level ratio estimators. We have already considered such estimators in the case of stratified sampling in Section 10.6. This estimator can also be used if there is reason to believe that the estimators \hat{X} and \hat{Y} are highly correlated for certain parts of the population. Suppose the population can be classified in k divisions or *post-strata*, within each of which the efficiency of a ratio estimator is high due to the higher correlation between the estimators of the numerator and the denominator and let (y_1, y_2, \dots, y_k) and (x_1, x_2, \dots, x_k) be the estimators of the total for y and x in these k post-strata. In this case the post-stratified ratio estimator is given by

$$\hat{Y}_R = \sum_{s=1}^k \frac{y_s}{x_s} X_s \quad \dots \quad (10.73)$$

provided the values of $\{X_s\}$ are known beforehand. Such an estimator has also been termed *component-wise ratio estimator*. This type of ratio estimator has been discussed by Olkin (1958) and Robson and Vithaysai (1961).

This estimator is likely to be more efficient than the *combined ratio estimator* $\left(\sum_{s=1}^k y_s / \sum_{s=1}^k x_s \right) \hat{X}$, when the sample size available in each of the post-strata is not small and when the division has been so done that y_s and x_s are more linearly related than \hat{Y} and \hat{X} and the lines of regression of y on x in the different post-strata pass through the origin. For instance, in a crop yield survey if the geographical area under different soil-types is known for the domain of study, then estimates of production of a specified crop may be obtained by types of soil and a component-wise ratio estimator may be built up. Similarly, in the case of a demographic survey, if we have data on the number of persons by age and sex, then the data relating to marital status, employment, etc., can be tabulated for these categories and a component-wise ratio estimator can be used.

10.12b CHAIN RATIO ESTIMATOR

Another type of ratio estimator, termed *chain ratio estimator* which is commonly used in crop surveys, consists in obtaining the overall ratio estimator by considering ratio estimators based on different stages of survey. Suppose a two stage design is adopted for estimating the acreage under a particular crop in a region with villages as the first stage units and plots as the second stage units and let a_{ij} and g_{ij} be the crop acreage and the geographical area for the j th sample plot in the i th sample village. The chain ratio estimator in this case is

$$\hat{A} = G \left\{ \frac{\sum_{i=1}^n g_i r_i}{\sum_{i=1}^n g_i} \right\}, \quad r_i = \frac{\sum_{j=1}^{m_i} a_{ij}}{\sum_{j=1}^{m_i} g_{ij}} \quad (10.74)$$

where g_i is the geographical area of the i th sample village and G is the total geographical area (cf Sub section 15.20 of Chapter 15). The expressions for the bias and the variance of the estimator in (10.74) are rather complicated. It may, however, be mentioned that the bias of this estimator may not necessarily be small even when the total sample size is large due to the fact that the ratio estimators built up at lower levels such as village are based on rather inadequate sample sizes and hence this estimator is to be used with caution.

10.12c DOUBLE RATIO ESTIMATOR

When we are interested in estimating the ratio of two ratios, such as the ratio of per capita income between two successive years in a country, the ratio of birth rates over a period of one year in a region etc., the question of double ratio estimation arises. In this case, if r_1 and r_2 are the ratio estimators for the ratios R_1 and R_2 , then we may take r_1/r_2 as an estimator of R_1/R_2 which estimator is termed a *double ratio estimator*. Let (y_1, x_1) and (y_2, x_2) be unbiased estimators of (Y_1, X_1) and (Y_2, X_2) , the population totals of y and x respectively.

for the current period and the previous period. Then the double ratio estimator of $R(=R_1/R)$ is of the form

$$\hat{R} = \frac{r_1}{r_2} = \left(\frac{y_1}{x_1} \right) \Big/ \left(\frac{y_2}{x_2} \right) = \frac{y_1}{x_1} \frac{x_2}{y_2}. \quad \dots \quad (10.75)$$

For obtaining the expected value and the variance of this estimator, we may proceed as before by substituting $y_1 = Y_1(1+e_1)$, $y_2 = Y_2(1+e_2)$, $x_1 = X_1(1+e'_1)$ and $x_2 = X_2(1+e'_2)$ in the estimator, that is,

$$\begin{aligned} \hat{R} &= (R_1/R_2)\{(1+e_1)(1+e'_2)(1+e'_1)^{-1}(1+e_2)^{-1}\} \\ &= (R_1/R_2)\{1+(e_1+e'_2-e'_1-e_2)+\dots\}. \end{aligned} \quad \dots \quad (10.76)$$

By taking the expected values of $(\hat{R}-R)$ and $(\hat{R}-R)^2$, we find that the bias is zero to the first order of approximation, that is, when the second and higher powers of terms in e 's are neglected and that the mse, which is the same as the variance for the first order of approximation, is

$$\begin{aligned} V(\hat{R}) &= \left(\frac{R_1}{R_2} \right)^2 \left[\frac{1}{Y_1^2} \{V(y_1)-2R_1 \text{ Cov}(x_1, y_1)+R_1^2 V(x_1)\} \right. \\ &\quad + \frac{1}{Y_2^2} \{V(y_2)-2R_2 \text{ Cov}(x_2, y_2)+R_2^2 V(x_2)\} \\ &\quad + 2 \left\{ \frac{\text{Cov}(y_1, x_2)}{Y_1 X_2} + \frac{\text{Cov}(y_2, x_1)}{Y_2 X_1} \right. \\ &\quad \left. \left. - \frac{\text{Cov}(y_1, y_2)}{Y_1 Y_2} - \frac{\text{Cov}(x_1, x_2)}{X_1 X_2} \right\} \right]. \end{aligned} \quad \dots \quad (10.77)$$

It is of interest to note that the double ratio estimator can also be used to improve the estimator of the total Y_1 , if the population totals Y_2 , X_2 and X_1 are known. For instance, in an industrial enterprise survey, if information on the number of workers (X_1) is available from a current census and if data on the number of workers (X_2)

and output (Y_2) are available from an earlier census then for the estimation of the current total output (Y_1) we may consider the double ratio estimator which is of the form

$$\hat{Y}_1 = \left\{ \left(\frac{y_1}{x_1} X_1 \right) \Big/ \left(\frac{y_2}{x_2} X_2 \right) \right\} Y_2 = \left(\frac{y_1}{x_1} \frac{x_2}{y_2} \right) \left(\frac{Y_2}{X_2} X_1 \right) \quad (10.78)$$

The bias of \hat{Y}_1 is zero for the first order of approximation and the variance of the estimator can be obtained by multiplying the expression for $V(\hat{R})$ in (10.77) by $(Y_2 X_1 / X_2)^2$, (cf Problem 10.20, p 403). This extension of the ratio method to double ratio method of estimation is due to Keyfitz (Yates, 1960, p 343).

10.12d MULTIPLE AUXILIARY VARIABLES

When adequate supplementary information is available, it is possible to improve on the ratio and product estimators. For instance, if the study variable y is positively correlated to k available auxiliary variables $\{x_t\}$ then Olkin (1958) has suggested a composite estimator of the form

$$\hat{Y}_r = \sum_{t=1}^k w_t (\hat{Y}/\hat{X}_t) X_t \quad \left(\sum_{t=1}^k w_t = 1 \right), \quad (10.79)$$

where \hat{Y} and $\{\hat{X}_t\}$ are the estimators of Y and the totals $\{\hat{X}_t\}$ of the k auxiliary variables respectively. An estimator of the form

$$\hat{Y}_p = \sum_{t=1}^k u_t (\hat{Y}/\hat{X}_t)/X_t \quad \left(\sum_{t=1}^k u_t = 1 \right), \quad (10.80)$$

can be used, if the estimators $\{\hat{X}_t\}$ are negatively correlated with \hat{Y} . The estimators (10.79) and (10.80) may be termed *multi variable ratio estimator* and *multi variable product estimator* respectively. The possibility of using a linear combination of ratio and product estimators may be resorted to when estimators of the totals of some auxiliary variables are highly positively correlated and the others highly negatively correlated with \hat{Y} .

Robson and Vithyasai (1961) have suggested the splitting of the study variable y into the sum of k variables $\{y_i\}$ positively correlated with the available auxiliary variables $\{x_i\}$, $i = 1, 2, \dots, k$, and using the estimator of the form

$$\hat{Y}_r^* = \sum_{i=1}^k (\hat{Y}_i / \hat{X}_i) X_i, \quad \dots \quad (10.81)$$

where (\hat{Y}_i, \hat{X}_i) are the estimators of the totals (Y_i, X_i) of the split variable y_i and the i -th auxiliary variable x_i , and such an estimator is termed *component-wise ratio estimator*. If the variables $\{y_i\}$ are negatively correlated with $\{x_i\}$, the possibility of using a *component-wise product estimator* of the form

$$\hat{Y}_p^* = \sum_{i=1}^k (\hat{Y}_i \hat{X}_i) / \bar{X}_i \quad \dots \quad (10.82)$$

can be explored. Also a linear combination of ratio and product estimators can be used if some y_i 's are positively and the others negatively correlated with the corresponding x_i 's.

Ratio cum Product Estimators

M. P. Singh (1965, 1967) has suggested the following *ratio-cum-product* estimators for estimating the population total of the study variable when supplementary information is available on two auxiliary variables :

$$\hat{Y}_{R1}^* = \left(\frac{\hat{Y}}{\hat{X}_1} X_1 \right) \frac{\hat{X}_2}{X_2}$$

and

$$\hat{Y}_{R2}^* = \left(\frac{\hat{Y}}{\hat{X}_2} X_2 \right) \frac{\hat{X}_1}{X_1}.$$

The approximate expressions for the mse's of these two estimators are given by

$$M(\hat{Y}_{R1}^*) = M(\hat{Y}_R) + Y^2[C_2^2 + 2\rho_{02}C_0C_2 - 2\rho_{12}C_1C_2] \quad \dots \quad (10.83)$$

and

$$M(\hat{Y}_{R2}^*) = M(\hat{Y}_R) + Y^2[C_2^2 - 2\rho_{02}C_0C_2 + 2\rho_{12}C_1C_2], \quad \dots \quad (10.84)$$

where $M(\hat{Y}_R) = Y^2[C_0^2 - 2\rho_{01}C_0C_1 + C_1^2]$ (cf. (10.11)), C_0 , C_1 and C_2 are the rse's of \hat{Y} , \hat{X}_1 and \hat{X}_2 , and ρ_{01} , ρ_{02} and ρ_{12} are the correlation coefficients between (\hat{Y}, \hat{X}_1) , (\hat{Y}, \hat{X}_2) and (\hat{X}_1, \hat{X}_2) respectively. It can be easily shown that the estimators \hat{Y}_{R1}^* and \hat{Y}_{R2}^* would respectively be more efficient than \hat{Y}_R if

$$\rho_{02}(C_0/C_2) - \rho_{12}(C_1/C_2) < -\frac{1}{2} \quad (10.85)$$

and

$$\rho_{02}(C_0/C_2) - \rho_{12}(C_1/C_2) > +\frac{1}{2} \quad (10.86)$$

It may be noted that though these conditions do not depend on ρ_{01} , they would get changed according to the signs of \hat{Y} , \hat{X}_1 and \hat{X}_2 .

M P Singh has also considered the question of improving the ratio estimator $\hat{R} = \hat{Y}/\hat{X}_1$ by the estimators $\hat{R}_1^* = \hat{R}(\hat{X}_2/\bar{X}_2)$ and $\hat{R}_2^* = \hat{R}(\bar{X}_2/\hat{X}_2)$ subject to the conditions given in (10.85) and (10.86) respectively (cf Problem 10.23, p 404). Further in the case of improving the estimator $\hat{P}(-\hat{Y}\hat{X}_1)$ of the product YX_1 by the estimator $\hat{P}_1^* = (\hat{P}\hat{X}_2)/\bar{X}_2$ and $\hat{P}_2^* = (\hat{P}\bar{X}_2)/\hat{X}_2$ the conditions for \hat{P}_1^* and \hat{P}_2^* to be better than \hat{P} become

$$\rho_{02}(C_0/C_2) + \rho_{12}(C_1/C_2) < -\frac{1}{2} \quad (10.87)$$

and

$$\rho_{02}(C_0/C_2) + \rho_{12}(C_1/C_2) > +\frac{1}{2} \quad (10.88)$$

respectively. If $C_0 = C_1 = C_2$ the conditions (10.85) to (10.88) may be illustrated with the help of the configurations given in Figure 10.1

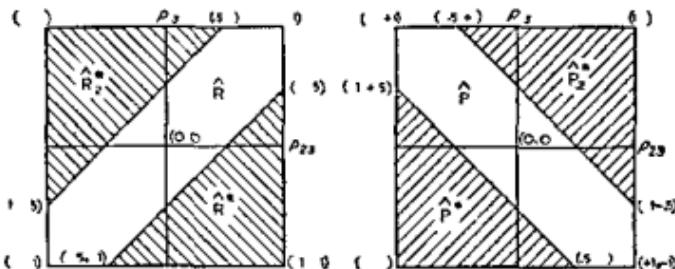


Figure 10.1 Configurational representation of the regions of preference for
(i) \hat{R} , \hat{R}_1^* and \hat{R}_2^* and (ii) \hat{P} , \hat{P}_1^* , and \hat{P}_2^*

10.13 TWO-PHASE SAMPLING

We have considered in Section 9.12 of Chapter 9 (p 349) the technique of two phase sampling where a large initial sample is selected for observing the values of a suitable auxiliary variable, and then collecting data on the study variable for a sub sample of the initial sample. The supplementary information collected for the first phase sample may either be utilized for selecting the subsequent sub sample in an

efficient manner or used at the estimation stage for getting an efficient estimator and here we consider the latter possibility. Suppose n_1 units are selected in the first phase and n_2 units in the second phase according to any specified sample design. Let \hat{X}_1 be an unbiased estimator of X based on the first phase sample and \hat{Y}_2 and \hat{X}_2 be unbiased estimators of Y and X based on the second phase sample. Then an unbiased estimator of Y is given by

$$\hat{Y} = \frac{\hat{Y}_2}{\hat{X}_2} \hat{X}_1. \quad \dots \quad (10.89)$$

By substituting $\hat{Y}_2 = Y(1+e_2)$, $\hat{X}_2 = X(1+e'_2)$, $\hat{X}_1 = X(1+e'_1)$ in (10.79), we get

$$\begin{aligned}\hat{Y} &= Y\{(1+e_2)(1+e'_1)(1+e'_2)^{-1}\} \\ &= Y\{1+(e_2+e'_1-e'_2)+\dots\}.\end{aligned}$$

Taking the expected values of $(\hat{Y} - Y)$ and $(\hat{Y} - Y)^2$, we see that the bias is zero to the first order of approximation and the mse, which is the same as the variance for the first order of approximation, is given by

$$\begin{aligned}V(\hat{Y}) &= V(\hat{Y}_2) - 2R \text{ Cov}(\hat{Y}_2, \hat{X}_2) + R^2 V(\hat{X}_2) + R^2 V(\hat{X}_1) \\ &\quad + 2R \text{ Cov.}(\hat{Y}_2, \hat{X}_1) - 2R^2 \text{ Cov}(\hat{X}_1, \hat{X}_2). \quad \dots \quad (10.90)\end{aligned}$$

Noting that in the case of two-phase sampling with srswr at both the phases the expressions for the variances and the covariances occurring in (10.90) are given by

$$V(\hat{Y}_2) = N^2 \sigma_y^2 / n_2, \quad V(\hat{X}_2) = N^2 \sigma_x^2 / n_2, \quad V(\hat{X}_1) = N^2 \sigma_x^2 / n_1,$$

$$\text{Cov}(\hat{Y}_2, \hat{X}_2) = N^2 \rho \sigma_x \sigma_y / n_2, \quad \text{Cov}(\hat{Y}_2, \hat{X}_1) = N^2 \rho \sigma_x \sigma_y / n_1, \quad \text{Cov}(\hat{X}_1, \hat{X}_2) = N^2 \sigma_x^2 / n_1,$$

where ρ is the correlation coefficient between x and y , we get the expression for the variance as

$$V(\hat{Y}) = \frac{N^2}{n_2} \left\{ \sigma_y^2 - 2R\rho\sigma_x\sigma_y + R^2\sigma_x^2 \right\} + \frac{N^2}{n_1} \left\{ 2R\rho\sigma_x\sigma_y - R^2\sigma_x^2 \right\}. \quad \dots \quad (10.91)$$

It may be noted that if the first and the second samples had been selected independently from the whole population $\text{Cov}(\hat{Y}, \hat{X}_1)$ and $\text{Cov}(\hat{X}_1, \hat{X}_2)$ would be zero in (10.90). In this case (10.91) becomes

$$V(\hat{Y}) = \frac{N^2}{n_2} \left\{ \sigma_y^2 - 2R\rho\sigma_x\sigma_y + R^2\sigma_x^2 \right\} + \frac{N^2}{n_1} R^2\sigma_x^2 \quad (10.97)$$

10.14 AN EMPIRICAL STUDY

To study the efficiency of the ratio estimator empirically, the village wise crop acreage data for three police station areas (administrative units in West Bengal) are used. The geographical area is taken as the supplementary variable for estimating the area under jute and paddy on the basis of a sample of villages. The following sampling schemes are considered here:

- (i) simple random sampling with replacement,
- (ii) ratio estimation in the case of srswr, and
- (iii) sampling with ppswr, size being area

The relative efficiencies of the estimators in (i) and (ii) compared to that of (iii) are presented in Table 10.3. From this table it may be observed that neither the pps estimator nor the ratio estimator is more efficient than the other uniformly for all the populations studied here though both of them are more efficient than the srs unbiased estimator.

TABLE 10.3 EFFICIENCIES OF SRS,
PPS AND RATIO ESTIMATORS

police station area	srs estimator		pps unbiased estimator
	unbiased	ratio	
(1)	(2)	(3)	(4)
<i>jute</i>			
1	54	118	100
2	77	84	100
3	69	99	100
<i>autumn paddy</i>			
1	42	79	100
2	26	180	100
3	26	58	100
(1) Haringhata , (2) Hanskhali , (3) Santipur)			

REFERENCES

- COCHRAN, W. G. (1940) : The estimator of the yields of the cereal experiments by sampling for the ratio of grain to total produce; *J. Agr. Sci.*, 37, 199–212.
- DES RAJ (1954) : Ratio estimation in sampling with equal and unequal probabilities; *J. Ind. Soc. Agr. Stat.*, 6, 127–138.
- DURBIN, J. (1959) : A note on the application of Quenouille's method of bias reduction to the estimation of ratios; *Biometrika*, 46, 477–480.
- GOODMAN, L. A. and HARTLEY, H. O. (1958) : The precision of unbiased ratio-type estimators; *J. Amer. Stat. Assn.*, 53, 491–508.
- GOODMAN, L. A. (1960) : On the exact variance of products; *J. Amer. Stat. Assn.*, 55, 708–713.
- HARTLEY, H. O. and ROSS, A. (1954) : Unbiased ratio estimators; *Nature*, 174, 270–271.
- LAHIRI, D. B. (1951) : A method of sample selection providing unbiased ratio estimates; *Bull. Inter. Stat. Inst.*, 33, (2), 133–140.
- MICKEY, M. R. (1959) : Some finite population unbiased ratio and regression estimators; *J. Amer. Stat. Assn.*, 54, 594–612.
- MIDZUNO, H. (1952) : On a sampling system with probability proportional to sum of sizes; *Ann. Inst. Stat. Math.*, 3, 99–107.
- MURTHY, M. N. and NANJAMMA, N. S. (1959) : Almost unbiased ratio estimates based on interpenetrating sub-samples; *Sankhyā*, 21, 381–392.
- MURTHY, M. N. (1964) : Product method of estimation; *Sankhyā*, 26, (A), 69–74.
- NANJAMMA, N. S., MURTHY, M. N. and SETHI, V. K. (1959) : Some sampling systems providing unbiased ratio estimators; *Sankhyā*, 21, 299–314.
- NIETO DE PASCUAL, J. (1961) : Unbiased ratio estimators in stratified sampling; *J. Amer. Stat. Assn.*, 56, 70–87.
- OLKIN, I. (1958) : Multi-variate ratio-estimation for finite populations; *Biometrika*, 45, 154–165.
- QUENOUILLE, M. H. (1956) : Note on bias in estimation; *Biometrika*, 43, 353–360.
- RAO, T. J. (1966) : On certain unbiased ratio estimators; *Ann. Inst. Stat. Math.*, 18, 117–121.
- ROBSON, D. S. (1957) : Applications of multi-variate polykays to the theory of unbiased ratio-type estimators; *J. Amer. Stat. Assn.*, 52, 511–522.
- ROBSON, D. S. and VITHYASAI, C. (1961) : Unbiased component-wise ratio estimation; *J. Amer. Stat. Assn.*, 56, 350–358.
- SEN, A. R. (1952) : Present status of probability sampling and its use in estimation of farm characteristics (an abstract); *Econometrica*, 20, 103.
- SINGH, M. P. (1965) : On the estimation of ratio and product of the population parameters; *Sankhyā*, 27, (B) 321–328.

- SINGH, M P (1967) Ratio cum product method of estimation, *Metrika*, 12
- SWAIN, A K P C (1964) The use of systematic sampling in ratio estimate, *J Ind Stat Assn*, 2, 160-161
- WILLIAMS, W H (1961) Generating unbiased ratio and regression estimators, *Biometrika*, 48, 267-274
- YATES, F (1960) *Sampling Methods for Censuses and Surveys*, Third Edition, Charles Griffin & Co, London

COMPLEMENTS AND PROBLEMS

10.1 Using the data given in Table 10.4 for a hypothetical population of six units, compare the efficiency of a ratio estimator of the population total Y based on two units selected with srs wrt with that of the usual unbiased estimator by enumerating all possible samples

TABLE 10.4 VALUES OF y AND x FOR SIX UNITS

variable	U_1	U_2	U_3	U_4	U_5	U_6
(1)	(2)	(3)	(4)	(5)	(6)	(7)
x	0	1	3	5	8	10
y	1	3	11	18	29	46

10.2 For estimating the percentage of absentees in the 325 factories situated in a district, a sample of 43 factories was drawn with srs wrt Utilizing the data given in Table 10.5 estimate the percentage of absentees (R) and its rse, stating the assumptions involved in the calculations

TABLE 10.5 NUMBER OF WORKERS (x) AND NUMBER OF ABSENTEES (y) FOR THE 43 SAMPLE FACTORIES

sr no	x	y									
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	95	9	12	89	4	23	75	6	34	159	18
2	79	7	13	57	5	24	69	8	35	54	13
3	30	3	14	132	13	25	63	5	36	69	14
4	45	2	15	47	4	26	83	7	37	61	1
5	28	3	16	43	9	27	124	13	38	164	35
6	142	8	17	116	12	28	31	2	39	132	21
7	125	9	18	65	8	29	96	23	40	82	5
8	81	10	19	103	9	30	42	13	41	33	4
9	43	6	20	52	8	31	85	18	42	86	11
10	53	2	21	67	14	32	91	14	43	41	10
11	148	16	22	64	6	33	73	7			

10.3 A sample of 34 villages was selected from a population of 170 villages with ppswr, size being cultivated area in 1961, for estimating the area under wheat in the region during 1964. Later it was found that the figures for area under wheat in 1963 were also available for all the villages in the region. The relevant data are given in Table 10.6.

(i) Estimate the area under wheat in 1964 by the method of ratio estimation using the information on wheat area for 1963 and estimate its rse.

(ii) Determine the efficiency of the ratio estimate as compared to that of the usual unbiased estimate.

TABLE 10.6. CULTIVATED AREA IN 1961 (x_1) AND AREA UNDER WHEAT IN 1963 (x_2) AND IN 1964 (y) FOR 34 SAMPLE VILLAGES.

sr. no.	x_1	x_2	y	sr. no.	x_1	x_2	y
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1.	401	70	50	18.	186	45	27
2.	630	163	149	19.	1767	564	515
3.	1194	320	284	20.	604	238	249
4.	1170	440	381	21.	700	92	85
5.	1065	250	278	22.	524	247	221
6.	827	125	111	23.	571	134	133
7.	1737	558	634	24.	962	131	144
8.	1060	254	278	25.	407	129	103
9.	360	101	112	26.	715	190	175
10.	946	359	355	27.	845	363	335
11.	470	109	99	28.	1016	235	219
12.	1625	481	498	29.	184	73	62
13.	827	125	111	30.	282	62	79
14.	96	5	6	31.	194	71	60
15.	1304	427	339	32.	439	137	100
16.	377	78	80	33.	854	196	141
17.	259	75	105	34.	820	255	263

(x_1 , x_2 and y are in acres; total cultivated area in 1961 : 78,000 acres; total area under wheat in 1963 : 21,288 acres; 1 acre : 0.4047 hectares).

10.4 In an experimental study in a large paddy field, the weight of grain with straw (x) and the grain yield (y) were obtained for a number of square area units selected at random over the field and the estimates of C_x^2 , C_y^2 , and C_{xy} were found to be 1.20, 1.24 and 0.81 respectively, where C_x and C_y are the coefficients of variation for the variables x and y and $C_{xy} = \rho C_x C_y$, ρ being the correlation coefficient between x and y . When the total of x is available for the whole field, determine the efficiency of the ratio estimate of the total of y based on the ratio of y to x in the sample area units compared to that of the unbiased estimate based on the average of the grain yield per sample area unit.

10.5 For estimating the total (T) of current population in a region, two subsamples of 6 villages each are selected systematically from each stratum with independent random starts. Using the data given in Table 10.7, obtain a ratio estimate for Y taking the previous census population (x) as the supplementary information and compare its efficiency with that of the usual unbiased estimate. Also examine whether the ratio estimate is approximately unbiased.

TABLE 10.7 TOTAL NUMBER OF VILLAGES (N) AND SAMPLE TOTALS OF x AND y

stratum number	no of villages N	sub sample 1		sub sample 2	
		x	y	x	y
(1)	(2)	(3)	(4)	(5)	(6)
1	2044	3722	3935	3456	3641
2	1304	3625	4033	4171	4649
3	1265	2769	3050	3746	4043
4	1252	3180	3498	4323	4722
5	4264	3522	3819	3314	3638
6	1598	2827	2936	3550	3652
7	810	8603	9596	7285	7935
8	567	8323	9135	9595	10024
9	500	9019	9772	11073	12152
10	486	7404	8185	6981	7690
total	14090	—	—	—	—

(total of x for the region 10,155,680)

10.6 From an urban area consisting of 1840 households with a total population of 8846 persons, a sample of 10% of households is drawn with srs w/o for estimating the total income in that area (Y). Using the data given in Table 10.8, obtain a ratio

TABLE 10.8 DISTRIBUTION OF SAMPLE HOUSEHOLDS BY INCOME GROUPS AND HOUSEHOLD SIZE

house hold size	income group (in rupees)					
	1-50	51-100	101-150	151-200	201-300	301-500
(1)	(2)	(3)	(5)	(5)	(6)	(7)
1	2	6	1	1	—	—
2	6	9	12	3	2	—
3	5	6	10	13	2	—
4	2	5	8	13	12	3
5	—	3	6	18	2	2
6	—	1	2	5	7	3
7	—	—	1	2	4	2
8	—	—	—	1	3	1

estimate for Y taking number of persons as the supplementary variable and estimate its rse. Also calculate the efficiency of the ratio estimator compared to the unbiased estimator.

10.7 For estimating the average land holding area in different land holding size classes, a sample of n holdings is selected from a population of N holdings with srs wr.

(i) Show that the sample means of the holding sizes obtained after classification of the sample into the specified holding size classes are approximately unbiased for large samples and obtain the sampling variance of the sample mean in one of the classes.

(ii) What modification would be needed in the expression for the variance, if the sample is selected with srs wor.

10.8 If y and x are unbiased estimators of the population totals Y and X , show that the ratio of the exact bias of the ratio estimator $(y/x)X$ to its standard error is not greater than C_x , the relative standard error of x . Also show that the bias relative to Y is less than C_x^2 if the rse of y/x is less than C_x .

10.9 If y and x are unbiased estimators of the population totals Y and X , show that the relative variance of the ratio estimator (y/x) can be approximated by $C^2(y) - C^2(x)$, when the correlation coefficient between y/x and x may be assumed to be negligibly small.

(Hansen, M. H., Hurwitz, W. N. and Madow, W. G., *Sample Survey Methods and Theory*, Vol. II, 1953, p. 204).

10.10 If \bar{y} and \bar{x} are unbiased estimators of the population means \bar{Y} and \bar{X} based on n units drawn with srs wr, derive the expressions M_1 and M_2 for the mse of the ratio \bar{y}/\bar{x} neglecting terms of (i) second and higher powers of $(1/n)$, and (ii) third and higher powers of $(1/n)$ when \bar{y} and \bar{x} are distributed in a bivariate normal form. Hence, show that the relative difference $(M_2 - M_1)/M_1$ is not greater than $9C_x^2/n$, where C_x is the coefficient of variation of the variable x in the population. Also show that when $C_x = C_y$ the above relative difference is approximately equal to $3(2-\rho)C_x^2/n$, where ρ is the correlation coefficient between the variables x and y .

(Sukhatme, P. V., *Sampling Theory of Surveys with Applications*, (1953), pp. 151-154; Cochran, W. G., *Sampling Techniques*, (1963), pp. 159-160).

10.11 (i) When the regression between the variable under study y and the supplementary variable x is perfectly linear, that is,

$$Y_i = \alpha + \beta X_i, \quad i = 1, 2, \dots, N,$$

derive the condition for the ratio estimator of the population mean \bar{Y} to be more efficient than the usual unbiased estimator in the case of srs wr. What is the effect of the line of regression passing through the origin on the efficiency of the ratio estimator?

(u) If, instead of the perfect linear regression, it is assumed that the population of N units is drawn from a super population following the model

$$E(y|x) = \alpha + \beta x \quad \text{and} \quad V(y|x) = \sigma^2 = \sigma_y^2(1 - \rho^2),$$

where ρ is the correlation coefficient between y and x , derive the condition for the ratio estimator to be more efficient than the usual unbiased estimator. From this deduce the condition for the case where the line of regression passes through the origin.

(Des Raj, *J. Ind. Soc. Agr. Stat.*, 6, (1954), 127-138)

10.12 Assuming that the population of N units is drawn from a super population with the model

$$E(Y_t|X_t) = \alpha + \beta X_t, \quad V(Y_t|X_t) = \alpha X_t^\beta \quad \text{and} \quad \text{Cov}(Y_t, Y_s | X_t, X_s) = 0, \quad (t \neq s)$$

where α and β are positive constants, compare the efficiency of ratio method of estimation for srswr with that of ppswr sampling for estimating the population mean \bar{Y}

(Des Raj, *J. Amer. Stat. Assn.*, 53, (1958), 98-101)

10.13 Suppose r_1 and r_2 are estimates of a population ratio at two points of time based on a common probability sample. Show that the bias of the difference $(r_1 - r_2)$ relative to its standard error is less than $\sqrt{2}C_{x_2}$ if $\rho(r_1, r_2) \leq 0$ and less than $\sqrt{2}C_{x_2}\sqrt{(\sigma_{r_1}^2 + \sigma_{r_2}^2)/(\sigma_{r_1}^2 - \sigma_{r_2}^2)}$ if $\rho(r_1, r_2) > 0$, where C_{x_2} is the rse for the estimator of the denominator on the second occasion and σ indicates the standard error

(Koop, *J. C. Metrika*, 5, (1962), 145-149)

10.14 Suppose a finite population of N units has NP_1 units belonging to a particular category, of which NP_2 units have a special characteristic, and it is proposed to estimate the population ratio P_2/P_1 on the basis of a sample of n units selected with s.w.r.

(i) If p_1 and p_2 are the sample proportions corresponding to P_1 and P_2 respectively, show that p_2/p_1 is approximately unbiased and derive its approximate variance in case of large samples, stating clearly the assumptions involved, if any

- (ii) If P_1 is known, prove that the estimator p_2/p_1 is more efficient than p_2/P_1
 (iii) When P_2 is known, derive the condition for p_2/p_1 to be more efficient than P_2/P_1

(Elkin, *J. Amer. Stat. Assn.*, 48, (1953), 128-130)

10.15 Show that even the substitution of the unbiased estimators $\hat{X}^2 - v(\hat{X})$ and $\hat{R}^2 - v(\hat{R})$ instead of \hat{X}^2 and \hat{R}^2 for X^2 and R^2 in (10.11) leads to the same variance estimator as given in (10.12)

10.16 Show that the ratio-type estimator given in (10.66) is more efficient than the usual biased ratio estimator \hat{Y}_R , if and only if the slope of the regression line of y on x is closer to $\frac{1}{N} \sum_{i=1}^N \frac{Y_i}{X_i}$ than to $\frac{\bar{Y}}{\bar{X}}$.

(Goodman, L. A. and Hartley, H. O., *J. Amer. Stat. Assn.*, 53, (1958), 491–508).

10.17 Suppose a simple random sample of n units is selected in the form of k independent sub-samples of m units each. Noting that the bias of the ratio estimator \hat{R} based on all the n observations is of the form $(C/n) + O(1/n^2)$, show that the estimator

$$\hat{R}_Q = k\hat{R} - \frac{k-1}{k} \sum_{i=1}^k \hat{R}_i,$$

where \hat{R}_i is the estimate computed after omitting the i -th sub-sample has bias of order $(1/n^2)$ at the most.

(Quenouille, M. H., *Biometrika*, 43, (1956), 353–360).

10.18 Show that if the ratio estimator is of the form (y/x) and if the regression of y on x is linear and x is normally distributed, \hat{R}_Q in Problem 10.17 has asymptotically a smaller variance than \hat{R} when $k = 2$.

(Durbin, J., *Biometrika*, 46, (1959), 477–480).

10.19 Derive the variance estimators given in (10.70) and (10.72) in connection with sampling schemes providing unbiased ratio estimators.

(Des Raj, *J. Ind. Soc. Agr. Stat.*, 6, (1954), 127–138).

10.20 Show that an approximate formula for the variance of the double ratio estimator (10.78) in the case of srs wor is given by

$$V(\hat{Y}_1) = \frac{N-n}{N-1} \frac{Y_1^2}{n} \frac{1}{N} \sum_{i=1}^N \left(\frac{Y_{1i}}{\bar{Y}_1} - \frac{X_{1i}}{\bar{X}_1} - \frac{Y_{2i}}{\bar{Y}_2} + \frac{X_{2i}}{\bar{X}_2} \right)^2$$

and derive a consistent estimator of this variance.

(Yates, F., *Sampling Methods for Censuses and Surveys*, Third Edition, (1960), p. 343).

10.21 Show that for a sample of n units selected with srs wor from a population of N units

$$B(\hat{R}_n \bar{X}) = \frac{n}{N} \frac{N-1}{n-1} E(\bar{x}(\hat{R}_n - \hat{R}_1)), \quad \hat{R}_1 = \frac{\bar{y}}{\bar{x}} \text{ and } \hat{R}_n = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i},$$

\bar{y} and \bar{x} being the sample means for the variables y and x .

Using the above result, show that (i) the combined estimator $\theta_1(\hat{R}_1 \bar{X}) + (1-\theta_1)(\hat{R}_n \bar{X})$, where θ_1 is a variable, is unbiased for \bar{Y} if $\theta_1 = \frac{n}{N} \frac{N-1}{n-1} \frac{\bar{x}}{\bar{X}}$ and (ii) the combined estimator $(1-\theta_2)(\hat{R}_1 \bar{X}) + \theta_2(\hat{R}_n \bar{X})$, where θ_2 is a variable, is almost unbiased if $\theta_2 = -\frac{(N-1)\bar{x}}{N(n-1)\bar{X}}$, (cf. (10.66), (10.68) of Section 10.10).

(Rao, T. J., *Ann. Inst. Stat. Math.*, 18, (1966), 117–121).

10.22 When the population of N units is assumed to be drawn from a super population following the model

$$E(Y_t|X_t) = aX_t, \quad V(Y_t|X_t) = \sigma^2 X_t^g \quad \text{and} \quad \text{Cov}(Y_t, Y_{t'}|X_t, X_{t'}) = 0, \quad (t' \neq t),$$

where $g > 1$, show that the expected variance of the ratio estimator in the case of the Midzuno-Sen scheme given in Sub-section 10.11a is more than that of the Horvitz-Thompson estimator (p 210) with probabilities of inclusion $\{II_t\}$ proportional to $\{X_t\}$

(Rao, T. J., *J. Roy. Stat. Soc. (B)*, 29, (1967))

10.23 Show that the composite estimator $\hat{R}_c^* = \omega \hat{R}_1^* + (1-\omega) \hat{R}_2^*$, where \hat{R}_1^* and \hat{R}_2^* are the two estimators of the ratio R considered in Sub-section 10.12d using the auxiliary variables x_2 and x_3 respectively, has the minimum mse correct to the first order of approximation if $\omega = \frac{1}{2}$ and that the minimum mse is given by

$$M(\hat{R}_c^*) = R^2 C^2 [2(1 - \rho_{01} + \rho_{02} - \rho_{03}) + \frac{1}{2}(1 - \rho_{23})],$$

where it is assumed that (i) the coefficients of variation C_0, C_1, C_2 and C_3 of the estimators $\hat{Y}, \hat{X}_1, \hat{X}_2$ and \hat{X}_3 are the same (C) and (ii) $\rho_{02} = \rho_{13}$ and $\rho_{03} = \rho_{12}$, ρ 's being the correlation coefficients between pairs of estimators such as $(\hat{Y}, \hat{X}_1), (\hat{X}_1, \hat{X}_2)$, etc.

(Singh, M. P., (1967), unpublished)

Difference and Regression Estimators

11.1 ESTIMATING A DIFFERENCE

In this chapter, we consider the question of estimation of change or difference over time or space between population characteristics or between populations for the same characteristic and of using the estimator of certain types of differences to obtain better estimates for the population totals and means of some characteristics. Suppose \hat{Y}_1 and \hat{Y}_2 are unbiased estimators of the population totals Y_1 and Y_2 at the two end-points of a specified period, then we get an unbiased estimator of the difference $D = Y_2 - Y_1$ as

$$d = \hat{Y}_2 - \hat{Y}_1. \quad \dots \quad (11.1)$$

This estimator may be termed a *difference estimator*. The sampling variance of d is given by

$$V(d) = V(\hat{Y}_1) - 2 \operatorname{Cov}(\hat{Y}_1, \hat{Y}_2) + V(\hat{Y}_2), \quad \dots \quad (11.2)$$

where the covariance term $\operatorname{Cov}(\hat{Y}_1, \hat{Y}_2)$ will be zero if the two samples selected at the two points of time are independent or when the estimators are otherwise uncorrelated. Thus we see that the variance of the difference estimator depends much on the correlation between \hat{Y}_1 and \hat{Y}_2 and that the variance decreases with increase in this correlation coefficient, showing that for efficient estimation of the difference, the correlation between the two estimators \hat{Y}_1 and \hat{Y}_2 should be positive and as large as possible. It is of interest to note that even when one of the population totals (Y_1 say) is known, it would be desirable to

use the estimator d given in (11.1) instead of the other difference estimator $d' = \hat{Y}_2 - Y_1$, if \hat{Y}_1 and \hat{Y}_2 are positively correlated.

Since in many practical situations, one is interested in estimating both averages and differences of parameter values over specified periods of time, it is to be noted that the conditions conducive for the efficient estimation of the averages and of the differences of parameters over time are not the same and in fact they are opposite in nature. For, in estimating the average of the population totals of the variable y over a period of time on the basis of a double sample, the estimator is of the form

$$\hat{Y} = \frac{1}{2} (\hat{Y}_1 + \hat{Y}_2) \quad (11.3)$$

and its variance is

$$V(\hat{Y}) = \frac{1}{4} [V(\hat{Y}_1) + 2 \operatorname{Cov}(\hat{Y}_1, \hat{Y}_2) + V(\hat{Y}_2)], \quad . \quad (11.4)$$

which shows that for the estimator \hat{Y} to be efficient, the correlation coefficient should not be positive and large. Hence, one has to strike a balance in arriving at the optimum value of the number of common units between the two samples. This question is discussed in Sub section 11.7b.

11.2 REGRESSION METHOD OF ESTIMATION

In Chapter 10, we considered the question of improving the conventional unbiased estimator \hat{Y} by multiplying it with the factor X/\hat{X} , where \hat{X} is an unbiased estimator of the total of a suitably chosen supplementary variable. Here we examine the possibility of improving upon \hat{Y} by considering the estimator $\hat{X} - X$, which is a *zero function* in the sense that its expected value is zero. For instance, an unbiased estimator of Y can be taken as

$$\hat{Y}' = \hat{Y} + \lambda(\hat{X} - X),$$

where λ is a constant. The value of λ can be so fixed as to minimize the variance of \hat{Y}' , which is given by

$$V(\hat{Y}') = V(\hat{Y}) + 2\lambda \operatorname{Cov}(\hat{X}, \hat{Y}) + \lambda^2 V(\hat{X})$$

Differentiating this with respect to λ and equating the partial derivative to zero, we get

$$\lambda = - \frac{\text{Cov}(\hat{X}, \hat{Y})}{V(\hat{X})}.$$

Thus we see that the optimum value of λ is given by $-\beta$, where β is the regression coefficient of \hat{Y} on \hat{X} and hence the estimator with the optimum value of λ is

$$\hat{Y}'_r = \hat{Y} + \beta(X - \hat{X}), \quad \dots \quad (11.5)$$

its variance being

$$V(\hat{Y}'_r) = V(\hat{Y})\{1 - \rho^2(\hat{X}, \hat{Y})\}, \quad \dots \quad (11.6)$$

where $\rho(\hat{X}, \hat{Y})$ is the correlation coefficient between \hat{X} and \hat{Y} . The estimator \hat{Y}'_r is termed the *regression estimator* and the procedure of estimation is known as *regression method of estimation*.

From $V(\hat{Y}'_r)$ given in (11.6), we see that this estimator would be efficient if the estimators \hat{Y} and \hat{X} are highly correlated. This estimator is more efficient than the conventional estimator \hat{Y} , if $\rho(\hat{X}, \hat{Y})$ is non-zero, which is usually satisfied in practice. This condition is less severe than that for the ratio estimator to be more efficient than \hat{Y} . In fact, comparing (11.6) with the variance of the ratio estimator, namely,

$$V(\hat{Y}_R) = V(\hat{Y}) - 2R \text{Cov}(\hat{X}, \hat{Y}) + R^2 V(\hat{X}),$$

we find $V(\hat{Y}_R) - V(\hat{Y}'_r) = \{\rho\sigma(\hat{Y}) - R\sigma(\hat{X})\}^2$, which shows that a regression estimator is more efficient than the corresponding ratio estimator in general and that they are equally efficient when $\beta = R$, that is, when the line of regression passes through the origin.

In actual practice the exact value of β may not be known and it may have to be estimated on the basis of a sample. If $\hat{\beta}$ is an estimator of β , we get

$$\hat{Y}_r = \hat{Y} + \hat{\beta}(X - \hat{X}). \quad \dots \quad (11.7)$$

The estimator (11.7) is generally biased for Y and its bias and variance are considered in Section 11.3.

It may be mentioned that for this estimator to be efficient it is not necessary to get the exact value or an estimate of β based on a current sample and that even an approximation to it available from previous survey or census may be sufficient. However, the closer the approximation, the higher will be the efficiency. So long as the value of β used is uncorrelated with \hat{X} , the regression estimator remains unbiased. Noting that $\hat{\beta} = \hat{Y}/\hat{X}$, when the line of regression of \hat{Y} on \hat{X} passes through the origin, we see that the regression estimator (11.7) reduces to the ratio estimator $\hat{Y}_R = (\hat{Y}/\hat{X})X$.

11.3 BIAS AND VARIANCE

The regression estimator given in (11.7) is biased since (i) the regression coefficient β is generally estimated by the ratio of an estimator of $\text{Cov}(\hat{X}, \hat{Y})$ to that of $V(\hat{X})$ and (ii) it involves the product of two estimators, namely, $\hat{\beta} \hat{X}$. Writing

$$\hat{Y} = Y(1+e), \quad \hat{X} = X(1+e') \quad \text{and} \quad \hat{\beta} = \beta(1+e'')$$

and substituting in (11.7), we get

$$\hat{Y}_r = Y + (eY - e\beta X) - e'e''\beta X \quad (11.8)$$

The bias of the regression estimator is given by

$$B(\hat{Y}_r) = E(\hat{Y}_r) - Y = -E(e'e'')\beta X = -\text{Cov}(\hat{X}, \hat{\beta}), \quad (11.9)$$

since $E(e) = E(e') = 0$. It may be noted that for a large sample size the bias is expected to be negligible, because usually $\text{Cov}(\hat{X}, \hat{\beta})$ will decrease as the sample size increases.

If the sample is selected in the form of m independent sub samples, then the bias can be unbiasedly estimated by

$$b(\hat{Y}_r) = \frac{1}{m-1} \sum_{i=1}^m (\hat{X}_i - \bar{\hat{X}})(\hat{\beta}_i - \bar{\beta}), \quad (11.10)$$

where $\hat{\beta}_i$ and \hat{X}_i are estimators of β and X based on the i -th subsample and $\hat{\beta} = \frac{1}{m} \sum_{i=1}^m \hat{\beta}_i$ and $\hat{X} = \frac{1}{m} \sum_{i=1}^m \hat{X}_i$, provided $\hat{\beta}$ in \hat{Y}_r is also taken as the mean of $\hat{\beta}_i$'s. The estimator of bias given in (11.10) can be used to make the regression estimator unbiased by correcting it for the bias (Mickey, 1959; Williams, 1961). The technique developed by Murthy (1962) for obtaining unbiased or almost unbiased estimators of any non-linear parametric function may also be applied to get an unbiased regression estimator.

When the sample size is large, the term involving $e'e''$ in (11.8) is likely to be negligible, in which case the bias of the estimator is zero and this amounts to assuming β to be known. Hence, the variance of \hat{Y} , to the first order of approximation becomes

$$V(\hat{Y}_r) = V(\hat{Y}) - 2\beta \text{Cov}(\hat{X}, \hat{Y}) + \beta^2 V(\hat{X}),$$

which reduces to (11.6), since $\beta = \text{Cov}(\hat{X}, \hat{Y})/V(\hat{X})$. A consistent estimator of the variance may be obtained by substituting estimators of $V(\hat{Y})$ and $\rho(\hat{X}, \hat{Y})$ in the variance expression.

The regression estimator is not commonly used in practice due to the fact that the calculation of the estimate of the regression coefficient in large-scale surveys becomes cumbersome and time-consuming. Further, since the regression line passes through the origin or close to the origin in most of the cases usually met with, the ratio estimator is generally used instead of the more complicated regression estimator.

The use of the regression method of estimation has been briefly mentioned by Mahalanobis (1941), and the theoretical basis of the regression method has been discussed in detail by Cochran (1942). The use of difference and regression estimators in double or multiple sampling on successive occasions has been considered, among others, by Bose (1943), Patterson (1950), Seal (1951), Tikkiwal (1953), Zarkovich (1956) and Kulldorf (1963).

11.4 SRS AND STRATIFIED SRS

In the previous sections we have considered the regression method of estimation in the general case for any sampling design. In this section, this method is applied to srs w/o r with and without stratification.

11.4a SIMPLE RANDOM SAMPLING

In sampling n units with srs w/o r, an estimator of β , the coefficient of the regression of \bar{y} on \bar{x} , which is the same as that of y on x , is given by

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad . \quad (11.11)$$

Substituting this in the regression estimator (11.7), we get

$$\hat{Y}_r = N[\bar{y} + \hat{\beta}(\bar{X} - \bar{x})] \quad (11.12)$$

The bias of the estimator is given by $-\text{Cov}(\hat{\beta}, \bar{x})$. Noting that $V(\bar{y}) = (1-f)\sigma_y^2/n$, where f is the sampling fraction, we get the variance of \hat{Y}_r to the first order approximation as

$$V(\hat{Y}_r) = N^2(1-f)\sigma_y^2(1-\rho^2)/n \quad (11.13)$$

An estimator of $V(\hat{Y}_r)$ is given by

$$v(\hat{Y}_r) = N^2 \frac{(1-f)}{n(n-1)} \sum_{i=1}^n \{(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\}^2 \quad (11.14)$$

It is of interest to note that when the regression of y on x is perfectly linear, that is, when $|\rho| = 1$, the variance becomes zero, and that if y and x are uncorrelated, the variance is the same as in the case of the conventional unbiased estimator.

11.4b STRATIFIED SRS WOR

Suppose the population is divided into K strata and in the s -th stratum n_s units are selected from N_s units with srs wor. As in the case of ratio estimation, we have two possible estimators in this case also, which are obtained by estimating the overall regression coefficient and those of the different strata. Thus we get the two estimators of Y as

$$\hat{Y}_r = \sum_{s=1}^K N_s \bar{y}_s + \hat{\beta}(X - \sum_{s=1}^K N_s \bar{x}_s), \quad \dots \quad (11.15)$$

where \bar{x}_s and \bar{y}_s are the sample means in the s -th stratum for the variables x and y , and $\hat{\beta}$ is given by

$$\hat{\beta} = \frac{\sum_{s=1}^K w_s \sum_{i=1}^{n_s} (y_{si} - \bar{y}_s)(x_{si} - \bar{x}_s)}{\sum_{s=1}^K w_s \sum_{i=1}^{n_s} (x_{si} - \bar{x}_s)^2},$$

where $w_s = N_s^2 (1-f_s)/n_s(n_s-1)$ and $f_s = n_s/N_s$, and

$$\hat{Y}_r^* = \sum_{s=1}^K N_s \{\bar{y}_s + \hat{\beta}_s (\bar{X}_s - \bar{x}_s)\}, \quad \dots \quad (11.16)$$

where

$$\hat{\beta}_s = \frac{\sum_{i=1}^{n_s} (y_{si} - \bar{y}_s)(x_{si} - \bar{x}_s)}{\sum_{i=1}^{n_s} (x_{si} - \bar{x}_s)^2}.$$

The estimator \hat{Y}_r is termed a *combined regression estimator* and the other estimator \hat{Y}_r^* may be termed *separate regression estimator*.

The sampling variances of these two estimators are approximately given by

$$V(\hat{Y}_r) = \sum_{s=1}^K N_s^2 (1-f_s)(1-\rho^2) \frac{\sigma'_{sy}^2}{n_s}, \quad \dots \quad (11.17)$$

where ρ is the overall correlation coefficient and $\sigma'_{sy}^2 = N_s \sigma_{sy}^2 / (N_s - 1)$, σ_{sy}^2 being the variance of y in the s -th stratum, and

$$V(\hat{Y}_r^*) = \sum_{s=1}^K N_s^2 (1-f_s)(1-\rho_s^2) \frac{\sigma'_{sy}^2}{n_s}, \quad \dots \quad (11.18)$$

where ρ_s is the correlation coefficient between y and x in the s th stratum. Though the combined regression estimator is likely to be less efficient than the separate regression estimator from the consideration of sampling variance, the former is likely to have lesser bias than the latter. Approximate estimators of the variances given in (11.17) and (11.18) are given by

$$v(\hat{Y}_r) = \sum_{s=1}^K \frac{N_s^2(1-f_s)}{n_s(n_s-1)} \sum_{i=1}^{n_s} \{(y_{si}-\bar{y}_s) - \hat{\beta}(x_{si}-\bar{x}_s)\}^2 \quad (11.19)$$

and

$$v(\hat{Y}_r^*) = \sum_{s=1}^K \frac{N_s^2(1-f_s)}{n_s(n_s-1)} \sum_{i=1}^{n_s} \{(y_{si}-\bar{y}_s)^2 - \hat{\beta}_s^2(x_{si}-\bar{x}_s)^2\} \quad (11.20)$$

11.5 TWO-PHASE SAMPLING

The regression estimator assumes the knowledge of the population total X of the supplementary variable x . There are situations where the population total of x is not known in advance and in such cases a two phase sampling scheme can be used for getting a regression estimator, where a sample of n_1 units is selected in the first phase according to a suitable sampling scheme and the value of the supplementary variable x is observed for all these n_1 sample units and then in the second phase a sub sample of n_2 units is selected from the first phase sample of n_1 units and the value of the study variable y is observed for each of these n_2 units. This selection procedure has been briefly discussed earlier in Sections 9.12 (p. 349) and 10.13 (p. 394). Let \hat{X}_1 be an estimator of X based on the first phase sample and \hat{X}_2 , \hat{Y}_2 and $\hat{\beta}_2$ be estimators of X , Y and β respectively based on the second phase sample. Then a regression estimator of Y is given by

$$\hat{Y}_r = \hat{Y}_2 + \hat{\beta}_2(\hat{X}_1 - \hat{X}_2) \quad (11.21)$$

11.5a VARIANCE OF ESTIMATOR

If n_2 is fairly large, the bias in the above estimator is likely to be negligible and the variance of this estimator is approximately equal to that of

$$\hat{Y}_r = \hat{Y}_2 + \beta(\hat{X}_1 - \hat{X}_2). \quad \dots \quad (11.22)$$

The variance of \hat{Y}_r is to be obtained over the two phases and is given by

$$V(\hat{Y}_r) = V_1 E_2(\hat{Y}_r) + E_1 V_2(\hat{Y}_r),$$

where E_1 and V_1 denote the unconditional expected value and variance over first phase sampling and E_2 and V_2 denote the conditional expected value and variance over second phase sampling for a given first phase sample, (cf. Section 2.8 of Chapter 2, p. 41). Noting that

$$V_1 E_2(\hat{Y}_r) = V(\hat{Y}_1) \text{ and } V_2(\hat{Y}_r) \doteq V_2(\hat{Y}_2)(1-\rho^2),$$

where \hat{Y}_1 is an unbiased estimator of Y that would be obtained if y were observed for all the units in the first phase sample, we get

$$V(\hat{Y}_r) \doteq V(\hat{Y}_1) + E_1 V_2(\hat{Y}_2)(1-\rho^2).$$

Assuming that sampling is done with replacement in the first phase, in which case $V(\hat{Y})$ and $V(\hat{X})$ are of the form V_y/n and V_x/n , where V_y and V_x are the variance for a sample of one unit for the characteristics y and x respectively, and that the second phase sample has been selected with srswr from the first phase sample, we find

$$E_1 V_2(\hat{Y}_r) \doteq \left(1 - \frac{n_2}{n_1}\right) V(\hat{Y}_2)(1-\rho^2),$$

for

$$E_1 V_2(\hat{Y}_2) = V(\hat{Y}_2) - V_1 E_2(\hat{Y}_2) = \left(1 - \frac{n_2}{n_1}\right) V(\hat{Y}_2).$$

Substituting the relevant values in $V(\hat{Y}_r)$, we get

$$V(\hat{Y}_r) = \left[1 - \left(1 - \frac{n_2}{n_1} \right) \rho^2 \right] V(\hat{Y}_2) = \left[\frac{\rho^2}{n_1} + \frac{1-\rho^2}{n_2} \right] V_y \quad (11.23)$$

Comparing (11.23) with (11.6) which in this case is $(1-\rho^2) V_y/n_1$, we see that the variance in two phase sampling is more than that in uni phase sampling. But it may be noted that the collection of information on y for all the n_1 units in the first phase may be costly and hence the cost of two phase sampling is likely to be less than that of uni-phase sampling. Hence, there is the need to strike a balance between the cost and variance aspects in practice.

11.5b COST ASPECT

If the cost of surveying a unit for x is C_1 and that for y is C_2 , the cost of survey in two phase sampling is given by

$$C = C_0 + n_1 C_1 + n_2 C_2, \quad (11.24)$$

where C_0 is the overhead cost. The optimum values of n_1 and n_2 may be obtained by minimizing the variance (11.23) for a fixed cost, or minimizing the cost for a fixed variance. In the former case, the optimum values of n_1 and n_2 are given by

$$n_1 = \frac{C' - C_0}{C_1} \left\{ 1 + \sqrt{\frac{C_2}{C_1} \frac{1-\rho^2}{\rho^2}} \right\}^{-1} \quad (11.25)$$

and

$$n_2 = n_1 \sqrt{\frac{1-\rho^2}{\rho^2} \frac{C_1}{C_2}}, \quad (11.26)$$

where C' is the fixed cost and the minimum variance is given by

$$V_m(\hat{Y}_r) = \frac{V_y}{C' - C_0} \left(\sqrt{C_1 \rho^2} + \sqrt{C_2 (1-\rho^2)} \right)^2 \quad (11.27)$$

In the latter case where the variance is pre-fixed at V_0 , we get the optimum values of n_1 and n_2 which minimize the cost as

$$n_1 = \frac{V_y}{V_0} \rho^2 \left\{ 1 + \sqrt{\frac{1-\rho^2}{\rho^2} \frac{C_2}{C_1}} \right\} \quad \dots \quad (11.28)$$

and

$$n_2 = n_1 \sqrt{\frac{1-\rho^2}{\rho^2} \frac{C_1}{C_2}}. \quad \dots \quad (11.29)$$

The minimum cost in this case is

$$C_m = C_0 + \frac{V_y}{V_0} \left\{ \sqrt{C_1 \rho^2} + \sqrt{C_2 (1-\rho^2)} \right\}^2. \quad \dots \quad (11.30)$$

11.6 SAMPLING ON SUCCESSIVE OCCASIONS

In repetitive surveys for estimating the same characteristic at different points of time, it is possible to use the information collected on the previous occasion to improve upon the conventional estimator for the current period by using the difference method of estimation. Generally the main objective of repetitive surveys is to estimate the change from period to period. The method of sampling units on successive occasions, termed *multiple sampling*, consists in selecting the samples on the different occasions such that they have none, some or all units common with the samples selected on the previous occasions, and when there are only two occasions involved, the sampling scheme is known as *double sampling*. If the sampling on successive occasions is done with partial replacement of the sample units according to a specific pattern, it is termed *rotation sampling* (Hansen, Hurwitz, Nisselson and Steinberg, 1955; J. N. K. Rao and Graham, 1964).

Suppose samples of size n units are drawn on each of two occasions according to the same sampling scheme such that m units are common between the two samples and let $u = n-m$ be the number of uncommon units in the two samples. If \hat{Y}_{im} , \hat{Y}_{iu} and \hat{Y}_{in} are the estimators of Y_i , the population total on the i -th occasion based on the

common, uncommon and all sample units respectively and if $\hat{\beta}$ is the estimator of regression coefficient based on the common part of the samples we get

$$\hat{Y}'_{2m} = \hat{Y}_{2m} + \hat{\beta}(\hat{Y}_{1n} - \hat{Y}_{1m}) \quad (11.31)$$

and assuming the variances of the estimators of the population totals on the two occasions based on one sample unit to be the same V_y , we get its variance as

$$V(\hat{Y}'_{2m}) = \left\{ \frac{\rho^2}{n} + \frac{1-\rho^2}{m} \right\} V_y \quad (11.32)$$

Another estimator of Y is given by \hat{Y}_{2u} and its variance is

$$V(\hat{Y}_{2u}) = V_y/u \quad . \quad (11.33)$$

Weighting \hat{Y}_{2m} and \hat{Y}_{2u} by the inverse of their variances, we get

$$\hat{Y}_{2n} = \frac{V(\hat{Y}_{2u})\hat{Y}_{2m} + V(\hat{Y}_{2m})\hat{Y}_{2u}}{V(\hat{Y}_{2u}) + V(\hat{Y}_{2m})} \quad (11.34)$$

and its variance is approximately given by

$$V(\hat{Y}'_{2n}) = \frac{n-u\rho^2}{n^2-u^2\rho^2} V_y \quad (11.35)$$

It is of interest to note that the above variance reduces to V_y/n in the case of complete matching ($m = n$) and in the case of no matching ($m = 0$). The optimum matching proportion or repetition factor is obtained by equating to zero the partial derivative of the variance expression (11.35) with respect to (m/n) and solving for (m/n) . The optimum value of (m/n) is

$$\frac{m}{n} = \frac{\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} \quad (11.36)$$

and the minimum variance is given by

$$V(\hat{Y}_{2n}) = \{1 + \sqrt{1-\rho^2}\} \frac{V_y}{2n} \quad . \quad (11.37)$$

The expression (11.36) shows that if $\rho = 0$, $(m/n) = \frac{1}{2}$ and if $\rho = \pm 1$, $(m/n) = 0$. It means that the matching proportion should not exceed $\frac{1}{2}$.

11.7 ESTIMATION OF RELATIVE CHANGE

In the previous paragraphs we have seen how and under what circumstances the estimator of the second period can be improved with the help of the regression method of estimation. Another important problem in sampling on two occasions is to estimate the rate of change in the total value of the characteristic under study during the period considered. In this section the problem of estimating both the relative change and the average of a population parameter over time based on samples selected on successive occasions is briefly discussed. The results given in this section are based on the work of Parthasarathy (1961). The estimator \hat{R} of the rate of change and its approximate variance are given by

$$\hat{R} = (\hat{Y}_2 - \hat{Y}_1)/\hat{Y}_1 \quad \dots \quad (11.38)$$

and

$$\begin{aligned} V(\hat{R}) &= \{V(\hat{Y}_2) + (R+1)^2 V(\hat{Y}_1) - 2(R+1) \text{ Cov}(\hat{Y}_2, \hat{Y}_1)\}/Y_1^2 \\ &= \{V_2 + (1+R)^2 V_1 - 2(1+R) f V_{12}\}/n Y_1^2, \end{aligned} \quad \dots \quad (11.39)$$

where $f = m/n$ and it is assumed that $|(\hat{Y}_1 - Y_1)/Y_1| < 1$, and $V(\hat{Y}_1)$, $V(\hat{Y}_2)$ and $\text{Cov}(\hat{Y}_2, \hat{Y}_1)$ are of the form V_1/n , V_2/n and fV_{12}/n , V_1 and V_2 being the variance for a sample of one unit for the characteristic on the first and the second occasions and V_{12} being the covariance between the estimators of the characteristic on the two occasions based on one sample unit.

11.7a CHANGE IN A PROPORTION

If the rate of change in a population proportion P is to be estimated, then in the case of srswr the formulae (11.38) and (11.39) reduce to

$$\hat{R} = (\hat{P}_2 - \hat{P}_1)/\hat{P}_1 \quad \dots \quad (11.40)$$

and

$$V(\hat{R}) = [P_1 Q_2 + P_2 Q_1 - 2f(P - P_1 P_2)](P_2/n P_1^3), \quad \dots \quad (11.41)$$

where $f = m/n$ and P is the proportion of units having a particular characteristic on both the occasions. It is presumed that the total number of units in the two periods is the same.

The relative variance of \hat{R} is given by

$$\frac{V(\hat{R})}{R^2} = \frac{(P_2/P_1)}{n_1(P_2 - P_1)^2} [P_1Q_2 + P_2Q_1 - 2f(P - P_1P_2)] \quad (11.42)$$

If we know the approximate values of P_1 and P_2 , we can get upper and lower bounds of the relative variance of \hat{R} by using the inequality $0 \leq P \leq \min(P_1, P_2)$ for any particular f .

11.7b PROPORTION OF COMMON UNITS

The next problem is to determine that value of f which is suitable for estimating the average as well as the growth rate. When the matching proportion between the samples selected on two successive occasions is f , the variances of the estimators of rate of change

$\hat{R} = (\hat{P}_2 - \hat{P}_1)/\hat{P}_1$ and of the average

$$\hat{A} = (\hat{P}_1 + \hat{P}_2)/2 \quad (11.43)$$

are respectively given by (11.41) and

$$V(\hat{A}) = [P_1Q_1 + P_2Q_2 + 2f(P - P_1P_2)]/4n \quad (11.44)$$

When $P > P_1P_2$, we get the value of f , that minimizes the variances of estimators of the rate of change and of the average to be 1 and 0 respectively and when $P < P_1P_2$, this situation regarding optimum value of f gets reversed. One criterion for choosing f is that the relative efficiencies of the design with respect to the two estimators should be the same. The efficiency here is defined as the ratio of the optimum variance to the variance for any given f . The efficiencies of the design with a specified f for the estimation of rate of change and of the average are given by

$$\text{Eff}(\hat{R}) = \frac{V_0(\hat{R})}{V(\hat{R})} = \frac{P_1Q_2 + P_2Q_1 - 2(P - P_1P_2)}{P_1Q_2 + P_2Q_1 - 2f(P - P_1P_2)} \quad (11.45)$$

and

$$\text{Eff}(\hat{A}) = \frac{V_0(\hat{A})}{V(\hat{A})} = \frac{P_1Q_1 + P_2Q_2}{P_1Q_1 + P_2Q_2 + 2f(P - P_1P_2)} \dots \quad (11.46)$$

Equating (11.45) and (11.46), we get

$$f = \frac{P_1Q_1 + P_2Q_2}{(P_1 + P_2)(Q_1 + Q_2) - 2(P - P_1P_2)} \dots \quad (11.47)$$

Knowing the approximate values of P_1 and P_2 , we can get the upper and the lower bounds of f by using the inequality $0 \leq P \leq \min(P_1, P_2)$. Taking some arbitrary values for P_1 and P_2 , it is found that in some empirical situations the minimum value for f turns out to be about $\frac{1}{2}$.

11.8 MULTI-VARIABLE REGRESSION ESTIMATOR

When data on more than one auxiliary variable are available, it will be advantageous to build up a regression estimator which utilizes all the available information. In this connection, B. Ghosh (1947) has suggested the use of an estimator of the form

$$\hat{Y}_G = \hat{Y} + \sum_{i=1}^k \hat{\beta}_i (\bar{X}_i - \hat{X}_i), \quad \dots \quad (11.48)$$

where \hat{Y} is the usual unbiased estimator for the total of the study variable, \hat{X}_i is an unbiased estimator of the total (X_i) of the i -th auxiliary variable and $\hat{\beta}_i$ is an estimate of the regression coefficient of \hat{Y} on \hat{X}_i , $i = 1, 2, \dots, k$. It can be seen that (11.48) is similar to the arithmetic mean of the k regression estimators that can be obtained by using the k auxiliary variables.

Des Raj (1965) has proposed a weighted difference estimator of the form

$$\hat{Y}_D = \sum_{i=1}^k w_i \{ \hat{Y} + \lambda_i (\bar{X}_i - \hat{X}_i) \}, \quad \dots \quad (11.49)$$

where the weights $\{w_i\}$ add up to unity and λ_i 's are known constants. It has been suggested that approximate values of $\{Y/X_i\}$, $i = 1, 2, \dots, L$, derived from past surveys, may be taken as $\{\lambda_i\}$. Though this would be the appropriate procedure to follow when the lines of regression of \hat{Y} on $\{\hat{X}_i\}$ pass through the origin, this is likely to be a fairly good procedure even when the regression lines do not pass through the origin. As is to be expected, when $\lambda_i = Y/X_i$, the variance of \hat{Y}_D is exactly the same as the approximate variance of the multi-variable ratio estimator (10.79) considered in Chapter 10 (p. 392). The case of using two auxiliary variables to build up a multiple regression estimator is considered in Problem 11.14, (p. 424).

REFERENCES

- BOSE, C (1943) Note on the sampling error in the method of double sampling, *Sankhya*, 6, 329-330
- COCHRAN, W G (1942) Sampling theory when sampling units are of unequal sizes, *J Amer Stat Assn*, 37, 199-212
- DES RAJ (1965) On a method of using multi auxiliary information in sample surveys, *J Amer Stat Assn*, 60, 270-277
- GHOSH, B (1947) Double sampling with many auxiliary variates, *Bull Cal Stat Assn*, 1, 91-93
- HANSEN, M H, HURWITZ W N, NISSLERSON, H and STEINBERG, J (1955) The redesign of the census current population survey, *J Amer Stat Assn*, 50, 701-719
- KULLDORF, G (1963) Some problems of optimum allocation for sampling on two occasions, *Rev Inter Stat Inst* 31, (1), 24-57
- MAHALANOBIS, P C (1941) *Report on the Sampling Techniques for Forecasting the Bark yield of Cinchona Plants* Experiments Series B, Indian Statistical Institute
- MICKEY, M R (1959) Some finite population unbiased ratio and regression estimators *J Amer Stat Assn*, 54, 594-612
- MURTHY, M N (1962) Almost unbiased estimators based on interpenetrating sub samples, *Sankhya*, 24, (A), 303-314
- PARTHASARATHY, G (1961) Sampling on successive occasions and self weighting design *Thesis submitted for M Stat degree of the Indian Statistical Institute*
- PATTERSON, H D (1950) Sampling on successive occasions with partial replacement of units *J Roy Stat Soc, (B)*, 12, 241-255
- RAO, J N K and GRAHAM, J E (1964) Rotation designs for sampling on repeated occasions, *J Amer Stat Assn*, 59, 492-509,

- SEAL, K. C. (1951) : On errors of estimates in various types of double sampling procedures; *Sankhyā*, 11, 125-144.
- TIKKIWAL, B. D. (1960) : On the theory of classical regression and double sampling estimation; *J. Roy. Stat. Soc., (B)*, 22, 131-138.
- WILLIAMS, W. H. (1961) : Generating unbiased ratio and regression estimators; *Biometrics*, 17, 267-274.
- ZARKOVICH, S. S. (1956) : An illustration of some characteristic situations in the application of the difference estimate; *Rev. Inter. Stat. Inst.*, 24, 52-63.

COMPLEMENTS AND PROBLEMS

11.1 For the hypothetical population of six units given in Table 10.4(p.398) obtain the rse of the regression estimator of the population total Y based on a sample of two units selected with srs wr by enumerating all possible samples and compare the efficiency of the regression estimator with those of the conventional unbiased estimator and of the ratio estimator.

11.2 It is proposed to use two-phase sampling for estimating the total volume of timber (Y) in a forest area. A sample of n_1 plots of 1/10th acre and a sub-sample of n_2 plots from this sample are selected with srs wr for obtaining eye-estimates (x) and actual measurements (y) of the volume of timber respectively. Considering the regression estimator of Y got by regarding x as the supplementary variable, obtain the approximate variance and determine the optimum values of n and n' when the cost function is given by $C = 1000 + 5n_1 + 40n_2$, C being fixed at Rs. 10000, assuming the population coefficient of variation for y to be 100% and the correlation coefficient between x and y to be 0.7. Find also the minimum value of the rse.

11.3 For estimating the total cattle population in a given area, a sample of 24 villages was selected from the 1238 villages in that region with srs wr. The number of cattle obtained in the survey is given for each sample village in Table 11.1 together with the corresponding census figure relating to a previous period. Using this information, compare the efficiency of regression estimator with that of ratio estimator.

11.4 Using the data given in Table 10.8 (p.400), compare the efficiency of the regression estimator with those of the ratio estimator and the usual unbiased estimator.

11.5 When a sample is selected with srs wr from a finite population of N units, which itself is assumed to be a sample from an infinite population following the model

$$E(Y_i | X_i) = \alpha + \beta X_i, \quad V(Y_i | X_i) = a X_i^g, \quad \text{Cov}(Y_i, Y_{i'} | X_i, X_{i'}) = 0, \quad i' \neq i,$$

where a and g are positive constants, show that the regression estimator of the population mean \bar{Y} is always more efficient than the ratio estimator when x is used as the supplementary variable. (*Hint* : Obtain the expected variances in the two cases and compare them.)

(Des Raj, *J. Amer. Stat. Assn.*, 53, (1958), 98-101).

TABLE II.1 NUMBER OF CATTLE FOR 24 SAMPLE VILLAGES

sample village	number of cattle		sample village	number of cattle		sample village	number of cattle	
	census	survey		census	survey		census	survey
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	623	654	9	161	210	17	330	375
2	690	696	10	298	555	18	218	212
3	534	530	11	2045	2110	19	160	147
4	293	315	12	1069	592	20	210	297
5	69	78	13	706	707	21	262	401
6	842	640	14	1795	1890	22	204	252
7	475	692	15	1406	1123	23	185	199
8	371	292	16	118	115	24	574	564

11.6 Show that if the proportional increase in the variance of the regression estimator (11.5) arising due to the use of a constant λ instead of β is to be less than α then the relative deviation of λ from β , that is $|(\lambda - \beta)/\beta|$ should be less than $\sqrt{\alpha(1-\rho^2)/\rho^2}$.

(Cochran, W. G., *Sampling Techniques*, (1963), p. 192)

11.7 Derive the expressions given in (11.25) to (11.30) relating to the optimum values of first phase and second phase samples and the minimum values of variance and cost

11.8 To estimate the total yield of guava in a district, a two phase sampling design was adopted. First a sample of n_1 villages was selected with ppsswr size being area under orchards in the previous year, for obtaining a the current total number of orchards and b the current number of guava trees. From each of a sub sample n_2 villages (e.g. the first n_2 villages of the initial sample of n_1 villages), m orchards were selected with srs wr for observing the yield of the guava crop. Let y_{ij} denote the yield of the j th sample orchard in the i th sample village, $j = 1, 2, \dots, m$, $i = 1, 2, \dots, n_2$. Suggest a suitable regression estimator for the total yield of guava fruit in that district. Derive its approximate sampling variance stating clearly the assumptions involved and give a suitable consistent estimator of the variance.

11.9 In two phase sampling where n_1 units are selected in the first phase with srswr for getting the value of x , the auxiliary variable, and in the second phase a sub sample of n_2 units is selected with srswr for obtaining the value of y the study variable, find the condition for the regression estimator of \bar{Y} obtained by using the optimum

values of n_1 and n_2 to be more efficient than the usual unbiased estimator of \bar{Y} based on n_2 units selected with srsrwr directly from the population, when the cost is fixed, the cost function being of the form $C = n_1C_1 + n_2C_2$, where C_1 and C_2 are the costs of obtaining the values of x and y respectively.

(Hansen, M. H., Hurwitz, W. N. and Madow, W. G., *Sample Survey Methods and Theory*, Vol. II, (1953), p. 256).

11.10 Derive the variance expression given in (11.35) in connection with double sampling and also obtain the optimum value of the proportion of common units given in (11.36) and the minimum variance (11.37).

11.11 Suppose a large sample of n units is drawn from a finite population of N units with srsrwr in the first phase to obtain data on the supplementary variable x and then in the second phase, a sub-sample of n_1 units is drawn from the first phase sample with srsrwr and a direct sample of n_2 units is selected independently from the whole population with replacement for obtaining information on the study variable y . Considering two estimators of the population mean \bar{Y} namely, (i) regression estimator based on the two-phase sample of (n, n_1) units and (ii) an unbiased estimator based on the direct sample of n_2 units, obtain the best linear estimator and derive its approximate sampling variance.

(Patterson, H. D., *J. Roy. Stat. Soc., (B)*, 12, (1950), 241–255).

11.12 In sampling on two occasions, n units are selected with srsrwr on the first occasion. On the second occasion a sub-sample of $n\alpha$ of these units, selected with srsrwr, is retained for the second occasion and it is supplemented by a sample of $n(1-\alpha)$ units directly selected with srsrwr. Suppose it is proposed to estimate the population mean on the second occasion \bar{Y}_2 by using an unbiased linear estimator of the form

$$\lambda_1\bar{y}_1 + \lambda_2\bar{y}_2 + \lambda_3\bar{y}'_1 + \lambda_4\bar{y}'_2,$$

where \bar{y}_i and \bar{y}'_i , $i = 1, 2$, are the sample means for the i -th occasion based on the common $n\alpha$ units and on the $n(1-\alpha)$ freshly selected units respectively and λ 's are constants. Determine the optimum values of λ 's and α , which would minimize the sampling variance of the estimator, and find the expression for the minimum variance. Also verify that the estimator of \bar{Y}_2 is similar to that derived in Problem 11.11.

11.13 In problem 11.12, if it is intended to estimate (i) the sum of the means and (ii) the difference in the means on the two successive occasions using estimators of the form given therein, determine the values of λ 's, which would minimize the corresponding sampling variances. Also obtain the expressions for the minimum variance in the two cases when the population variance σ^2 remains the same for the two occasions.

Problems 11.12 and 11.13—Hansen, M. H., Hurwitz, W. N. and Madow, W. G., *Sample Survey Methods and Theory*, Vol. II, (1953), pp. 269–270.

11.14 In the case of using two auxiliary variables x_1 and x_2 to build up a multi-variable regression estimator (11.49) for estimating the population mean \bar{Y} of the study variable on the basis of a sample selected with srs w/o, show that the optimum weights are given by $w_1 = w_2 = \frac{1}{2}$, when the coefficients of variation of x_1 and x_2 are equal to C , the correlation coefficients between y and x_i , $i = 1, 2$, are the same ρ_0 and λ_0 is taken as \bar{Y}/\bar{X}_i , $i = 1, 2$. Derive the variance of the multi-variable regression estimator in this case and show that this estimator would be more efficient than the usual unbiased estimator \bar{y} , if $\rho_0 > \frac{1}{2}(1+\rho)/(C/C_0)$, where ρ is the correlation coefficient between x_1 and x_2 and C_0 is the coefficient of variation of y .

(Deo Raj, *J. Amer. Stat. Assn.*, 60, (1965), 270-277).

Self-weighting Design

12.1 NEED FOR EQUAL WEIGHTS

So far we have discussed many sampling schemes and estimation procedures without emphasizing the need for simplicity in the computation of the various estimates at the tabulation stage. In this chapter we shall first consider the usual estimators and then describe some of the techniques commonly used to simplify the calculation of estimates, thereby reducing the work-load at the tabulation stage. Since in a sample survey we are estimating the value of a characteristic for the whole population on the basis of data on a part of it selected as sample, the sample observations are weighed with certain *weights* for obtaining an estimate of the population parameter. The estimators commonly used for estimating the population total of a variable are of the form

$$\hat{Y} = \sum_{i=1}^n w_i y_i, \quad \dots \quad (12.1)$$

where n is the total number of ultimate sample units and y_i and w_i are respectively the value of the variable y and the corresponding weight for the i -th sample unit. The weights $\{w_i\}$ depend on the selection and estimation procedures and they are usually chosen such that the estimator becomes unbiased.

The weights $\{w_i\}$ in (12.1) are known as *multipliers*, *inflation factors*, or *raising factors*, since they are used to inflate or raise the sample observations to get an estimate of the population total. For

instance, in srswr, srs wor and circular systematic sampling, the weight is common for all the sample units and is given by

$$w_t = N/n, \quad (12.2)$$

which is the inverse of the sampling fraction, and in linear systematic sampling the weight is L , the sampling interval used, instead of (N/n) . But in ppswr sampling, the weight generally differs from unit to unit, since for the t th selected unit it is given by

$$w_t = \frac{1}{n} \cdot \frac{1}{p_t}, \quad (12.3)$$

where p_t is the probability of selection of that sample unit at each draw. In a two stage design, where n fsu's are selected with ppswr (or with pps circular systematically), and m_t ssu's are selected from M_t ssu's with srswr or srs wor or circular systematically, the weight is of the form

$$w_t = \frac{1}{n} \cdot \frac{1}{p_t} \cdot \frac{M_t}{m_t} \quad (12.4)$$

Since the weights do not depend on the individual sample observations, they are calculated first and then these weights are applied to multiply the sample values of the various characteristics under study to obtain the estimates. In large scale surveys, where a number of parameters are to be estimated, the calculation at estimation stage becomes heavy, time consuming and costly owing to the need for weighting the sample observations for each characteristic by the weights. Hence, from the consideration of ease, speed and economy at the tabulation stage, it is desirable to have a sampling design which gives rise to a single common weight for all the sample units. Such a sample design is known as *self weighting design*, since the design itself takes care of the problem of weighting the sample observations eliminating the need for calculation of weights for each

of the sample units. Such a design may also be termed *equi-weighting design*. In this case, estimator (12.1) reduces to

$$\hat{Y} = w \sum_{i=1}^n y_i, \quad \dots \quad (12.5)$$

where w is the common weight for all the sample units.

It is possible to make a design self-weighting by suitably specifying the sampling scheme. That is, the selection of the units is so done as to make all the w_i 's equal to one another. Such a design is termed *self-weighting design at field stage*. For instance, a multi-stage design can be made self-weighting by appropriately specifying the number of ultimate stage units to be selected from the selected penultimate stage units. There may be some practical situations, where the design cannot be easily made self-weighting at field stage due to some technical or operational difficulties. In such cases, certain techniques are available to reduce the number of weights to be used at the tabulation stage. A design, where self-weighting is achieved by the adoption of some device at the tabulation stage, is termed *self-weighting design at tabulation stage*. Usually it is desirable to make the design self-weighting at the field stage itself and self-weighting at the tabulation stage is to be resorted to only if it is not possible to do so.

In this chapter both the situations of making a design self-weighting at the field and at the tabulation stages are considered. Though it is ideal to have only one common weight for all the sample units, the saving in time and cost at the tabulation stage is likely to be substantial even if two or more common weights are used, provided the number of such common weights is fairly small. In the latter case the design may be said to be *partially self-weighting*. This point arises because in practice such situations, where a small number of common weights is to be used instead of one such, are likely to occur often owing to the restrictions usually imposed on the design by technical and operational considerations.

The question of making the design self-weighting at the field stage has been discussed among others, by Hansen, Hurwitz and

W G Madow (1953) Lahiri (1954) and Som (1959) and the procedure of making the design self weighting at the tabulation stage has been considered by Murthy and Sethi (1959 1961) Sukhatme and Panse (1951) and Church (1954) have considered the implications of using unweighted estimates in case of non self weighting designs In the next few sections the technique of making a design self weighting at the field stage is briefly discussed for some of the commonly used sampling designs and the procedures of making a non self weighting design self weighting at the tabulation stage are considered in Sections 12.5 and 12.6 The former aspect is also discussed in Section 15.31 of Chapter 15

12.2 STRATIFIED UNI-STAGE SAMPLING

In this section the question of making the design self weighting in cases of stratified sampling with srs and pps sampling within the strata is discussed

12.2a SRS WITHIN STRATA

In stratified sampling with srswr srs wr and circular systematic sampling the design is not self weighting as such for the estimate of Y is given by

$$\hat{Y} = \sum_{s=1}^S \frac{N_s}{n_s} \sum_{i=1}^{n_s} y_{si} \quad (12.6)$$

where N_s and n_s are respectively the total number of units and sample size in the s th stratum and y_{si} is the value of the i th selected unit in that stratum Here the multiplier N_s/n_s differs from stratum to stratum unless the allocations to the strata are proportional to N_s Thus we see that the design becomes self weighting for proportional allocation in which case

$$(N_s/n_s) = (N/n) \quad (12.7)$$

which is the inverse of the overall sampling fraction This shows that the use of self weighting design imposes a restriction on the

allocation to the strata, since we have to have proportional allocation even if we have knowledge about the strata variances, as use of optimum allocation makes the design non-self-weighting. In such a case, a decision regarding the use of proportional or optimum allocation is to be arrived at after considering the relative magnitude of the gain in tabulation cost by having a self-weighting design against that of the loss in precision of the estimator due to using proportional allocation instead of optimum allocation.

In proportional allocation, the number of units to be selected from the s -th stratum, namely $n(N_s/N)$, may not be an integer and rounding off n_s to the nearest integer would make the design non-self-weighting. A convenient way of avoiding this difficulty is to select the units in each of the strata linear systematically with the same sampling interval (N/n) and random starts from 1 to (N/n) . Here also there may be rounding off difficulty if N/n is not an integer. One possibility is to take the interval as the integer nearest to N/n , but this is likely to lead to slightly more (or less) sample size in each stratum. Another possibility is to select the sample systematically with the fractional interval N/n (cf. Section 5.3 of Chapter 5, p. 141). A third procedure is to make the interval in each stratum take the values $[N/n]$ or $[N/n]+1$ such that the expected value of a random variable taking the values $(N/n)/[N/n]$ and $(N/n)/\{[N/n]+1\}$ is unity, (cf. Problem 12.7, p. 447).

In large-scale surveys it is sometimes operationally convenient to have equal work-load in the different strata. It is of interest to note that even in the cases, where the use of self-weighting design makes the work-load in the strata unequal, it is possible to equalize the work-load by appropriately changing the sampling intervals to be used in them to multiples or sub-multiples of the common interval. But when this procedure is adopted, the design can only be made partially self-weighting and not fully self-weighting, since the inflation factors to be used in some strata would be multiples or sub-multiples of the common inflation factor. This situation has been illustrated with an example in Section 12.4.

12.2b PPS SAMPLING WITHIN STRATA

In a stratified design where the units in the strata are selected with ppswr, size being the value of x , the estimator of Y is given by

$$\hat{Y} = \sum_{s=1}^K \frac{X_s}{n_s} \sum_{i=1}^{n_s} \frac{y_{si}}{x_{si}}, \quad (12.8)$$

where y_{si} and x_{si} are the value of the study variable y and of the auxiliary variable respectively for the i th sample unit in the s th stratum and X_s and n_s are the total size and the sample size in the s th stratum. If the ratio y_{si}/x_{si} (say, r_{si}) can be readily observed in the field or can be reported by the investigator without much difficulty, the design would become self weighting if the allocation is done in proportion to the total size of the s th stratum, that is, if $n_s = nX_s/X$. In this case, the estimator in (12.8) reduces to

$$\hat{Y} = \frac{X}{n} \sum_{s=1}^K \sum_{i=1}^{n_s} r_{si}, \quad (12.9)$$

where X is the total size and n is the overall sample size.

A practical example of the use of the estimator in (12.9) is provided by a crop survey where y and x may stand for area under a particular crop and geographical area respectively, in which case the ratio y/x is the proportion of the area under the crop in the sample plot or field. In this case if the allocation to the strata is done in proportion to their total geographical area and if the plots or fields in the strata are selected with ppswr, size being geographical area, then the estimator is of the form given in (12.9), (cf Sub section 7.8b of Chapter 7, p 257). Similarly in a socio economic survey, where the sample unit is a household, y and x may stand for expenditure on a given item and household size respectively. If the households are selected with ppswr, size being x , and if the per capita expenditure for each sample household is reported, then the sample design becomes self weighting for estimating the per capita expenditure or total expenditure in the population.

12.3 STRATIFIED TWO-STAGE SAMPLING

In a stratified two-stage design, where the fsu's are selected with ppswr (or systematically), size being x , and the ssu's are selected linear systematically, an unbiased estimator of Y is given by

$$\hat{Y} = \sum_{s=1}^K \frac{X_s}{n_s} \sum_{i=1}^{n_s} \frac{I_{si}}{x_{si}} \sum_{j=1}^{m_{si}} y_{sij}, \quad \dots \quad (12.10)$$

where y_{sij} is the value of y for the j -th sample ssu in the i -th selected fsu of the s -th stratum, x_{si} , I_{si} and m_{si} are the size, the sampling interval and the number of sample ssu's for the i -th sample fsu in the s -th stratum, and X_s and n_s are the total size and the sample size for the s -th stratum. The weight in this case is

$$w_{sij} = \frac{X_s}{n_s} \cdot \frac{I_{si}}{x_{si}}. \quad \dots \quad (12.11)$$

To make the design self-weighting, we have to fix I_{si} or m_{si} such that the weight is a constant λ . For any given constant λ the design becomes self-weighting, if

$$I_{si} = \lambda n_s x_{si}/X_s. \quad \dots \quad (12.12)$$

Noting that the number of sample ssu's in the i -th sample fsu is given by

$$m_{si} = \frac{M_{si}}{I_{si}} = \frac{1}{\lambda} \cdot \frac{X_s}{n_s} \cdot \frac{M_{si}}{x_{si}}, \quad \dots \quad (12.13)$$

we see that if λ is made large the sample size in terms of the number of ssu's would be reduced since m_{si} would be small, and conversely if λ is taken to be small then the number of ssu's in the sample would be larger than the requirement. Hence, λ is to be so determined as to get the required sample size in terms of ssu's. If nm sample ssu's are required on the average then we get

$$E \left\{ \sum_{s=1}^K \sum_{i=1}^{n_s} m_{si} \right\} = \frac{1}{\lambda} \sum_{s=1}^K \sum_{i=1}^{N_s} M_{si} = nm \quad \dots \quad (12.14)$$

and hence

$$\lambda = \frac{1}{nm} \sum_{s=1}^K \sum_{i=1}^{N_s} M_{si} \quad (12.15)$$

Thus we see that the constant λ should be taken as the inverse of the overall sampling fraction for the ssu's. If the values of $\{M_{si}\}$ for all the fsu's are not available at the stage of fixing the constant λ an idea of the total number of ssu's obtained on the basis of a previous survey may be used in determining the constant λ . The sample size actually obtained would depend on the accuracy of the estimate of the total number of ssu's used in the calculation of λ . Once λ is determined the design can be made self weighting by fixing the sampling interval I_{st} as specified in (12.12). If I_{st} is not an integer it may be suitably rounded off as mentioned in Section 12.2a. It may be noted that for this procedure it is not necessary to know the values of $\{M_{si}\}$ in advance even for the selected fsu's provided λ can be determined on the basis of a previous census or survey.

If in a stratified two stage design the fsu's are selected with pps size being the number of ssu's and the ssu's are selected with srs or systematically the design can be made self weighting with equal work load in the selected fsu's by allocating the sample size in terms of fsu's to the strata in proportion to the total stratum size and selecting the same number of ssu's from each selected fsu (cf Problem 12.5 p 446). With this type of design the work load in a stratum can be equalized by forming strata with approximately the same total size. The above discussion shows that it is possible to make a given design self weighting by suitably determining the sampling interval to be used in the selected penultimate stage units and to ensure at the same time the required sample size. Here also two or more common weights may be used if found necessary from field operational considerations. An example to illustrate this situation is given in Section 12.4.

12.4 SOME EXAMPLES

In this section two examples are given to illustrate the procedure of making the design self-weighting and of modifying this design with a view to equalizing the work-load within strata and within the sample fsu's, when the self-weighting design leads to unequal work-load. The first example relates to selection of 20 factories from a population of 160 factories divided into 6 regional strata on the basis of their location. In this case, the design can be made self-weighting by selecting sample factories from each stratum linear systematically with 8 as the sampling interval. This procedure may lead to variation in work-load over the strata. But this situation can be avoided by taking multiples or sub-multiples of the sampling interval in those strata, where the work-load is considerably more (or less) than the average work-load of 3 to 4 sample factories. However, this modified design ceases to be completely self-weighting, since the weights in some strata get changed due to changes in the sampling interval, thereby giving rise to a partially self-weighting design. The relevant values of the sampling interval and the expected sample size for the different strata in the two cases under consideration are given in Table 12.1.

TABLE 12.1. STRATUM-WISE SAMPLING INTERVAL AND EXPECTED SAMPLE SIZE FOR TWO DESIGNS.

sr. no. of stratum	no. of factories	design 1		design 2	
		I_s	$E(n_s)$	I_s	$E(n_s)$
(1)	(2)	(3)	(4)	(5)	(6)
1.	52	8	6.500	16	3.250
2.	14	8	1.750	4	3.500
3.	25	8	3.125	8	3.125
4.	28	8	3.500	8	3.500
5.	12	8	1.500	4	3.000
6.	29	8	3.625	8	3.625

design 1 : self-weighting with unequal work-load; design 2 : partially self-weighting with equal work-load.

From Table 12 1, it is seen that in the case of self weighting design with one common multiplier, the expected number of sample factories in each stratum is proportional to the total number of factories in that stratum and that it varies from 1 to 7. In the case of a partially self weighting design, where the sampling intervals in three strata have been changed we find that the number of sample factories in the strata varies only from 3 to 4. The former design with proportional allocation is expected to be more efficient from the point of view of sampling variance than the latter, where the allocation is approximately equal though the latter may have to be adopted if there is considerable advantage in making the work load in the strata approximately equal.

The second example illustrates the procedure of making a stratified two stage design self weighting for a demographic survey, taking villages as the fsu's and households as the ssu's. Considering a hypothetical population of 16 villages divided into four strata let us suppose that from each stratum one village is to be selected with pps and that on the average a total sample of 80 households is to be selected systematically from all the selected villages. The data on population and number of households for all the 16 villages are given in Table 12 2. Noting that the total number of households is 7198 and that the total number of sample households required is 80 on the average we determine the constant inflation factor λ as $(7198/80)$, which is 90 approximately. For each of the 16 villages, the interval to be used for sampling households if it gets selected, can be worked out on the basis of equation (12 12) taking n_s to be 1. The values of sampling interval together with the expected number of sample households have been worked out for each village and these are also shown in Table 12 2.

From Table 12 2, we find that the work load in terms of sample households is approximately the same for all the villages in a stratum, though this work load varies over the strata due to the non proportional allocation adopted. However, the work load can be approximately equalized, as shown in columns (8) and (9) of Table 12 2,

by making the sampling intervals in the villages of stratum 3 half the values given in column (5) and doubling the intervals in stratum 2. This would, of course, lead to three multipliers, thereby making the design only partially self-weighting. The question of fractional intervals can be dealt with, as mentioned in Sub-section 12.2a, p.429

TABLE 12.2. SELF-WEIGHTING DESIGN IN THE CASE OF STRATIFIED TWO-STAGE SAMPLING.

sr. no. of stratum	sr. no. of village	no. of persons	no. of house- holds	design 1			design 2		
				I_{st}	m_{st}	λ	I_{st}	m_{st}	λ
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	1	1618	360	21.14	17.0	90	21.14	17.0	90
	2	1402	255	18.31	13.9	90	18.31	13.9	90
	3	699	140	9.14	15.3	90	9.14	15.3	90
	4	3168	704	41.47	17.0	90	41.47	17.0	90
2	1	4006	728	21.99	33.1	90	43.98	16.6	180
	2	7917	1588	43.44	36.6	90	86.88	18.3	180
	3	2122	472	11.65	40.5	90	23.30	20.2	180
	4	2355	428	12.92	33.1	90	25.84	16.6	180
3	1	490	98	12.67	7.7	90	6.34	15.4	45
	2	533	118	13.79	8.6	90	6.90	15.2	45
	3	284	52	7.34	7.1	90	3.67	14.2	45
	4	2172	434	56.20	7.7	90	28.10	15.4	45
4	1	684	155	6.88	22.5	90	6.88	22.5	90
	2	2656	483	26.75	18.1	90	26.75	18.1	90
	3	2726	545	27.45	19.9	90	27.45	19.9	90
	4	2872	638	28.92	22.1	90	28.92	22.1	90

(total population : 35704, total number of households : 7198).

design 1 : self-weighting with unequal work-load; design 2 : partially self-weighting with equal work-load; I_{st} : sampling interval; m_{st} : expected sample size.

12.5 SELF-WEIGHTING AT TABULATION STAGE

One of the difficulties in making the sample design self weighting at the field stage is that the work load in the penultimate stage units may become unequal, which may not be desirable from the point of view of administrative considerations in large scale surveys. The methods which make the sample design self weighting at the field stage ensuring at the same time equal work load within penultimate stage units impose rather severe restrictions on the allocation and selection procedures. Because of these considerations it may be neither desirable nor feasible to adopt a self weighting design at the field stage in certain situations. In such cases it may be advisable to make the design self weighting at the tabulation stage. The following five procedures of making the design self weighting at the tabulation stage are briefly considered here.

- (i) rounding off of the weights to the nearest multiple of a convenient number such as hundred, thousand, etc.,
- (ii) substituting each weight by the mean of the weights,
- (iii) random rounding off of the weights to an optimum set of weights,
- (iv) sub sampling with ppswr, size being the weights, and
- (v) sub sampling with pps systematically size being the weights.

Procedures (i) and (ii) give rise to biased estimates with possible decrease in the variance of the estimate under certain circumstances, while procedures (iii), (iv) and (v) provide unbiased estimates with some increases in the variance.

12.5a ROUNDING OFF TO COMMON WEIGHTS

Suppose we have a series of four digit weights. The usual procedure of decreasing the number of weights is to round off the weights to the nearest thousand. Thus the n four digit weights are replaced by at most ten rounded off weights. The estimate obtained by using this method will generally be biased. The magnitude of

bias depends on the weights as well as on the values of the characteristics. In practice, it is difficult to get an idea of the sign and magnitude of the bias unless one works out the biased estimate as well as the estimate using the actual weights for at least some selected characteristics.

12.5b ROUNDING OFF TO AVERAGE WEIGHTS

In this method the common weight is taken as the average of the weights, and in this case the estimate is given by the product of the mean of the weights and the sum of the sample values for the characteristic. This estimate is also biased, but the bias will be negligible if in the sample the covariance between the weights and the values of the sample units is very small. For instance, in a particular per capita expenditure class, it is felt that the expenditure on cereals and food, and the total expenditure of a household might not depend much on its weight. If so, the sample households can be classified according to the per capita expenditure class and then the suggested biased estimate can be found separately for each of the classes. The sum of these biased estimates will be the estimate of the population total and the bias can be expected to be small. This method can be used only if we are fairly certain that the values of the study variable are uncorrelated with the weights.

12.5c RANDOMIZED ROUNDED-OFF WEIGHTS

The general solution would be to round off each of the weights to a certain number of weights which may be called *rounded-off weights* with such probabilities that the expected value is the original weight. As these weights are at our choice, we can choose them such that the increase in variance is minimized. Further, a prespecified level of increase in the variance at the tabulation stage can be achieved by taking a sufficient number of rounded-off weights in the optimum fashion. As the optimum solution for a specified number of rounded-off weights depends on the values of the characteristics in question, it is not feasible in practice to get the optimum

solution for each characteristic, though some method which would give us a solution near about the optimum can be devised. Another disadvantage is that even the determination of the approximate optimum rounded off weights becomes increasingly difficult as the number of such weights is sought to be increased. This procedure is considered in detail by Murthy and Sethi (1961) (cf Problem 12.8 p 448).

12.5d SUB-SAMPLING WITH PPSWR

This method consists in taking a sub sample of n' units from the field sample of size n with ppswr, size being the original weights. The sum of the values in the sub sample multiplied by a constant (the ratio of the sum of n weights to n') gives an unbiased estimator of $\sum_{i=1}^n w_i y_i$. The sub sample size can be so fixed as to avoid rejections of the sample units provided such a procedure does not lead to a large number of repetitions of many sample units.

The increase in variance at the tabulation stage is given by

$$E_f \left[\frac{1}{n'} \left\{ \left(\sum_{i=1}^n w_i \right) \left(\sum_{i=1}^n w_i y_i^2 \right) - \left(\sum_{i=1}^n w_i y_i \right)^2 \right\} \right], \quad (12.16)$$

where E_f is the expected value over the field sample. It can be seen that this method results in rounding off each weight to one of the weights $\left\{ \frac{j}{n} \sum_{i=1}^n w_i \right\}$, ($j = 0, 1, 2, \dots, n$) with certain probabilities such that the estimate remains unbiased and that the sum of the rounded off weights in the sub sample is equal to the sum of the initial weights.

12.5e SUB-SAMPLING PPS SYSTEMATICALLY

In this method the units in the field sample are arranged in a suitable order, the weights are cumulated and a systematic sample of n' units is drawn with pps, size being the weights, using the interval $I = w/n'$, where w is the sum of w_i 's. The estimator is given by the product of the sum of the values in the sub sample and the

sub-sampling interval. An expression for the increase in variance is difficult to find in this case, (cf. Sub-section 6.11c of Chapter 6, p.215). This method results in rounding off the weight w_i to $[w_i/I]I$ or $\{[w_i/I]+1\}I$ with certain probabilities such that the estimate remains unbiased and that the sum of the rounded-off multipliers in the sub-sample is equal to the sum of weights in the field sample.

A sort of *without replacement* element is present in the pps systematic selection. For, when the units are arranged at random and the sizes are equal, this procedure amounts to srs wor while in this case ppswr sampling amounts to srswr. Further, drawing of a sub-sample pps systematically is likely to take less time than in ppswr sampling. The estimate in the case of pps systematic sampling can be improved upon by arranging the units in a suitable order and by devising appropriate balancing procedures.

12.6 AN EMPIRICAL STUDY

As it is not possible to compare the efficiencies of the procedures suggested in Section 12.5 theoretically, an empirical study is conducted to assess their merits and demerits. For this purpose the data on consumer expenditure statistics collected in a large-scale sample survey in the State of Uttar Pradesh are used. The object of this study is to compare the efficiencies of the procedures (i), (ii), (iv) and (v) as well as their practicability in large-scale operations for estimating (a) expenditure on cereals, (b) expenditure on food and (c) total expenditure.

12.6a SAMPLING DESIGN OF THE SURVEY

The design of the survey was a stratified three-stage one with tehsils as first stage units, villages as second stage units and households as third stage units. From each stratum two tehsils were selected with ppswr, size being the previous census population (p). From each selected tehsil two villages were drawn with ppswr, size being p , and from each sample village five households were selected systematically with a random start for the consumer expenditure

enquiry It is to be noted that for each stratum total we get two independent estimates, one from each of the two tehsils selected from that stratum The sample households belonging to the tehsils selected first are considered as belonging to field sample 1 and the other sample households constitute field sample 2

12.6b SELF-WEIGHTING PROCEDURES

The sample households belonging to each of the field samples are grouped into fourteen classes on the basis of their per capita total expenditure In each per capita expenditure class the households are arranged according to the village tehsil and stratum to which they belong For procedures (iv) and (v), from each per capita expenditure class a sample of size equal to the number of sample households in that class is selected to estimate the expenditure on cereals and food and the total expenditure This sample size was taken in each class since in that case the efficiencies of procedures (iv) and (v) become comparable with those of procedures (i) and (ii) The variances of the estimates obtained by using procedures (iv) and (v) are calculated assuming the field sample to be the population These variances can be considered as the estimates of the increase in the variances of the estimates due to sub sampling The bias of the estimates based on procedures (i) and (ii) are also calculated

Let w_{ij} and y_{ij} be the weight and the value of y for the j th sample household in the i th per capita expenditure class The increase in the variance of the estimate based on a sub sample taken with ppswt, size being the weight, is estimated by

$$v_{pps} = \sum_{i=1}^{14} \left[\frac{1}{n_i} \left\{ \left(\sum_{j=1}^{n_i} w_{ij} \right) \left(\sum_{j=1}^{n_i} w_{ij} y_{ij}^2 \right) - \left(\sum_{j=1}^{n_i} w_{ij} y_{ij} \right)^2 \right\} \right] \quad (12.17)$$

The variance of the estimate for pps systematic sampling is obtained by enumerating all the possible samples The estimates in the case of procedures (i) and (ii) are respectively

$$\hat{Y}' = \sum_{i=1}^{14} \sum_{j=1}^{n_i} r_{ij} y_{ij} \quad (12.18)$$

and

$$\hat{Y}'' = \sum_{i=1}^{14} \frac{1}{n_i} \left(\sum_{j=1}^{n_i} w_{ij} \right) \left(\sum_{j=1}^{n_i} y_{ij} \right), \quad \dots \quad (12.19)$$

where r_{ij} is the nearest multiple of 10000 of the weight w_{ij} . It is expected that for procedure (ii) the bias in the estimate of the population totals considered here is likely to be small if the sample households are stratified according to per capita expenditure class at the tabulation stage. The differences between these estimates and $\sum_{i=1}^{14} \sum_{j=1}^{n_i} w_{ij} y_{ij}$ estimate the biases.

12.6c RESULTS OF THE EMPIRICAL STUDY

The results of the empirical study are given in Table 12.3. From this table we find that the pps systematic estimate is much better than ppswr estimate in all the cases. It is possible to improve upon the pps systematic estimate by suitable arrangement of the sample units before sub-sampling. Of course, in pps systematic sampling, the variance of the estimate cannot be estimated unbiasedly from one sample. However, this difficulty can be overcome by taking two or more sub-samples with independent random starts.

The comparison of procedures (i) and (ii) with procedure (iv) shows that the biased procedures may be preferred. Procedures (i) and (ii) should be used with considerable caution as they give rise to biased estimates and it is difficult to assess the magnitude of bias in practice. Procedure (ii) will be a good method, as has been pointed out earlier, only if the correlation coefficient between the weights and the values of the study variable is nearly equal to zero. Since the pps systematic method (v) is more efficient than (iv) and it has the added advantage of being unbiased, it may be preferred to (i), (ii) and (iv).

TABLE 12.3 RELATIVE BIASES OF PROCEDURES (i) AND (ii) AND
RELATIVE STANDARD ERRORS OF PROCEDURES (iv) AND (v)

proce dure	relative bias or rse (%)	cereal expenditure		food expenditure		total expenditure	
		sample 1	sample 2	sample 1	sample 2	sample 1	sample 2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(i)	bias	+1.47	+0.69	+1.51	+1.63	+1.04	+2.36
(ii)	bias	-3.36	-2.36	-2.96	0	-1.63	+0.13
(iv)	rse	3.10	5.79	3.07	4.72	3.61	4.89
(v)	rse	1.03	1.94	1.16	1.66	1.62	2.34

(number of sample households—sample 1 358 sample 2 341)

12.7 REPETITION OF SAMPLE OBSERVATIONS

We have seen in the previous sections that it is possible to make the sample design at least partially self weighting, when the design could not be made completely self weighting due to operational considerations. In a partially self weighting design, there is not one single common weight, but a set of common weights. One procedure of getting the required estimate in such a case is to obtain sample totals for the groups of units having the common weights and then to compute the overall estimate by weighting the group totals with the appropriate weights. But in large scale surveys, where a large number of items are to be estimated with detailed cross classifications, this procedure may become cumbersome. To avoid this, it may be desirable to take the least weight as standard and to repeat the sample observations having other common weights in such a manner that for the repeated observations the standard weight becomes the appropriate weight. For instance, if there are only two weights, one being twice the other, then after suitable duplication all the observations will have the same weight. If one of the weights is 1.5 times the other, it is necessary to duplicate a sample of half the observations having this weight and to retain as such the other half of the observations without repetitions. For instance, in

the example illustrated in Table 12.2, we can have the common weight 45 for all the sample observations by repeating twice those having the weight 90 and four times those having the weight 180.

While making the design self-weighting either at the field stage or at the tabulation stage, it is desirable to make the weights multiples and sub-multiples of each other, or at least multiples of a common weight. For instance, in rounding off the weights to multiples of a convenient number like 100, 1000, etc., we would get the weights as multiples of the selected convenient number and in this case complete self-weighting is achieved by repeating some of the observations twice, thrice, etc. depending on the ratio of their weights to the convenient number chosen. It may be mentioned that even in the case of making the design self-weighting by selecting a sub-sample with pps, size being the weight, the design need not necessarily be completely self-weighting as some sample units may get selected more than once and hence the weight to be used for a selected sample unit is $r (w/n')$, where r is the number of repetitions of this unit in the sub-sample. In this case the sample observations can be made to have the same weight by repeating them r times.

It is to be noted that the advantage of self-weighting would be lost if the least weight is so small as to lead to numerous repetitions of the sample observations having other weights. In such a case, it is advisable to choose some reasonably small weight (not necessarily the minimum weight) and to repeat or to reject the sample observations in suitable proportions. Suppose in a self-weighting design we have 5% of the sample observations with the minimum weight 50 and all the other observations have weights, which are multiples of 100. Then it may be economical to reject half of the sample observations having the weight 50, thereby making the weight of the retained observations 100 and to repeat suitably those observations having weights greater than 100.

REFERENCES

- CHURCH, B M (1954) Problems of sample allocation and estimation in an agricultural survey, *J Roy Stat Soc, (B)* 16, 223-235
- HANSEN, M H, HURWITZ, W N and MADOW, W G (1953) *Sample Survey Methods and Theory* Volume I John Wiley & Sons, New York
- LAHIRI, D B (1954) Technical paper on some aspects of the development of the sample design, *Sankhya*, 14, 264-316
- MURTHY, M N and SETHI, V K (1959) Self weighting design at tabulation stage National Sample Survey Working Paper 5 also *Sankhya*, 27, (B), (1965), 201-210
- MURTHY, M N and SETHI, V K (1961) Randomized rounded off multipliers, *J Amer Stat Assn*, 56, 328-334
- SOM, R K (1959) Self weighting sample design with an equal number of ultimate stage units in each of the selected penultimate stage units, *Bull Cal Stat Assn*, 8, 59-66
- SUKHATME, P V and PANSE, V G (1951) Crop surveys in India-II, *J Ind Soc Agr Stat*, 3 97-168

COMPLEMENTS AND PROBLEMS

12.1 To estimate the total number of persons (P) and the average household size (P/H) in an urban area a sample of 24 urban blocks is selected in the first stage with ppswr, size being the previous census population and from each sample block a sample of households is selected linear systematically with a random start. The sampling interval to be used in each sample block for selecting households is so specified that the sampling design becomes self weighting with the constant weight 480. Using the data given in Table 12.4, estimate P unbiasedly and obtain its rse by estimating its sampling variance unbiasedly. Also estimate the ratio P/H and estimate its rse.

12.2 For estimating the total agricultural population (Y) in a region, a sample of villages was selected from each stratum with ppswr size being the previous census population, and a sample of households was selected from each sample village linear systematically. The sampling intervals used in the sample villages were so specified that the sampling design was self weighting with 250 as the constant inflation factor for each sample household. Using the data given in Table 12.5, estimate Y unbiasedly and obtain its rse.

TABLE 12.4. NUMBER OF SAMPLE HOUSEHOLDS (h) AND NUMBER OF PERSONS (p) IN THEM FOR 24 SAMPLE BLOCKS.

block	h	p	block	h	p	block	h	p
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	8	35	9	5	19	17	1	6
2	7	40	10	9	35	18	13	54
3	5	22	11	7	36	19	0	0
4	6	32	12	6	32	20	6	18
5	5	16	13	5	26	21	5	27
6	6	28	14	10	38	22	4	20
7	2	8	15	7	28	23	5	21
8	9	32	16	8	29	24	11	47

TABLE 12.5. AGRICULTURAL POPULATION IN SAMPLE HOUSEHOLDS FOR 18 SAMPLE VILLAGES.

stratum	no. of sample villages	sample village						
		1	2	3	4	5	6	7
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	7	57	48	72	63	71	54	62
2	5	48	35	76	54	30	—	—
3	6	25	22	34	45	68	55	—

12.3 For studying the living conditions of the working class population residing in an industrial area, a stratified two-stage sampling design is proposed, in which from each stratum a sample of factories is to be drawn systematically with pps, size being the number of workers in an earlier period (x), and a sample of workers is to be selected from each sample factory linear systematically with a random start using the current payroll. The relevant data are given in Table 12.6.

- (i) Determine the constant weight to be used in a self-weighting design for ensuring a total sample size of about 1000 workers.
- (ii) Specify the sampling interval to be used in each sample factory for achieving a self-weighting design, using the constant weight determined in (i).
- (iii) Also find the approximate number of workers expected to be selected from each sample factory, thereby determining approximately the total sample size.

TABLE 12.6 NUMBER OF WORKERS IN SAMPLE FACTORIES

stratum	sample factory	number of workers		stratum	sample factory	number of workers	
		past	current			past	current
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	1	99	163	3	1	2697	2839
	2	523	465		2	4667	6255
	3	110	64		3	1423	1158
	4	741	829		4	1064	1150
2	1	4200	3504	4	1	90	91
	2	3187	2527		2	618	416
	3	2215	2186		3	150	131
	4	5322	5285		4	266	282

(number of workers in an earlier period—stratum 1 5896, stratum 2 43096,
 stratum 3 31625 stratum 4 10774)

12.4 In a socio economic survey 6 villages were selected from each stratum with ppswr, size being the 1961 census population. In case of big villages, the village was divided into a given number of divisions, each having approximately the same number of households and one of them was selected with srs. Then all the households in the selected division of the sample village were listed for facilitating selection of sample households. Using the data given in Table 12.7, make the sampling design self weighting by suitably specifying the sampling intervals to be used in the sample villages (or divisions of sample villages) so as to get 3 households per village on the average and determine the value of the common inflation factor for all the strata taken together.

12.5 Show that a two stage sampling design, where n fsu's are selected with ppswr, size being the number of ssu's in them and a constant number (m) of ssu's are selected from each sample fsu with srs w/o r, is self weighting for estimating the population total. Derive the sampling variance and obtain an unbiased estimator for it.

12.6 Suppose a non self weighting design was adopted in a survey for drawing a sample of n ultimate stage units. To reduce the cost of tabulation, a sub sample of n' units is selected from the original sample of n units with ppswr size being their inflation factors. Show that the estimator of the population total Y based on the sub sample would be self weighting and using the sub sample obtain an unbiased estimator of the increase in variance due to sub sampling at the tabulation stage.

TABLE 12.7. VILLAGE-WISE DATA ON PROBABILITY OF SELECTION AND NUMBER OF HOUSEHOLDS FOR THE SAMPLE VILLAGES/DIVISIONS.

stratum	sample village	probability of selection	number of divisions	households in sample division
(1)	(2)	(2)	(4)	(5)
1	1	0.002155	2	202
	2	0.003115	2	248
	3	0.000531	1	172
	4	0.000995	1	162
	5	0.002157	2	177
	6	0.006709	7	203
2	1	0.005372	2	195
	2	0.007223	2	210
	3	0.006160	2	184
	4	0.004572	2	166
	5	0.003665	1	208
	6	0.006156	2	201
3	1	0.005059	3	182
	2	0.002135	2	114
	3	0.027408	15	175
	4	0.004787	3	183
	5	0.015665	9	200
	6	0.008445	5	194

(total rural population in 1951 census—stratum 1 : 899725;
stratum 2 : 261372; stratum 3 : 574639).

- 12.7 Suppose in making a stratified multi-stage sampling design self-weighting, the sampling interval to be used in a penultimate stage sample unit becomes a fractional integer. Show how the sampling interval can be randomized to an integer retaining the unbiasedness of the estimator and the self-weighting nature of the sampling design.

12.8 Suppose $\{y_i\}$ and $\{a_i\}$ are the values of the study variable for the sample units and their corresponding weights such that $\sum_{i=1}^n a_i y_i$ is unbiased for the population total Y . If the weights $\{a_i\}$ are substituted by the random variables $\{r_i\}$, $i = 1, 2, \dots, n$, taking the values of the rounded off weights $\{b_i\}$, $b_1 < b_2 < \dots < b_k$, such that $E\left(\sum_{i=1}^n r_i y_i\right) = \sum_{i=1}^n a_i y_i$, show that the weights $\{b_i\}$ which minimize $V\left(\sum_{i=1}^n r_i y_i\right)$ are given by $b_1 = \min\{a_i\}$, $b_k = \max\{a_i\}$ and by

$$\sum_{b_j < a_i \leq b_{j+1}} y_i^2 \leq \frac{1}{(b_{j+1} - b_{j-1})} \sum_{b_{j-1} < a_i \leq b_{j+1}} (a_i - b_{j-1}) y_i^2 \leq \sum_{b_j \leq a_i \leq b_{j+1}} y_i^2$$

for $j = 2, \dots, k-1$.

(Murthy, M. N. and Sethi, V. K., *J. Amer. Stat. Assn.*, 56, (1961), 328-334)

Non-Sampling Errors

13.1 STUDY OF NON-SAMPLING ERRORS

In the previous chapters we have developed the theory of sampling assuming that the *true* value of each unit in the population can be obtained and tabulated without any error. Accordingly one would expect that a complete enumeration of all the units in the population would give rise to data free from errors. This is not usually the case in practice. For instance, it is difficult to completely avoid errors of observation or ascertainment. So also in the processing of data, tabulation errors may be committed affecting the final results. Errors arising in this manner are termed *non-sampling errors*, as they are due to factors other than the inductive process of inferring about the population from a sample. Under the conditions usually obtaining in large-scale census and survey work, occurrence of non-sampling errors is not only possible, but is also unavoidable. Thus, the data obtained in a census by complete enumeration, although free from sampling error, would still be subject to non-sampling error, whereas the results of a sample survey would be subject to sampling error as well as non-sampling error. In some situations the non-sampling errors may be large, and deserve greater attention than the sampling error. While, in general, sampling error decreases with increase in sample size, non-sampling error tends to increase with the sample size. In the case of complete enumeration non-sampling error and in the case of sample surveys both sampling and non-sampling errors require to be controlled and reduced to a level, at which their presence does not vitiate the use of the final results.

In recent years there has been a growing recognition of the need for assessing and controlling the non sampling errors that are apt to arise at the various stages of collection and tabulation of statistical data in large scale censuses and surveys. The increasing awareness of the existence of such errors owes much to the fairly widespread use of the sampling method, one of the main advantages of which is that it provides an opportunity for greater control of non sampling errors as well. Most of the sources and types of non sampling errors as also the techniques for assessment and control of these errors, considered in the subsequent sections, are applicable to both complete enumeration and sample surveys.

Problems of measurement and control of non sampling errors have been studied and various techniques have been developed by several authors. Mahalanobis (1940, 1944, 1946), Mahalanobis and Sengupta (1951), Mahalanobis and Lahiri (1961), Birnbaum and Sirken (1950), Durbin (1954) and Lahiri (1958a, 1958b) have given a number of techniques for assessing and controlling errors in censuses and surveys. Hansen and others (1946, 1951, 1961) and Sukhatme and Seth (1952) have examined the question of non sampling errors in census and survey work and they have furnished mathematical models for such errors. Post-enumeration checks and re interview surveys are increasingly being used in censuses and surveys as a means of assessing non sampling errors. A brief review of contributions to this field is given by Murthy (1963) and a fairly detailed one is available in a publication by Zarkovich (1965).

13.2 SOURCES OF NON-SAMPLING ERRORS

Non sampling errors can occur at every stage of planning and execution of the census or survey. The chief causes of non sampling errors are lack of proper specification of the domain of study and scope of the investigation, incomplete coverage of the population or sample, faulty definitions, defective methods of data collection, and tabulation errors. More specifically, non sampling errors may arise from one or more of the following factors:

- (i) data specification being inadequate and inconsistent with respect to the objectives of the census or survey;
- (ii) omission or duplication of units due to imprecise definition of the boundaries of area units, incomplete or wrong identification particulars of units or faulty methods of enumeration;
- (iii) inaccurate or inappropriate methods of interview, observation or measurement with inadequate or ambiguous schedules, definitions or instructions;
- (iv) lack of trained and experienced investigators;
- (v) difficulties involved in actual data collection arising from recall error and other types of errors on the part of respondents (including non-response);
- (vi) lack of adequate inspection and supervision of primary staff;
- (vii) inadequate scrutiny of the basic data;
- (viii) errors in data processing operations such as coding, punching, verification, tabulation, etc.; and
- (ix) errors committed during presentation and printing of tabulated results, graphs, etc.

These sources are not exhaustive, but are given to indicate some of the possible sources of error. In a sample survey, non-sampling errors may also arise due to defective frames and faulty selection of sampling units.

Non-sampling errors may be broadly classified into three categories (a) *specification errors*, (b) *ascertainment errors*, and (c) *tabulation errors* corresponding to the three stages (planning, field work and tabulation) of census or survey work. Specification errors at the planning stage can occur for reasons such as (i) to (iii) listed above. Ascertainment errors may arise in data collection due to factors such as (iv) to (vi) and tabulation errors are due to factors such as (vii) to (ix).

Ascertainment errors may be further sub divided into (i) *coverage errors* owing to over or under-enumeration of the population or sample, resulting from duplication or omission of units, and from non response, and (ii) *content errors* relating to wrong entries due to errors on the part of investigators and respondents. The same division can be made in the case of tabulation errors also, as there is a possibility of missing or repeating some data at the tabulation stage and thereby giving rise to coverage errors, and also of errors in coding, punching, calculations etc., which give rise to content errors.

13.3 TREATMENT OF NON-SAMPLING ERRORS

In the next few sections, a mathematical treatment of non sampling errors is given to bring out the components which together make up the total error in an estimate based on a sample survey. The conceptual background needed for this purpose is given in this section.

The difference between the sample survey estimate and the parametric true value being estimated may be termed *total error*. If complete accuracy can be ensured in the procedures such as determination, identification and observation of sample units and the tabulation of the collected data then the total error would consist only of the error due to sampling, termed *sampling error*. A measure of the sampling error is supplied by the mean square error which is the expected value of the square of the difference between the estimator and the true value. This mean square error is composed of two parts—square of *sampling bias* and *sampling variance*. If the results are also subject to non sampling error, then the *total error* would consist of both sampling and non sampling errors. In the following, a measure of the non sampling errors in terms of *non sampling bias* and *non sampling variance* is developed.

The *true value* of a unit is to be conceived of as a characteristic of the unit independent of the survey conditions. But, the value *reported* for a unit may be affected by survey conditions. The age of a person at a particular point of time, the income of a person

during a particular period or the number of persons in a country at a specified time are examples of characteristics for which the true values can be clearly defined. But there are many items of information, such as intelligence of a person, attitude to some social measures, consumer preference to certain articles, etc. for which it is very difficult even to conceive of the true values. In such cases some suitable conceptually defined value will have to be taken as the true value. For the definition of a true value to be useful in practice, it should serve the purpose of the survey by being well defined and observable under *reasonable conditions of survey* relating to subject coverage, method of enquiry, survey period, reference period and method of tabulation.

Suppose a sample has been chosen to be canvassed under reasonable conditions of survey and that there are two populations, one of investigators suitable for data collection and the other of computors (clerks) qualified for processing work. If we were to carry out the survey repeatedly on the same sample of units with different samples of investigators and computors chosen with a probability design, we may get different results because of the various sources of error present under the usual operational conditions. Here three steps of randomization could be visualized : selection of units, investigators and computors. The difference between the expected value of the estimator taken over all the three steps of randomization and the true value may be termed *total bias*. This consists of both *sampling bias* and *non-sampling bias*. The variance of the estimator, taken over all the three steps of randomization, measures the divergence of the estimator from its expected value and comprises sampling variance, variance between investigators, variance between computors and some interactions between the three sources of error. Thus we see that the total error consists of sampling bias and variance, non-sampling bias and variance and some interactions between the sample and the sources of non-sampling errors.

13.4 NON-SAMPLING BIAS

For the sake of simplicity, let us assume only two steps of randomization one for selecting the sample of units and the other for selecting the survey personnel. Here we consider the survey personnel as a whole instead of as investigators and computers. Let \hat{Y}_{sr} be the estimate of the population mean \bar{Y} based on the s th sample of units supplied by the r th sample of the survey personnel. The conditional expected value of \hat{Y}_{sr} taken over the second step of randomization for a fixed sample of units is given by

$$E_r(\hat{Y}_{sr}) = \hat{Y}_s, \quad (13.1)$$

which may be different from the estimate \hat{Y}_s based on the true values of the units in the sample. The expected value of \hat{Y}_s over the first step of randomization gives

$$E_s(\hat{Y}_s) = \bar{Y}', \quad (13.2)$$

which is the value for which an unbiased estimator can be had by the specified survey process. This value \bar{Y}' may be different from the true population mean \bar{Y} and the difference

$$B_t(\hat{Y}_{sr}) = \bar{Y}' - \bar{Y} \quad (13.3)$$

may be termed *total bias*.

It may be noted that the sampling bias is given by

$$B_s(\hat{Y}) = E_s(\hat{Y}_s) - \bar{Y}, \quad (13.4)$$

which is the difference between the expected value of the estimator based on the true values of units and the true value of the population mean. Since the total bias is the sum of sampling and non sampling biases, the non sampling bias is given by

$$B_r(\hat{Y}_{sr}) = B_t(\hat{Y}_{sr}) - B_s(\hat{Y}_s) = \bar{Y}' - E_s(\hat{Y}_s) = E_s(\hat{Y}_s - \bar{Y}), \quad (13.5)$$

which is the expected value of the non-sampling deviation. In a complete enumeration, there is no sampling bias and hence the total bias consists only of non-sampling bias. In the case of sample surveys also, the total bias will consist only of non-sampling bias, if as it is usually done estimators are used which are unbiased from the point of view of sampling of units.

There are a number of techniques available for the assessment of non-sampling bias (Lahiri, 1958a, 1958b). The survey figure may be compared with a figure separately obtained by some other agency or by the same agency on some other occasion after making the necessary adjustments for differences in coverage, definitions, survey period, etc. Such a comparison termed *external aggregate check* may provide a broad check on the survey figures. A better check would be to have unit by unit comparison of the survey data with the corresponding values in some other survey. This method is termed *external unitary check*. There would, however, be considerable difficulties in matching the units for this type of check. In these checks the assumption is that one source of data is more reliable than the other. If this assumption does not hold, it would be difficult to conclude which set of figures is subject to more bias in case of disagreement. Another technique of assessing non-sampling bias is to draw the sample in the form of two or more interpenetrating sub-samples and to get them surveyed by separate groups of investigators possibly with different intensity of training and experience.

The non-sampling bias in a census can be estimated by surveying a sample of units in the population using better techniques of data collection and compilation than those adopted under general census conditions. Surveys called *post-enumeration surveys*, usually conducted just after the census for studying the quality of the census data, may be used for this purpose. In a large-scale sample survey also, the ascertainment bias can be estimated by resurveying a sub-sample of the original sample using better survey techniques. Another method of checking survey data is to compare the values of the units obtained in two surveys and to reconcile discrepant figures by further investigation. This method of checking is termed *reconciliation*.

(check) surveys The different procedures available for assessing non-sampling errors are discussed in some detail in Section 13.10

13.5 NON-SAMPLING VARIANCE

The mean square error of the estimator \hat{Y}_{sr} , based on the s th sample of units and supplied by the r th sample of the survey personnel, is by definition

$$M(\hat{Y}_{sr}) = E_{sr} (\hat{Y}_{sr} - \bar{Y})^2, \quad (13.6)$$

where \bar{Y} is the true value being estimated. This is a measure of the divergence of the estimator from the true value, taking into account both sampling and non sampling errors. This measure consists of bias and variance, that is,

$$M(\hat{Y}_{sr}) = V(\hat{Y}_{sr}) + B^2(\hat{Y}_{sr}) = E(\hat{Y}_{sr} - \bar{Y}')^2 + (\bar{Y}' - \bar{Y})^2, \quad (13.7)$$

where \bar{Y}' is the expected value of the estimator taken over both the steps of randomization. The variance of the estimator is a measure of the divergence of the estimator from its expected value and $(\bar{Y}' - \bar{Y})$ is the bias. Taking the variance over the two steps of randomization, we get

$$V_{sr}(\hat{Y}_{sr}) = V_s E_r(\hat{Y}_{sr}) + E_s V_r(\hat{Y}_{sr}) = V_s(\hat{Y}_s) + E_s E_r(\hat{Y}_{sr} - \hat{Y}_s)^2 \quad (13.8)$$

From (13.8) we see that the variance can be split up into two parts, sampling variation and non sampling variation. The second term on the right hand side of (13.8) stands for the expected value of the square of the response deviations of the sample estimates from their expected value taken over both the stages of randomization. This term can be further split up by proceeding as follows

$$\hat{Y}_{sr} - \hat{Y}_s = (\hat{Y}_{sr} - \hat{Y}_s - \bar{Y}_r + \bar{Y}') + (\hat{Y}_r - \bar{Y}'),$$

where $\hat{Y}_r = E_r(\hat{Y}_{sr})$, and hence we get

$$E_{sr}(\hat{Y}_{sr} - \hat{Y}_s)^2 = E_{sr}(\hat{Y}_{sr} - \hat{Y}_s - \hat{Y}_r + \bar{Y}')^2 + E_r(\hat{Y}_r - \bar{Y}')^2 \quad . \quad (13.9)$$

The first term on the right hand side of (13.9) is the interaction between the sampling and non-sampling errors and the second term is the variance between survey personnel. Thus we see that the mean square error of the estimator consists of sampling variance, interaction between sampling and non-sampling errors, variance between survey personnel and square of the sum of the sampling and non-sampling biases. In a complete census the mean square error is composed of only the non-sampling variance and square of the non-sampling bias.

13.6 SIMPLE RANDOM SAMPLING

Let us consider an example where a simple random sample of n units drawn with replacement from a population of N units is divided at random into k equal sub-samples of m units each and these sub-samples are surveyed by k investigators selected with equal probability from a large population of K investigators qualified for this work. Let Y_{ij} and Y_i be the value reported by the j -th investigator ($j = 1, 2, \dots, K$) for the i -th population unit ($i = 1, 2, \dots, N$), and its true value respectively. Suppose y_{ij} is the value reported for the i -th sample unit ($i = 1, 2, \dots, m$) by the j -th selected investigator ($j = 1, 2, \dots, k$). An estimator of the population mean is given by

$$\bar{y} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^m y_{ij}, \quad (n = km). \quad \dots \quad (13.10)$$

The expected value of the estimator taken over the two steps of randomization is

$$E(\bar{y}) = \frac{1}{N} \sum_{i=1}^N Y'_i, \quad \left(Y'_i = \frac{1}{K} \sum_{j=1}^K Y_{ij} \right), \quad \dots \quad (13.11)$$

and the total bias, which in this case consists wholly of response bias, is

$$B_t(\bar{y}) = B_r(\bar{y}) = \frac{1}{N} \sum_{i=1}^N (Y'_i - Y_i) = \bar{Y}' - \bar{Y}. \quad \dots \quad (13.12)$$

The variance of the estimator is given by

$$V_{sr}(\bar{y}) = V_s E_r(\bar{y}) + E_s V_r(\bar{y}),$$

where the subscripts denote the steps of randomization. The conditional expected value of \bar{y} over the second step of randomization for a fixed set of sample units is

$$E_r(\bar{y}) = \frac{1}{n} \sum_{i=1}^n y_i, \quad \left(y_i = \frac{1}{K} \sum_{j=1}^K y_{ij} \right)$$

The unconditional variance of this over the first step of randomization is given by

$$V_s E_r(\bar{y}) = \frac{\sigma_d^2}{n}, \quad \sigma_d^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (13.13)$$

The conditional variance of \bar{y} over the second step of randomization for a fixed sample of units is

$$\begin{aligned} V_r(\bar{y}) &= \frac{1}{k} E_r \left(\frac{1}{m} \sum_{i=1}^m y_{ij} - \frac{1}{m} \sum_{i=1}^m y_i \right)^2 \\ &= \frac{1}{km^2} \frac{1}{K} \sum_{j=1}^K \left[\sum_{i=1}^m (y_{ij} - y_i)^2 + \sum_{i=1}^m \sum_{i' \neq i} (y_{ij} - y_i)(y_{i'} - y_i) \right], \end{aligned}$$

for $V_r\left(\frac{1}{m} \sum_{i=1}^m y_{ij}\right)$ is the same for all j . Taking the unconditional expected value of $V_r(\bar{y})$ over the first step of randomization, we get

$$\frac{1}{km^2} \frac{1}{K} \sum_{j=1}^K \left[\frac{m}{N} \sum_{i=1}^N (Y_{ij} - Y_i)^2 + \frac{m(m-1)}{N(N-1)} \sum_{i=1}^N \sum_{i' \neq i} (Y_{ij} - Y_i)(Y_{i'} - Y_i) \right],$$

that is,

$$E_s V_r(\bar{y}) = \frac{1}{km} \sigma_d^2 \{1 + (m-1)\rho_c\}, \quad . \quad (13.14)$$

where σ_d^2 is termed *simple or uncorrelated response variance* and is given by the variance of individual response deviations, that is,

$$\sigma_d^2 = \frac{1}{KN} \sum_{i=1}^N \sum_{j=1}^K (Y_{ij} - Y_i)^2 \quad (13.15)$$

and ρ_c is the intraclass correlation among the response deviations in a sample canvassed by one investigator (*intra-investigator correlation*), and is given by

$$\rho_c \sigma_d^2 = \frac{1}{KN(N-1)} \sum_{j=1}^K \sum_{i=1}^N \sum_{i' \neq i}^N (Y_{ij} - Y'_i)(Y_{i'j} - Y'_{i'}). \quad \dots \quad (13.16)$$

Hence, the variance and the mean square error of \bar{y} are

$$V(\bar{y}) = \frac{\sigma_s^2}{n} + \frac{\sigma_d^2}{n} \{1 + (m-1)\rho_c\} \quad \dots \quad (13.17)$$

and

$$M(\bar{y}) = V(\bar{y}) + (\bar{Y}' - \bar{Y})^2. \quad \dots \quad (13.18)$$

In the case of complete enumeration, the sampling variance will be zero and hence the variance and mse of the census figure \bar{y}' are given by

$$V(\bar{y}') = (\sigma_d^2/N)\{1 + (m-1)\rho_c\} \quad \dots \quad (13.19)$$

and

$$M(\bar{y}') = V(\bar{y}') + (\bar{Y}' - \bar{Y})^2. \quad \dots \quad (13.20)$$

The result (13.17) shows the contribution to the total variance from the response variation and it also brings out the impact of the intra-investigator correlation on the response variance. The intraclass correlation will be positive if the response deviations for the different units have a consistent tendency to be in one direction for an investigator. Even when this correlation is small, the contribution to the response variation may be considerable if m , the number of units surveyed by each investigator is large.

An unbiased estimator of $V(\bar{y})$ in (13.17) is given by

$$v(\bar{y}) = \frac{1}{k(k-1)} \sum_{j=1}^k (\bar{y}_{.j} - \bar{y})^2, \quad (\bar{y}_{.j} = \frac{1}{m} \sum_{i=1}^m y_{ij}), \quad \dots \quad (13.21)$$

for $E\left(\sum_{i=1}^k \bar{y}_{.j}^2 - k\bar{y}^2\right) = k[kV(\bar{y}) + \bar{Y}'^2 - V(\bar{y}) - \bar{Y}'^2] = k(k-1)V(\bar{y})$.

This result shows that if k independent samples are surveyed by k investigators selected with equal probability from a large population of investigators, then it is possible to get an unbiased estimator of the total variance (but not of the total mse). This procedure is known as the method of *interpenetrating sub samples*, which is considered in Sub section 13.10g. The variance between investigators is

$$\sigma_r^2 = \frac{1}{K} \sum_{j=1}^k (\bar{Y}_j - \bar{Y}')^2 = \sigma_d^2 \rho_c \quad (13.22)$$

For

$$\begin{aligned} \sigma_r^2 &= \frac{1}{K} \sum_{j=1}^k \left[\frac{1}{N} \sum_{i=1}^N (Y_{ij} - \bar{Y}_j)^2 \right]^2 \\ &= \frac{1}{KN^2} \sum_{j=1}^k \sum_{i=1}^N (Y_{ij} - \bar{Y}_j)^2 + \frac{1}{KN^2} \sum_{j=1}^k \sum_{i=1}^N \sum_{i' \neq i}^N (Y_{ij} - \bar{Y}_j)(Y_{i'j} - \bar{Y}_i) \\ &= \frac{\sigma_d^2}{N} + \frac{N-1}{N} \sigma_d^2 \rho_c = \sigma_d^2 \rho_c, \end{aligned}$$

if N is large. In this case an unbiased estimator of σ_r^2 is

$$\hat{(\sigma_r^2)} = k \bar{e}(\bar{y}) - \frac{1}{km(m-1)} \sum_{j=1}^k \sum_{i=1}^m (y_{ij} - \bar{y}_j)^2, \quad (13.23)$$

for, taking the conditional expected value of the second term in (13.23), we get

$$\frac{1}{mk} \sum_{j=1}^k \frac{1}{N} \sum_{i=1}^N (Y_{ij} - \bar{Y}_j)^2$$

and the expected value of this expression over the sample of investigators, is given by

$$\frac{1}{m} \frac{1}{NK} \sum_{j=1}^k \sum_{i=1}^N (Y_{ij} - \bar{Y}_j)^2 = \frac{1}{m} (\sigma^2 - \sigma_r^2),$$

where σ^2 is the total variance in the population and is given by

$$\sigma^2 = \frac{1}{KN} \sum_{j=1}^K \sum_{i=1}^N (Y_{ij} - \bar{Y})^2 = \sigma_s^2 + \sigma_d^2. \quad \dots \quad (13.24)$$

Hence,

$$E(\hat{\sigma}_r^2) = k \left\{ \frac{\sigma^2}{mk} + \frac{(m-1)}{mk} \sigma_r^2 \right\} - \frac{1}{m} (\sigma^2 - \sigma_r^2) = \sigma_r^2.$$

13.7 ESTIMATION OF A PROPORTION

It is interesting to consider the question of response variance in estimating a population proportion. Let Y_{ij} be 1 or 0 according as the j -th investigator reports the i -th unit in the population as belonging to a particular class or not and let P'_i be the proportion of the investigators reporting the i -th unit in the population as belonging to that class. Suppose a sample of n units is drawn with srswr from a population of N units to be surveyed by a sample of k persons selected with equal probability from a large population of K persons qualified for this work. An estimator of the population proportion P is given by

$$\hat{P} = \frac{1}{mk} \sum_{j=1}^k \sum_{i=1}^m y_{ij}, \quad (n = mk). \quad \dots \quad (13.25)$$

The expected value of this estimator over both the stages of randomization is

$$E(\hat{P}) = \frac{1}{N} \sum_{i=1}^N P'_i = P', \quad \dots \quad (13.26)$$

and the bias, which in this case consists of only the ascertainment bias, is $(P' - P)$. In this case σ_s^2 and σ_d^2 defined in (13.13) and (13.15) respectively are given by

$$\sigma_s^2 = \frac{1}{N} \sum_{i=1}^N (P'_i - P')^2 \quad \dots \quad (13.27)$$

and

$$\sigma_d^2 = \frac{1}{N} \sum_{i=1}^N P_i Q_i, \quad (Q_i = 1 - P_i) \quad (13.28)$$

The variance of \hat{P} is

$$V(\hat{P}) = \frac{1}{Nn} \sum_{i=1}^n (P_i - \bar{P})^2 + \frac{1}{Nn} \sum_{i=1}^N P_i Q_i \{1 + (m-1)\rho_e\} \quad (13.29)$$

From (13.21) it can be seen that an unbiased estimator of the total variance given in (13.29) is

$$v(\hat{P}) = \frac{1}{k(k-1)} \sum_{j=1}^k (p_j - \bar{p})^2, \quad (13.30)$$

where p_j is the sample proportion reported by the j th selected investigator in the sample assigned to him and \bar{p} is the overall sample proportion. From (13.23) we see that when N is large, an unbiased estimator of the variance between investigators σ_r^2 is given by

$$(\hat{\sigma}_r^2) = k v(\hat{P}) - \frac{1}{k(m-1)} \sum_{j=1}^k p_j q_j, \quad (q_j = 1 - p_j) \quad (13.31)$$

If the intraclass correlation is assumed to be 0, then the variance given in (13.29) reduces to

$$V(\hat{P}) = P Q / n, \quad (Q = 1 - P') \quad (13.32)$$

This result is interesting because it shows that the expression which is normally used as the sampling variance of a sample proportion includes not only the sampling variance but also the uncorrelated response variance (Hansen, Hurwitz and Bershad, 1961). An unbiased estimator of the variance is

$$v(\hat{P}) = pq/(n-1), \quad (q = 1 - p) \quad (13.33)$$

since $E(pq) = E(p)E(q) = P - V(p) - P^2 = (n-1)V(p)$. Here again we see that the variance estimator of a sample proportion, generally used, is unbiased for the total variance, which includes both the sampling variance and the uncorrelated response variance.

13.8 COST FUNCTION

Let us consider the problem of determining optimum values of k the number of investigators, and m the number of units to be assigned to each investigator, so as to minimize the total variance for a given fixed cost. Let the cost function be

$$C = C_0 + kC_1 + nC_2, \quad \dots \quad (13.34)$$

where C_0 is the overhead cost, C_1 is the cost of recruiting and training one investigator, C_2 that of surveying one unit and $n = km$. The total variance of the estimator \bar{y} of \bar{Y} , given in (13.17), may be written as

$$V(\bar{y}) \doteq \frac{\sigma^2 - \sigma_r^2}{n} + \frac{\sigma_r^2}{k}, \quad \dots \quad (13.35)$$

since $\sigma_r^2 \doteq \rho_c \sigma_d^2$ and $\sigma^2 = \sigma_s^2 + \sigma_d^2$. Minimizing the variance in (13.35) with respect to n and k subject to the cost restriction (13.34), we get

$$k = \frac{C - C_0}{\sqrt{C_1 \sigma_r^2} + \sqrt{C_2 (\sigma^2 - \sigma_r^2)}} \sqrt{\left(\frac{\sigma_r^2}{C_1}\right)} \quad \dots \quad (13.36)$$

and

$$m = \sqrt{\frac{C_1}{C_2}} \sqrt{\frac{\sigma^2 - \sigma_r^2}{\sigma_r^2}}. \quad \dots \quad (13.37)$$

13.9 NON-RESPONSE ERROR

One of the sources of error in censuses and surveys mentioned earlier is incomplete coverage in respect of units. This may occur due to refusal by respondents to give information, or their being *not at home*, sample units being inaccessible, and so on. The error in this case would arise because the set of units getting excluded may have characteristics so different from the set of units actually surveyed as to make the results biased. This type of error is termed *non-response error*, since it arises from the exclusion of some of the

anticipated units in the population or sample. Obviously, the non-response error is not important if the characteristics of the non responding units are similar to those of the responding units. But such similarity of characteristics between the two types of units is not always obtained in practice. For instance, if a particular questionnaire is mailed to all farmers in a region, the non response rate may be higher among the farmers holding smaller areas of land. While non response cannot be completely eliminated in practice, it could be overcome to a great extent by persuasion through repeated visits, or by some other methods.

One way of dealing with the problem of non response is to make all efforts to collect information from a sub sample of the units not responding in the first attempt (Hansen and Hurwitz, 1946). Suppose out of n units, selected with srs wr from a population of N units, n_1 units respond and $n_2 (= n - n_1)$ units do not respond in the first attempt. Let a sub sample of n'_2 units be selected from the n_2 non responding units with srs wr for making special efforts to collect the information. If \bar{y}_1 and \bar{y}_2 are sample means based on the n_1 units responding in the first attempt and on the sub sample of n'_2 units respectively, then an unbiased estimator of \bar{Y} is given by

$$\hat{\bar{Y}} = \frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}_2) \quad (13.38)$$

There are two phases of randomization in this case sampling of n_1 units and sub sampling of n'_2 units from the n_2 units not responding in the first attempt. Taking the variance of the estimator given in (13.38) over the two phases of randomization, we have

$$V_{12}(\hat{\bar{Y}}) = V_1 E_2(\hat{\bar{Y}}) + E_1 V_2(\hat{\bar{Y}})$$

The conditional expected value and variance over the second phase of randomization are given by

$$E_2(\hat{\bar{Y}}) = \frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}_2) = N \bar{y}, \quad (\bar{y}_2 = \frac{1}{n_2} \sum_{t=1}^{n_2} y_{2t})$$

and

$$V_2(\hat{Y}) = \frac{1}{n^2} \{n_2(n_2 - n'_2)\} \frac{s_2^2}{n'_2}, \quad s_2^2 = \frac{1}{(n_2 - 1)} \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2,$$

where \bar{y} is the sample mean based on all the n units in the sample and y_{2i} is the value of the i -th non-responding unit in the sample. Hence, the variance of the estimator is given by

$$V(\hat{Y}) = V_1(\bar{y}) + E_1 \left\{ \frac{1}{n} \left(\frac{n_2}{n} \right) (k-1) s_2^2 \right\},$$

where $k = n_2/n'$. Since the conditional expected value of s_2^2 given n_2 is $N_2 \sigma_b^2 / (N_2 - 1)$ and the expected value of (n_2/n) is (N_2/N) , where σ_b^2 is the variance between the units in the population not responding in the first attempt and N_2 is the number of such non-responding units in the population, the variance becomes

$$V(\hat{Y}) = \frac{(N-n)}{(N-1)} \frac{\sigma^2}{n} + \frac{1}{Nn} (k-1) \frac{N_2^2 \sigma_b^2}{(N_2-1)}, \quad \dots \quad (13.39)$$

where σ^2 is the variance between the N units in the population.

A suitable cost function in this case is

$$C = C_1 n + C_2 n P + C_3 \frac{n}{k} Q, \quad \dots \quad (13.40)$$

where C_1 is the cost per unit for the first attempt at data collection, C_2 the cost per unit for tabulation, C_3 the cost per unit sampled from the non-responding units (for obtaining data by additional efforts and for tabulation), P is the proportion of units in the population that would have responded in the first attempt, and $Q = 1 - P$. The optimum value of n and k which would minimize the cost, ensuring at the same time a given value V for the variance of the estimator, are given by

$$n = n' \left\{ 1 + (k-1) Q^2 \frac{(N-1)\sigma_b^2}{(N_2-1)\sigma^2} \right\} \quad \dots \quad (13.41)$$

and

$$L = \sqrt{\left\{ \frac{N^2(N_2-1)\sigma^2}{N_2^2(N-1)\sigma_b^2} - 1 \right\}} \frac{C_3 Q}{C_1 + C_2 P}, \quad (13.42)$$

where $n' = N\sigma^2/\{\sigma^2 + (N-1)V\}$ is the sample size required for ensuring the value V for the variance if there were complete response. If it is assumed that $\sigma^2 = \sigma_b^2$ and that $N/(N-1)$ and $N_2/(N_2-1)$ are nearly equal to unity, the optimum values on n and L reduce to

$$n = n'\{1 + (L-1)Q\}, \quad (13.43)$$

and

$$L = \sqrt{\frac{C_3 P}{C_1 + C_2 P}} \quad \dots \quad (13.44)$$

An interesting device in dealing with *not at home* cases has been considered by Politz and Simmons (1949). The procedure consists in ascertaining from the responding informants the chance of their being at home at a particular point of time and weighting the results with the inverse of this chance. For instance, the households may be asked whether they were at home at some specified time during the previous five days. Then the households can be classified as being at home once in six visits, twice in six visits and so on, and the data obtained for the different classes may be weighted by the inverse of the respective probabilities of being at home, (cf Problem 13.5, p 478). In practice, some bias would still persist because of persons not at home during the entire investigation period.

13.10 MEASUREMENT AND CONTROL OF ERRORS

The adoption of suitable methods for assessing non sampling errors and the choice of adequate procedures for controlling them require careful consideration even before the initiation of the main census or sample survey, as some of the procedures will have to be incorporated in the design of the census or survey itself. Besides such built in procedures for error measurement and control, it may

be necessary to have a separate programme for estimating the different types of non-sampling errors that may be present in the final results. For these, it will be necessary to use more refined methods of data collection and compilation which might involve greater per-unit cost. An account of some of the common procedures of measuring and controlling non-sampling errors is given in the following sub-sections. Most of these procedures relate mainly to the assessment of non-sampling bias considered in Section 13.4.

Just as a substantial increase in sample size is required to effect even marginal reductions in the sampling error after a certain stage, enormous cost and effort would be needed if the last few traces of non-sampling errors in the final results were to be removed. This is due to the fact that major sources and types of non-sampling errors are easier to detect by using the method of sample check mentioned in later sub-sections and it is usually simpler to take steps to control them, whereas detection and rectification of minor sources of error will require considerably greater time and effort. A rational approach to the problem of controlling non-sampling errors will, therefore, be to try to reduce them as much as possible to levels at which the results will be usable for the purpose in view, but not to such extents as will render the efforts and costs to become incommensurate with the improvements achieved.

13.10a CONSISTENCY CHECKS

In designing the questionnaires or schedules, special care has to be taken to include certain items of information that will serve as a check on the quality of the data to be collected. If these additional items of information are simple to obtain, they may be canvassed for all the units covered in the census or survey; otherwise, they may be canvassed only for a sample (or sub-sample) of units. For instance, in a population census, where the *de jure* method is followed, it may be helpful to collect information on a *de facto* basis also, so that it will be possible to work out the number of persons temporarily present and the number of persons temporarily absent,

and a comparison of these two figures will give an idea of the quality of the census data. Similarly, the inclusion of items leading to certain stable ratios such as the sex ratio may be useful in assessing the quality of census and survey data.

Another technique for assessing the quality of data, the use of which permits the location of doubtful observations, consists in arranging the observations in increasing order of some basic variable, such as per capita expenditure in a consumer expenditure survey, and then plotting against each sample unit, the value of a related variable such as proportion of expenditure on food to total expenditure. This graph, which is expected to follow a certain pattern, would help in spotting out any discrepant values not conforming to the general pattern. Then the discrepant values may be checked for accuracy.

In the case of variables such as yield rate of crop, salt consumption, etc., whose coefficients of variation may be quite stable over time and space in a region, the quality of the data can be assessed by calculating the relative standard error of the estimate corresponding to each investigator or some other source and comparing it with a standard figure based on past surveys. This method is particularly useful in detecting the tendency on the part of some investigators to supply apparently consistent but incorrect data.

13.10b SAMPLE CHECK

One way of assessing and controlling non sampling errors in censuses and surveys is to independently duplicate the work at the different stages of operation with a view to facilitating the detection and rectification of errors. Since it may not be feasible to do this with complete coverage owing to considerations of cost, time, shortage of trained personnel and the like, the duplicate checking can only be carried out on a sample of the work by using a comparatively smaller group of trained and experienced staff. If the sample is properly designed and if the checking operation is efficiently carried out, it would be possible, not only to detect the presence of

non-sampling errors, but also to get an idea of their magnitude. Such a procedure may be termed *method of sample check* and a sample selected for purposes of checking the quality of data may be termed *check-sample*.

If it were possible to have a complete check of the census or survey work, the quality of the final results could be considerably improved. But in the case of a sample check, the rectification work can be carried out only on the sample checked and hence the errors in the non-sampled part would remain uncorrected. This difficulty is partly overcome by dividing the output at different stages of the survey (such as the filled-in schedules, coded schedules, punched cards, computation sheets, etc.) into lots and checking samples selected from each lot. In this case, when the error rate in a particular lot is more than a specified level, the whole lot may be checked and corrected for the errors, thereby improving the quality of the final results.

13.10c POST-CENSUS AND POST-SURVEY CHECKS

An important type of sample check, commonly used to assess non-sampling errors, consists in selecting a sample (or sub-sample) of the units covered in the census (or survey) and in re-enumerating or re-surveying it by using better trained and more experienced survey staff than those employed for the main investigation. This procedure is termed *post-census* or *post-survey check* depending on whether it is applied to complete enumeration or to a sample survey. For the check-survey to be effective, it is necessary to ensure : (i) that the re-enumeration or re-survey is taken up immediately after the main census or survey to avoid any possible *recall error*; and (ii) that steps are taken to minimize the *conditioning effect* that the main survey may have on the work of the check-survey.

The check-survey should be so designed as to facilitate assessment of both the *coverage* and the *content* errors. For this purpose, it is first desirable to re-enumerate all the units in a sample (or sub-sample) of higher stage area units (village, enumeration district, etc.), with a view to detecting coverage errors and then to re-survey only

a sample of ultimate units, ensuring proper representation for different parts of the population which have special significance from the point of view of non sampling errors

To ensure that the check survey is done effectively, the investigators may be provided with the original list of sample units and the detailed filled in questionnaires or schedules obtained in the main census or survey for at least a part of the check sample after incorporating a few dummy errors and the check survey list and data collected by them should be compared with the original ones. This would lead to an assessment of the quality of work of the check survey. The knowledge of the presence of artificially introduced errors would make the investigators more cautious. It is desirable to instruct the investigators to list also every unit adjacent to each check sample unit (for example the unit nearest to, or preceding or following the sample unit) and then to check whether such units were covered in the main census or survey.

A special advantage of a check survey is that it facilitates a *unitary check*, which consists in first matching the data obtained in the two enumerations for those units covered by the check sample and then analysing the observed individual differences. When discrepancies are found, efforts are made to find the cause of their presence to gain an insight into the nature and types of non sampling errors. This is effected by getting the discrepant units observed or investigated by specially trained and experienced staff through what is termed a *reconciliation survey*. However, it is difficult to use the unitary check in practice due to the difficulties involved in a large scale matching operation. Matching of two sets of data is generally a time consuming and costly operation and especially so when the address and other identification particulars are not very specific. But the special advantages of a unitary check may, under certain circumstances, make it worth while in spite of the difficulties involved.

If it is not possible to have a unitary check owing to cost and operational considerations, a simpler, but less effective procedure, termed *aggregate check*, may be used. This consists in comparing

estimates of parameters given by the check-survey data with those based on the main census or survey. Such an aggregate check gives only an idea of the *net error*, which is the resultant of positive and negative errors, whereas a unitary check provides information on both net error and *gross error*, the latter being the absolute total of both positive and negative errors ignoring their sign. Incidentally, net error can be considered to be a measure of the non-sampling bias in the census results and gross error as a measure of the non-sampling variation (cf. Problem 13.1, p. 477).

An important point in a check-survey is the determination of the sample size in such a way as to permit detection of any error greater than a specified value at a given level of confidence. The required sample size would depend on a number of factors such as the type of sampling design, type of check (unitary or aggregate), the level at which the estimates of error are required and the precision envisaged for these estimates. The data collected in a previous census or survey and its check-survey may be used to determine the sample size and if no such information is available, a pilot survey is to be undertaken on a moderate scale to obtain data useful in fixing the sample size. While publishing the results of the check-survey, variance estimates of the estimates of non-sampling errors should also be given to facilitate proper interpretation of the check-survey results.

It is evident that in a post-census or post-survey check, the same concepts and definitions as those used in the original census or survey be followed and that, at the same time, steps be taken to avoid the types of errors which are generally possible in a large-scale operation. For instance, a common tendency in post-enumeration surveys is to conduct them on more or less the same lines as the main census or survey using the same or only slightly different staff; this leads to a considerable under-estimation of the magnitude of the non-sampling errors; giving the users an erroneous picture of the real quality of the main survey. The need to guard against this cannot be over-emphasized. In fact, it is essential that the post-census or post-survey work is conducted by an independent set of specially trained personnel.

13.10d EXTERNAL RECORD CHECK

Another method for assessing non sampling errors, especially in census work, is to take a sample of relevant units from a different source, if available, and to check whether all the units have been enumerated in the main investigation and whether there are any discrepancies between the values when matched. It may be noted that the list from which the check sample is drawn for this purpose, need not be a complete one. For instance, in carrying out a check on a population census, the check sample may be formed by selection from recent birth registrations, school children, old age pensioners, etc. The above method is termed *external record check*. A main difficulty in this method will be to locate and identify the units of the check sample and match them with those of the main investigation. The method is of considerable use when alternative lists of units exist, whether complete or partial.

13.10e QUALITY CONTROL TECHNIQUES

There is ample scope for applying statistical quality control techniques to census and survey work, because of the large scale and repetitive nature of the operations involved in such work. Control chart and acceptance sampling techniques could be used with advantage in assessing the quality of data and in improving the reliability of the final results in large scale surveys and censuses. Among the various techniques, those which prescribe rules for initiating corrective action are of special interest. For an example of such a technique we may consider a routine census or survey operation in which the output of each operator can be objectively checked and the error-rate worked out. The work of each operator is checked hundred per cent for an initial period of time, but if the error rate falls below a specified level, only a sample of his work is verified, the sample size being suitably determined, if not, complete verification of his work is continued. The above decision as to the nature of verification is taken separately for each operator, and is based on his cumulated

error-rate over a past period. Techniques of this type are of considerable use in large-scale work, as they not only reduce the cost of verification operations, but also ensure specified quality levels for final results. These techniques are being used with much advantage in censuses and surveys by the United States Bureau of the Census and some other organizations (Murthy, 1964).

13.10f STUDY OF RECALL ERROR

Response errors arise owing to various factors such as the attitude of the respondent towards the survey, method of interview, skill of the investigators and *recall error*. Of these, recall error deserves particular attention, as it presents certain special problems often beyond the control of the respondents. Recall error depends on the length of the reporting period and on the interval between the reporting period and the date of survey. The second factor may be taken care of by choosing for the reporting period a suitable interval preceding the date of survey or as near to it as possible. The choice of reporting period and its effect on the quality of data due to recall errors need careful consideration.

One way of studying recall error is to collect and analyse data relating to more than one reporting period in a sample (or sub-sample) of units covered in the census or survey. The main difficulty in such a study is that there is a certain amount of *conditioning effect* possibly due to the data reported for one reporting period influencing those reported for other reporting periods. Moreover, it is difficult to decide which reporting period is to be accepted, when there is disagreement between the results based on different reporting periods, although this can be overcome to some extent by having a control sample and making special efforts to get reliable data for a period of time covering all the tentative reporting periods.

To avoid the conditioning effect mentioned above, data for the different periods under consideration may be collected from different samples of units. The only difficulty in this approach is that for getting an effective comparison a large sample size is required, if

independent samples of units are used for the different reporting periods. This difficulty can be obviated to a large extent by canvassing the different reporting periods in linked interpenetrating sub samples, which are so selected that the correlation coefficient between any two of the sub samples is positive and large.

Another method of studying recall error is to collect some additional information which will permit estimates for different reporting periods to be obtained. For instance, in a demographic survey, one may collect information not only on the number of births during the last year but also on the date, week or month of birth, at least for a sample of units. This will then enable a tabulation of birth rate classified according to the interval between the occurrence of the event and the reporting date, which will in turn reveal any recall error that may be present in the reported data on number of births, (Som, 1967).

13.10g INTERPENETRATING SUB-SAMPLES

The technique of interpenetrating sub samples, originally developed by P C Mahalanobis during the 1930's, consists in drawing the sample in the form of two or more sub samples, selected in an identical manner and each capable of providing a valid estimate of the population parameter. This technique helps in providing "a means of control (i.e appraisal) of the quality of the information," as the interpenetrating sub samples can be used 'to secure information on non sampling errors such as differences arising from differential interviewer bias, different methods of eliciting information, etc (United Nations, 1964). After the sub samples have been surveyed by different groups of investigators and processed by different teams of workers at the tabulation stage, a comparison of the final estimates based on the sub samples provides a broad check on the quality of the survey results. For instance, in comparing the estimates based on four sub samples surveyed and processed by different groups of survey personnel, if three estimates agree among themselves and the other estimate differs widely from them in spite of the sample size being large enough, then normally one would suspect the quality of work in the discrepant sub sample.

If the interest is mainly in assessment of the quality of data, then the sub-samples should be linked up in the sense of the estimates based on them having a high positive correlation. This can be ensured by selecting a sample of clusters of two or more homogeneous units and forming sub-samples by taking one unit from each cluster. This procedure would increase the effectiveness of comparison of sub-sample estimates with a view to estimating differential non-sampling errors. For instance, in the case of a population census, the work in a sample of enumeration units can be so arranged that one set of alternate households, dwelling units or houses are canvassed by one investigator and the other set by another investigator to bring out the differential non-sampling errors, which is the difference between their biases. It is to be noted that what is attained in this way is only an idea of the differential non-sampling error and not an idea of the magnitude of the non-sampling error itself. That is, if the magnitude and the direction of the biases of two investigators were of the same order, a comparison of the sub-sample figures would generally show an agreement even when the magnitude of the bias of each investigator is considerable. However, this point can be met by getting one of the sub-samples surveyed by specially trained and experienced investigators.

It would be possible to study the various components of the errors such as variation between investigators, variation between tabulating teams and possible interactions by selecting the sample in the form of a number of interpenetrating sub-samples and getting them surveyed and tabulated by different groups of investigating and tabulating staff. For instance, if there are two tabulating teams and four parties of investigators, a sample may be selected consisting of 8 (or a multiple of 8) sub-samples, one half of which will be tabulated by one team and one-fourth to be surveyed by one party of investigators. This scheme will help in analysing the total variation into its components, such as variation between tabulating teams, that between parties of investigators and the residual error. This analysis of variance is useful, as it helps in focussing attention on those stages of survey work where errors are preponderant and in

taking corrective action to reduce the non sampling variation at the different stages

It is desirable to apply the technique of interpenetrating sub-samples also in post-census, post-survey and other sample checks with a view to assessing their own effectiveness as checks

REFERENCES

- BIRNBAUM, Z W and SIRKEN, M G (1950) Bias due to non availability in sampling surveys, *J Amer Stat Assn*, 45, 98-111
- DURBIN, J (1951) Non response and call backs in surveys, *Bull Inter Stat Inst.*, 34, (2), 72-86
- HANSEN, M H and HURWITZ W N (1946) The problems of non response error in sample surveys *J Amer Stat Assn* 41 517-529
- HANSEN M H, HURWITZ, W N, MARSH E S and MAULDIN, W P (1951) Response errors in surveys, *J Amer Stat Assn*, 46, 146-190
- HANSEN, M H, HURWITZ W N and BERSHAD M A (1961) Measurement of errors in censuses and surveys, *Bull Inter Stat Inst*, 38, (2), 359-374
- LAHIRI D B (1958a) Recent developments in the use of techniques for assessment of errors in national surveys in India *Bull Inter Stat Inst*, 36, (2), 71-93
- LAHIRI D B (1958b) Observations on the use of interpenetrating samples in India, *Bull Inter Stat Inst*, 36 (3) 144-152
- MAHALANOBIS, P C (1940) A sample survey of the acreage under jute in Bengal, *Sankhya* 4, 511-530
- MAHALANOBIS P C (1944) On large scale sample surveys, *Phil Trans Roy Soc.*, 231, (B) 329-451
- MAHALANOBIS, P C (1946) Recent experiments in statistical sampling in the Indian Statistical Institute, *J Roy Stat Soc* 109, 325-370
- MAHALANOBIS, P C and SENGUPTA, J M (1951) On the size of sample cuts in crop cutting experiments in the ISI 1939-50, *Bull Inter Stat Inst*, 33, (2), 359-403
- MAHALANOBIS, P C and LAHIRI D B (1961) Analysis of errors in censuses and surveys with special reference to experience in India, *Bull Inter Stat Inst*, 38 (2), 401-433, reprinted in *Sankhya* 23, (A), 325-358
- MURTHY, M N (1963) Assessment and control of errors in censuses and surveys, *Sankhya* 25 (B), 263-282
- MURTHY, M N (1964) The work of the United States Bureau of the Census with emphasis on sample designs and control of errors in censuses and surveys, *Sankhya*, 26, (B), 257 300
- POLITZ, A N and SIMMONS, W R (1949) An attempt to get the not at homes into the sample without call backs, *J Amer Stat Assn*, 44 9-31
- SOM, R K (1967) *Recall Lapse in Demographic Enquiries*, Asia Publishing House, Bombay.
- SUKHATME, P V and SETH, G R (1952) Non sampling errors in surveys, *J Ind Soc Agr Stat*, 4, 5-41

- UNITED NATIONS (1964): *Recommendations for the Preparation of Sample Survey Reports*; Statistical Papers series C, No. 1, Rev. 2, New York.
- ZARKOVICH, S. S. (1965): *Sampling Methods and Censuses, Volume II, Quality of Statistical Data (Draft)*; Food and Agricultural Organization of the United Nations, Rome.

COMPLEMENTS AND PROBLEMS

13.1 Suppose in a post-census survey, it is proposed to estimate the non-sampling bias and variance in the case of a certain dichotomous variable by carefully surveying a sample of n persons selected with srswr. Let a persons be reported as 1 and d persons as 0 both in the census and in the survey, whereas let b persons be reported as 0 in the census and 1 in the survey, and c persons as 1 in the census and 0 in the survey.

Assuming that the differences between the responses in the census and the survey are uncorrelated, estimate the non-sampling bias and also estimate the variance of the response differences. Show that the net error estimates the non-sampling bias and that when n is fairly large the gross error estimates the response variance.

(United States Bureau of the Census, Technical
Paper No. 6, (1963), Washington, D.C.).

13.2 Suppose n units are selected with srswr from a population of N units and these are assigned at random to k investigators selected from a population of K investigators with srswr in such a manner that each investigator observes only m of the units and each sample unit is observed by the same number r of investigators. Assuming that the sample observation y_{ij} reported by the j -th investigator for the i -th sample unit follows the model

$$y_{ij} = y'_i + \alpha_j + e_{ij},$$

where y'_i is the true value of the i -th sample unit, α_j is the average bias of the j -th investigator and e_{ij} is a random variable with expected value 0 and with variance σ_e^2 , find the bias when the sample mean is considered as an estimator of the true population mean, namely, $\bar{Y}' = \frac{1}{N} \sum_{i=1}^N Y'_i$ and derive its sampling variance. Also obtain an unbiased variance estimator for the sample mean, when N and K are very large compared to n and k and when $r = 1$.

(Sukhatme, P. V., *Rev. Inter. Stat. Inst.*, 20, (2/3), (1952), 121-133).

13.3 In a mail enquiry, n units were selected with srswr and of these n_1 units responded. From the $n - n_1$ ($= n_2$) non-responding units, r units were selected with srswr and the required information was obtained by personal interview. Suggest a suitable unbiased estimator of the population mean \bar{Y} and obtain its variance. Suppose the results of a pilot study have shown that

- (i) the non-response rate $Q (= 1 - P)$ is about 25% at the first attempt;
- (ii) population coefficient of variation for the variable under study is 100%;

- (iii) the ratio of the variance of the non responding units to the overall population variance is 0.5, and
- (iv) the values of C_1 , C_2 and C_3 are respectively Rs 0.15, 1.00 and 4.00 in the cost function

$$C = C_1 n + C_2 n P + C_3 (n/L) Q$$

Find the optimum values of n and L that would minimize the total cost of the survey such that the rse of the estimator is the same as that of a sample of 100 units selected with srswr from the population in the event of complete response. Also find the minimum cost of the survey.

13.4 Suppose in a population, P_{11} and P_{12} are the proportions of persons who are available for interview during a specified period of time and will answer respectively yes and no to a question and let P_{21} and P_{22} be the proportions of persons who are not available and who would answer respectively yes and no, if contacted. Let n units be drawn with srswr for the survey and let p be the proportion of persons in the sample who answer 'yes' to the question out of the persons available for interview. If the parameter under consideration is $P_1 (= P_{11} + P_{21})$, find the expression for bias of p' and obtain lower and upper bounds for the bias by noting that $0 \leq P_{21}/(P_{21} + P_{22}) \leq 1$

(Burnbaum, Z. W. and Sirken, M. G., *J. Amer. Stat. Assn.*, 45, (1950), 98-111)

13.5 Suppose in a complete enumeration survey, it is proposed to survey a population of N families over a period of s distinct time points. The time of survey for every family is chosen at random out of the s occasions and at the time of the survey the number of occasions it was or would be available for data collection is ascertained from each family.

Assuming that each family is available for data collection at least on one of the s occasions, show that the following estimator is unbiased for the population total Γ

$$\hat{\Gamma} = \sum_{i=1}^s \frac{1}{s} \sum_{j=1}^{N_i} d_{ij} Y_{ij},$$

where Y_{ij} is the value reported by the j th unit available for data collection on s occasions, d_{ij} is 1 or 0 according as that unit is contacted on the randomly selected occasion or not and N_i is the number of units available for data collection on s occasions. Find the variance of this estimator and obtain an unbiased estimator of this variance.

(Durbin, J., *Bull. Inter. Stat. Inst.*, 34, (2), (1954), 72-86)

13.6 The data obtained in a post census survey relating to livestock numbers in India in 1956 are given in Table 13.1. Using this data and assuming that (i) the sample is drawn with srswr, (ii) the response differences are independent and (iii) the households for which census information was not available were reported as not having any livestock in the census, estimate the percentage of under or over enumeration in the census and also obtain its rse.

13.7 The method of *self validation* is defined as the estimation or compilation of the same information from two or more sets of data, one *direct* and the other *indirect*, at least a major portion of which is collected from two (or more) different sources and by two (or more) sets of investigators or enumerators in a single survey operation, and

TABLE 13.1. DISTRIBUTION OF HOUSEHOLDS BY THE NUMBER OF WORKING BULLOCKS REPORTED AS PER LIVESTOCK CENSUS (c) AND SAMPLE VERIFICATION (s) IN INDIA : 1956.

		n.a. total																			
		n.a. total																			
		n.a. total																			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
0	13110	199	260	8	13	3	5	-	-	-	-	-	-	-	-	-	-	-	1135	14733	
1	240	2308	261	19	12	1	-	-	-	-	-	-	-	-	-	-	-	-	327	3168	
2	337	196	5650	177	106	7	5	-	1	-	-	1	-	1	-	-	-	-	793	7275	
3	17	7	117	472	62	8	3	2	2	-	-	1	-	-	-	-	-	-	66	757	
4	23	4	102	52	849	33	27	-	2	-	1	-	-	-	-	-	-	-	99	1192	
5	-	10	2	31	91	17	1	1	-	-	-	-	-	-	-	-	-	-	6	169	
6	6	-	8	5	19	7	172	5	12	-	3	-	-	-	-	-	-	-	13	249	
7	-	1	-	-	-	-	7	26	3	-	1	-	-	-	-	-	-	-	38	38	
8	1	-	1	-	5	1	14	2	44	2	3	-	3	-	-	-	-	-	5	81	
9	-	1	-	1	1	1	1	1	7	1	-	-	-	-	-	-	-	-	1	16	
10	-	-	-	1	-	1	-	1	-	2	1	12	-	1	-	-	1	-	3	22	
11	-	-	-	-	-	-	-	-	-	-	-	1	6	1	-	-	-	-	7	7	
12	-	-	1	-	-	1	-	-	2	-	4	-	-	-	-	-	-	-	6	6	
13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	
14	-	-	-	-	-	-	-	-	-	-	1	1	-	5	-	2	-	-	9	9	
15	-	1	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	2	2	
20-40	1	-	-	-	-	-	-	-	-	-	-	-	-	1	-	10	1	1	13	13	
total	13734	2716	6411	735	1099	162	253	37	68	10	25	6	10	1	5	2	4	1	10	2450	27720

n.a.—not available.

Source : Mahalanobis, P. C. and Lahiri, D. B. (1961) : Analysis of errors in censuses and surveys with special reference to experience in India; *Bull. Inter. Stat. Inst.*, 38, (2), 401-433, reprinted in *Sankhyā*, 23, (4), 325-358.

comparison of the results based on direct and indirect information. An example of this is given here.

Suppose in a population census in a village, it is ascertained from each person at the time of enumeration the number of his (her) brothers and the number of his (her) sisters (i) who are staying with him (her), (ii) who are not staying with him (her), but are residing in the same village, and (iii) who are residing outside the village. Using this information, explain how it is possible to get a measure of the bias, if any, in the enumeration of persons in a census.

In Table 13.2 are given the results of a study of this type conducted in a village. Carefully examine the data given in this table and briefly discuss the nature of biases involved in the enumeration.

TABLE 13.2 RELATIONSHIPS OF VARIOUS TYPES BETWEEN BROTHERS AND SISTERS RESIDENT IN THE SAME VILLAGE BUT NOT IN THE SAME HOUSEHOLD

no	relationship	Scheduled caste Hindus	other Hindus	Muslims	total
(1)	(2)	(3)	(4)	(5)	(6)
1.1	younger brother—elder brother	9	80	19	108
1.2	elder brother—younger brother	9	73	18	100
2.1	younger brother—elder sister	3	26	7	36
2.2	elder sister—younger brother	3	27	6	36
3.1	younger sister—elder brother	1	20	7	28
3.2	elder brother—younger sister	1	22	2	25
4.1	younger sister—elder sister	1	16	3	20
4.2	elder sister—younger sister	1	32	6	39

Source Lohani, D. B. (1958). Recent developments in the use of techniques for assessment of errors in nation wide surveys in India, *Bull Inter Stat Inst*, 36, (2), 71-93.

13.8 Suppose the data on births in a region are independently collected by a registrar (*R*) through the method of registration and by an interviewer (*I*) through the method of enquiry from the households. Let there be

- (i) C births recorded both by *R* and *I*,
- (ii) N_1 births recorded by *R*, but not by *I*, and found to be correct on verification,
- (iii) N_2 births recorded by *I*, but not by *R*, and found to be correct on verification, and
- (iv) w births recorded by *R* or *I*, but not by both, and found to be incorrect after verification.

Assuming that there is complete independence between *R* and *I* in reporting or missing births, show that an estimator of N , the total number of births, which is unbiased when N is large, is given by

$$\hat{N} = C + N_1 + N_2 + (N_1 N_2 / C)$$

and its variance is approximately equal to $N_1 q_2 / p_1 p_2$, where p_1 and p_2 are the probabilities of *R* and *I* recording a correct birth, $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

(Sekhar, C. C. and Deming, W. E., *J Amer Stat Assn*, 44, (1949), 101-115)

Planning of Sample Surveys

14.1 SCOPE

This chapter presents in a nutshell various aspects of planning sample surveys. Most of the steps involved in planning a sample survey are common to those for a complete enumeration. Three major stages of a survey are planning, data collection and tabulation of data. Some of the important aspects requiring attention at the planning stage are the following :

- (i) formulation of data requirements;
- (ii) *ad hoc* or repetitive survey;
- (iii) method of data collection;
- (iv) questionnaire *versus* schedule;
- (v) survey, reference and reporting periods;
- (vi) problem of sampling frame;
- (vii) choice of sampling design;
- (viii) planning of pilot survey;
- (ix) field work;
- (x) processing of data; and
- (xi) preparation of report.

These aspects are considered in some detail in the following sections. In addition, the advantages of multi-subject surveys and of a permanent survey organization are also briefly discussed.

The different aspects listed above are interdependent and decisions regarding them have to be taken jointly. Decisions made

independently on various aspects will have to be coordinated and reviewed at appropriate stages. The discussion in this chapter is only of a broad nature. It is generally not possible to give rules of thumb regarding the actions to be taken at the different stages of a survey, since these would depend very much on the data requirements, available resources and the operational conditions of individual surveys.

14.2 FORMULATION OF DATA REQUIREMENTS

The users that is, the persons or organizations requiring the statistical information are expected to formulate the objectives of the survey. The users formulation of data requirements is not likely to be adequately precise from the statistical point of view. It is for the survey statistician to give a clearer formulation of the objectives of the survey and check up whether his formulation faithfully reflects the requirements of the users. The survey statistician's formulation of data requirements should include the following:

- (i) a clear statement of the desired information in statistical terms,
- (ii) specification of the domains of study ,
- (iii) the form in which the data should be tabulated,
- (iv) the accuracy aimed at in the final results, and
- (v) cost of survey

In formulating the data requirements, it is necessary to visualize as comprehensively as possible the nature of the statistical information required to satisfy the current and the future needs of the user. A list of subjects of importance in national planning for economic and social development and on which data can be collected through sample surveys, is given below for illustration.

- (i) population, births and deaths, migration, employment and unemployment,
- (ii) income and expenditure, cost and level of living, savings and indebtedness,
- (iii) housing health and education,

- (iv) small and large-scale manufacture, trade, transport, professions and services;
- (v) agriculture; and
- (vi) wholesale and retail prices.

One can also visualize a large number of other items such as consumer preferences and demands, public reaction to policies and issues, and so on, the data on which may be required for commercial, administrative or other purposes.

As a part of formulation of data requirements, it is necessary to indicate the population totals, means, proportions, etc., for which estimates are required, by specifying the items or groups of items for which (i) only aggregate estimates, (ii) estimates of trend over time, or (iii) both aggregate and trend estimates are required, together with the periodicity (monthly, quarterly, yearly, etc.) with which they are required. It is advisable to specify wider intervals over which changes are to be estimated, as the utility of obtaining reliable estimates of changes over short intervals of time may not be usually commensurate with the cost. Further, the geographical and classificatory breakdowns for which the estimates are required should also be clearly specified. It should, however, be noted that if quick results are to be obtained with limited resources, it will be necessary to restrict as much as possible geographical and classificatory breakdowns. The survey statistician may have to explain these points to the users and get their requirements reconsidered, if necessary.

The next step in formulating data requirements consists in giving careful thought to the problem of specifying the margins of error that can be tolerated in the estimates for the purposes in view and in evaluating the cost requirements. The specification of such margins of error, referred to as *permissible errors*, for the various characteristics should be done after studying the implications of actions that are likely to be taken on the basis of the survey results. Usually this is not an easy task because of the complex nature of the decision making process. But it is extremely important that efforts are made to specify the permissible errors in a rational manner after an examination

of the actions taken on the basis of past survey results and of the losses and gains that have resulted therefrom. An under- or over-specification of the permissible error could lead to loss arising either from wrong decisions or from increase in the cost of the survey. Here it is desirable not to be over-ambitious in specifying the permissible error, because after a certain point even large increases in survey cost may not yield appreciable increases in the precision of estimates; further, too precise an estimate may not be really necessary in many situations in practice. Often the situation is one in which the user specifies the cost he is prepared to incur for the survey, leaving it to the survey statistician to obtain as much precision for the estimates as possible subject to the cost restriction.

In formulating data requirements for a survey, one may accommodate some additional items of information, directly or indirectly related to the objectives of the survey, which would provide checks on the accuracy of data or assist in interpreting the results. Subject matter, to be covered in detail in a future survey, may also be included to gain some preliminary information for estimating the cost and variance functions. These additional items can be obtained at marginal cost in many types of large-scale surveys.

14.3 SURVEY : AD-HOC OR REPETITIVE

Whether the survey is to be an *ad-hoc* one or is going to be a repetitive one needs examination. An *ad-hoc survey* is one, which is conducted without any intention of or provision for repeating it, whereas a *repetitive survey* is one, in which data are collected periodically from the same, partially replaced or freshly selected sample units. If the aim is to study only the current situation, the survey can be an *ad-hoc* one. But, when changes or trends in some characteristics over time are of interest, it is necessary to carry out the survey repetitively. When a repetitive survey is conducted without any appreciable lapse of time between completion and beginning of two successive surveys, it is termed a *continuing survey*.

In a repetitive survey, the series of surveys should be so planned as to minimize the overall cost ensuring at the same time specified precisions for the estimates of aggregates and rates of change over time, or they may be planned so as to maximize the precisions of the required estimates for a given cost. In either case, the plan should be flexible enough to permit modification on the basis of the experience accumulating from the successive surveys.

The trend over time may be studied either by carrying out surveys repeated regularly at fixed intervals of time, such as 6 months, 2 years or 5 years, or through a continuing survey. A continuing survey permits unusual fluctuations to be better accounted for and enables a fuller study of trend by providing estimates for the intervening periods. Further, it necessitates the maintenance of a regular staff, which could improve the quality of data through the accumulated experience.

14.4 METHOD OF DATA COLLECTION

The different methods of data collection are (i) physical observation or measurement, (ii) personal interview, (iii) mail enquiry, (iv) registration, and (v) transcription from records. The first four methods relate to collection of primary data from the units or the respondents directly, whereas the last one relates to the extraction of secondary data, collected earlier generally by one or more of the other four methods. These methods have their respective merits and are briefly discussed here. It is particularly important to give sufficient thought to the selection of an appropriate method or methods of data collection in any survey. The entire planning and execution of the survey is influenced considerably by the method of data collection. The decision regarding the choice of the method of data collection briefly described here should be arrived at after a careful consideration of accuracy, practicability and cost from among the alternative methods.

14.4a OBSERVATION AND MEASUREMENT

Data collection by *physical observation* or *measurement* consists in physically examining the units or respondents and recording data as a result of personal judgement or using a measuring instrument by the investigator. For instance, in an anthropological survey, the required data, such as head breadths and nose lengths of a sample of persons are obtained by measurement. In a crop survey, the area under different crops may be obtained by visually judging the crop grown and the proportion of crop area in each sample plot or field, the land area of which would have been measured earlier using land survey methods. Data obtained by this method are likely to be more accurate than those obtained by other methods such as (iii) and (iv), although it may involve greater effort and cost. Hence, the cost involved and the possible improvements in the quality of data are to be taken into account before deciding to adopt this method, and subject to these two considerations, this method should have preference over the other methods. However, there may be situations, such as the study of incidence of a certain disease requiring a careful medical examination, where method (i) has necessarily to be applied, and there may also be situations, such as the study of public opinion where this method cannot be effectively applied.

14.4b PERSONAL INTERVIEW

The method of *personal interview* consists in contacting the respondents and collecting statistical data by questioning. This method is being fairly widely used in social and economic surveys, since in this case the investigator personally contacts the respondents and can obtain the required data fairly accurately. He can clearly explain to the respondents the objectives of the survey and the exact nature of the data requirements, and persuade them to give the required information, thus reducing the possibility of non response arising from non cooperation, indifference, etc. Further, this method is most suitable for collecting data on conceptually difficult items from respondents.

14.4c MAIL ENQUIRY

In a *mail enquiry*, data are collected by obtaining questionnaires filled in by the respondents, the questionnaires being sent and collected back through an agency such as the postal department. This method is likely to cost much less than methods (i) and (ii), as no on-the-spot contact with the respondents is envisaged in this method. The main difficulty with this method is that it might give rise to a high rate of non-response under certain circumstances, which in its turn would lead to difficulties in interpreting the survey results. Of course, the rate of non-response may not be high, if the respondents have the ability to comprehend the contents of the questionnaire and are willing to supply the required information through the mail. In planning a mail enquiry, one may consider the possibility of supplementing it by the method of personal interview to follow up cases of non-response. A general drawback of the method is that only items of information conceptually easy to understand can be included in the questionnaire.

14.4d METHOD OF REGISTRATION

In the *registration method*, the respondents are required to register the required information at specified places. The vital statistics registration system followed in many countries provides an illustration of the registration method. In this case, the residents of a region are required to register every birth and every death with the vital statistics authorities. This method is in a sense a combination of the methods of mail enquiry and personal interview, as the respondents are informed of the data requirements through suitable notifications and the registration authorities ascertain and record the desired statistical data when the respondents go and report to them. This method is also fairly widely used, as the cost involved in adopting it will be much less than that in case of methods (i) and (ii). The main difficulty with this method, as in the case of the mail enquiry, is the possibility of non-response due to

indifference, reluctance, etc., on the part of the informants to go to the place of registration and supply the required data. Because of this difficulty, this method is generally used only to collect statistical data covered by statutory regulations, as in such cases the response rate is likely to be better due to fear of possible legal action and penalty. When this method is adopted only for a sample of units, it is termed *sample registration*.

14.4e TRANSCRIPTION FROM RECORDS

The method of *transcription* from records is used when the data needed for a specific purpose are already available in registers maintained in one or more places, making it no more necessary to collect them directly from the original units at much cost and effort. The method consists in compiling the required information from the registers for the concerned units. This method is extensively used, since a good deal of government and business statistics are collected as by product of routine administrative operations. Obviously the quality of the data obtained through this method can at best be the quality of the original data.

14.5 QUESTIONNAIRE VERSUS SCHEDULE

The question as to whether the questionnaire or schedule approach is to be used in the survey for collecting the required information needs consideration. In the *questionnaire* approach, the informants or respondents are asked prespecified questions and their replies to these questions are recorded by themselves or by investigators. This approach presumes that the respondents are capable of understanding and answering the questions, since in this case the investigator is not supposed to influence the responses in any way by his interpretation of the terms used in the questions. The questionnaire method is widely used in mail enquiries. In the *schedule* approach, the exact form of the questions to be asked are not given and the task of questioning and eliciting information is left to the investigator, who backed by his training and instructions has to

use his ingenuity in explaining the concepts and definitions to the informant for obtaining reliable information. Detailed instructions are, however, given to the investigators about the concepts, definitions and procedures to be used in collecting data for the survey.

It may appear that the schedule method of enquiry is subject to more investigator bias than the questionnaire method, as there is added scope in it for the investigator to influence the responses of the informants. This need not be so, if well-trained and skilled investigators are employed for conducting the survey. On the other hand, the respondent bias may become substantial in a questionnaire approach, if the survey items are rather complicated and involve conceptual difficulties. In such a case, it is more expedient to train a batch of investigators for explaining the concepts and definitions involved than to burden the respondents with elaborate instructions and clarifications. It may also be difficult to visualize all the possible situations one might come across at the time of the enquiry with a view to preparing a series of questions that would adequately cover them. Given the items together with necessary elucidations of concepts, definitions and procedures involved, the investigator is likely to be in a better position to obtain the required information. However, the cost of adopting the questionnaire approach is generally less than the schedule approach. Hence a decision as to which of the two approaches is to be followed in a particular survey should be arrived at after carefully examining the possible effects of investigator and respondent biases and the costs involved.

Preparation of a schedule or questionnaire with suitable instructions needs to be given careful consideration in planning a survey, as the utility of the results of the survey depends to a large extent on this. The framing of questions or items should be done in a simple, unambiguous, interesting and tactful manner, and they should be so worded as not to influence the answers of the respondents. The sequence of items is important. Those likely to help the investigators in establishing a good rapport with the respondents should be put first, and items relating to a particular aspect of the survey should

come together in the questionnaire or schedule. The items should be so arranged as to facilitate the work of tabulation. If this procedure of arranging the questions is unsuitable for collection of data, clear instructions are to be given to the investigators as to the sequence in which the questions are to be put. As far as possible the items should be such that the answers can be recorded in numbers, or specific codes. When codes are to be used, it is desirable to give both the description and the codes for at least important items to permit checking of the codes used.

To reduce the non sampling errors arising from ambiguous definitions and misunderstanding of the questions by the investigators or respondents, it is necessary to give some typical examples, detailed explanatory notes and instructions for the items of information included in the questionnaire or schedule. The instructions should include the concepts and definitions that are to be used in the survey, with an indication how these are expected to serve the objectives of the survey. Clarification of the doubts raised by the investigators is to be so done that there is uniformity in the procedures followed by the different investigators. If the data are collected through mail enquiry, the explanatory notes and instructions should be lucid, concise and precise. If the data are collected by the method of interview or by actual physical observation, the instructions can be made more detailed. In this case also the instructions should be clear and unambiguous.

14.6 SURVEY, REFERENCE AND REPORTING PERIODS

Another aspect requiring special attention is the determination of (i) *survey period*, the time period during which the required data are collected, (ii) *reference period*, the time period to which the collective data for all the units should refer, and (iii) *reporting period* (which is a part or whole of the reference period), the time period for which the required statistical information is collected for a unit at a time. The reference period depends on the objectives of the survey. The reporting period is determined by the nature

of the items of information and the conditions under which the survey is to be conducted. It may be necessary to have different reference and reporting periods for different items. The reporting and the reference periods for an item could be the same. The reporting period should be decided after conducting suitable studies to examine recall errors and other non-sampling errors. For instance, on the basis of appropriate studies, a week has been accepted as the reporting period in employment and unemployment enquiries in some countries. The reporting period may be taken as one year in the case of items, which occur rarely and can be easily remembered.

For items of information subject to seasonal fluctuation, it is desirable to have one complete year as the survey and reference periods the data being collected every month or season with suitable reporting periods for the same or different sets of sample units. The same set of sample units are to be used if the main objective is to study the changes from season to season, and different sets if the main objective is to obtain an average for the year as a whole, provided the values of the characteristics at different points of time have a positive correlation. The scheme will get reversed, that is, different sets and same set of sample units will be used respectively, if the above correlation is negative. Besides the advantage of taking care of seasonal fluctuations, the spreading out of the survey over a period, such as one year, makes it possible to manage the work with a relatively small number of survey personnel at a time.

It may be noted that when a large-scale survey is conducted using a fixed team of investigators, the survey may have to be staggered and in such a situation it is difficult to adopt a fixed set of time periods as reporting periods in view of recall errors. It may be desirable in such cases to use a *moving reporting period*, where the reporting period is one of specified duration prior to the date of the survey. For example, in a survey of one year duration from January to December of a year, a one-month reporting period would move from December of the previous year to November of the current year, giving rise to a reference period of one year, namely, December to

November A larger moving reporting period such as six months or a year may also be used if the recall error is not likely to be appreciable

Although technical considerations may demand different reporting periods for different items, it is desirable to use the same reference period for as many items as possible from the operational view point. This is useful in providing some internal checks and for studying inter relationships between different items. A common reporting period will, of necessity, have to be a short one, such as a week or a month though, if recall error is not large a longer reporting period is desirable, as it would reduce the sampling error.

In staggering a survey, it is advisable to divide the survey period into shorter sub periods (of one two or three months) and to canvass a representative sub sample of the total sample in each such sub period with a view to obtaining valid estimates periodically even when the survey is in progress and also to ensuring an even representation of different parts of the total reference period.

14.7 PROBLEM OF SAMPLING FRAME

One of the main requirements for efficiently designing a sample survey is a well constructed sampling frame, which, among other things, should be up to date and provide adequate information on relevant auxiliary characteristics. In many situations, it constitutes a basic problem, as usable sampling frames are either not available at all or have to be compiled from material produced mainly for administrative purposes at different times and places. Even when a frame is available, steps should be taken to ensure that it is free from omissions, duplications and other inaccuracies, and that the units are clearly identifiable for the purpose of investigation.

After having ensured a *minimal frame*, furnishing a list or map of units with their identification particulars, one should explore the possibility of getting information on some suitable auxiliary characteristics for use in stratification, allocation, arrangement, selection and estimation. The information on the auxiliary variables need not be

very accurate, since such information will be used only in a general manner to effect improvements on the sampling design. For this purpose, data collected by different agencies may have to be obtained and compiled in spite of any difficulties in matching them with the basic frame. The *master frame* with adequate supplementary information so built up will be of considerable use in designing more than one sample survey.

In multi-stage sampling, the problem of securing a good sampling frame arises for each of the stages. Usually a frame for higher stage units, such as provinces, towns, urban blocks and villages, is more stable than one for lower stage units such as farms, establishments and households, which are more subject to changes. The frames for the first one or two stages can usually be compiled from the census or other administrative records, but a satisfactory frame of ultimate units, such as persons, households, shops, fields, etc., may not be readily available. The frame of ultimate units, contained in a sample of the penultimate stage units, may have to be prepared by fresh listing or by up-dating an already available list of ultimate units.

At the stage of preparing a sampling frame within the sample penultimate stage units, it is desirable to collect for each ultimate unit information on some auxiliary characteristics having a bearing on the subjects of enquiry. Such information can be used in designing the sample in such a way as to reduce the contribution of the ultimate stage units to the total variation, which would result in a relatively small sample at the ultimate stage to ensure a specified precision for the estimate.

14.8 CHOICE OF SAMPLING DESIGN

The process of evolving a suitable sampling design for a survey, utilizing information provided by the sampling frame presents many problems. The determination of an optimum sampling design taking into account the various technical, operational and cost considerations becomes quite difficult; often one has to be satisfied with a rational sampling design chosen from among a few alternatives.

The principle generally adopted in the choice of a design is either the reduction of overall cost including field and tabulation costs with prespecified permissible errors or the reduction of the margins of error of the estimates with the total cost fixed. In both the cases it is extremely important to take adequate steps to control non sampling errors which could be large enough to vitiate the survey results. Special attention should also be paid to the workability of the particular sampling design under the prevailing operational conditions.

Generally a stratified uni stage or multi-stage design (usually two stage or three stage one) is adopted for large-scale sample surveys. If information is available on two or more auxiliary variables connected with the characteristics under consideration multiple (or deep) stratification may be resorted to using that information. Such stratification is usually achieved by the formation of a few strata in an optimum or near optimum manner using one of the stratification variables resulting generally in the exploitation of most of the benefit of stratification and then by further sub stratification of the strata using another stratification variable. To facilitate data collection by field work it is desirable to make the strata geographically compact so that each stratum can conveniently form the investigation zone for an investigator. There may be situations where the strata are not compact geographical areas and in such cases special efforts may be needed to form investigation zones after the sample selection is made. The question of allocation of sample size to the strata may be tackled by considering the allocations based on the data on auxiliary variables and any other information on the variability of the characteristics that may be available from previous censuses or surveys and then arriving at a compromise allocation. In practice it is generally found that moderate deviations from a specific allocation do not affect significantly the overall variability.

The selection of the units within the strata may be done conveniently by systematic sampling or by some other scheme such as sampling with varying probabilities systematically after arranging

the units in a suitable order. In varying probability sampling, the question of the choice of the measure of size arises. For instance, if the interest lies in studying some social characteristic of the people in a region as a whole, the previous population census figure may be a suitable measure of size when the sampling unit is an area unit. However, if the size measure has already been considered as a stratification variable or if the sizes are more or less equal (as is usually the case with population census enumeration districts or blocks), sampling may be done with equal probability systematically. The possibility of selection with probability proportional to other suitable measures of size through an appropriate integrated method of selection may also be examined. Alternatively, the other size measures may be used to improve the estimates at the estimation stage through the use of ratio or regression estimators.

In two-stage sampling the second stage units in the sampled first stage units may be selected systematically from the whole frame or from a suitably constructed sub-frame (in the case of special enquiries covering only a part of the general population) after effecting a suitable arrangement or sub-stratification. This process may be continued to the subsequent stages also in a multi-stage sampling design. The possibility of using clusters of units as sampling units at the different stages of the sampling design should also be considered with a view to reducing the cost.

~ It is useful to select the sample in the form of two or more inter-penetrating sub-samples, each of which is capable of providing a valid estimate of the population parameter. The main advantage of this technique is that it permits easy calculation of the estimates of the sampling errors irrespective of any complicated procedures involved in selecting the sample and in arriving at the estimates. This is particularly important in the case of sampling designs, where the first stage units are selected with varying probabilities systematically or without replacement and where less simple estimators such as ratio and regression estimators are used; since, in these cases, the expressions for the sampling variance and its estimator are rather complicated and are difficult to compute numerically on the basis

of a single sample. When these sub samples are surveyed and processed by different teams of field and processing staff, a comparison of the results based on the sub samples provides a check on the quality of the survey operations. As described earlier (cf Sub section 13.10g of Chapter 13), this technique can also be used to study the non sampling errors, such as the differential investigator bias, the differential effects of different methods of investigation, etc.

Another point, which needs attention, is the question of making the sampling design self weighting, the advantages of which have been discussed in Chapter 12. In situations where the data are processed manually or with conventional tabulating equipment, weighting the sample observations with appropriate inflation factors (necessitated by the use of a non self weighting design) may create a bottleneck at the tabulation stage and may hold up quick tabulation. Hence efforts should be made to make the sampling design fully self weighting with a single common inflation factor, if this is not possible for operational or other reasons, the design should at least provide for partial self weighting with only a limited number of inflation factors.

In evolving a rational sampling design it is necessary to consider carefully the question of programming the work in the field and at the tabulation stage. The field programme should be so drawn up as to obtain the maximum information per unit of cost consistent with the required standards regarding the quality of data. There should be a balance between field and tabulation work loads the tabulation programme being so planned as to avoid bottlenecks and delays in the publication of results.

14.9 PILOT SURVEY

For planning a survey effectively, some prior information about the population under consideration and on the operational and cost aspects of data collection and tabulation will be needed. When such information is not readily available from past surveys, it is desirable to design and carry out a *pilot survey*, for obtaining some

preliminary information on the variability of the characteristics to be studied and on the nature of cost of data collection and tabulation under different schemes with a view to building up cost and variance functions useful in planning the main survey. The pilot survey is also useful for (i) testing out the provisional schedules or questionnaires and the related instructions, (ii) evolving suitable procedures for field and tabulation work, and (iii) training field and tabulation staff. The scope and scale of the pilot survey would depend on the nature of the main survey. It may be possible to phase the main survey itself over time, so that the knowledge and experience gained in any phase of the survey would be of much help in planning and carrying out the subsequent phases of the survey.

14.10 FIELD WORK

One way of organizing data collection applicable to individual subject-fields is to use an existing agency to collect the required information as a by-product of its normal administrative activity. In that case, the cost is likely to be only marginal, because of the saving in overhead cost and as costs involved in journey, contacting units, etc. are avoided. Unless the items of information are very simple, and are related to the normal duties of the staff, the quality of data collected by this approach may not be very satisfactory, since this method does not offer much scope for employing investigators of desired qualifications. Another method of organizing data collection is to have a permanent field organization with well-trained whole-time staff. In this system the experience gained by the staff in the earlier surveys will be reflected in the later surveys and also it would provide better opportunities to develop and establish efficient schemes of field scrutiny, inspection and supervision.

The different aspects of field work such as recruitment and training of investigators, inspection and supervision etc., should be given careful consideration in the light of the prevailing operational conditions, since the quality of the final survey results can at best be only as accurate as the basic data collected in the field. The investigators

should be trained well not only in the concepts and definitions of the terms to be used in collecting the data, but also in the art of eliciting correct information from various sources. In recruiting the investigators, special attention should be paid to their ability to withstand long and arduous travel, sometimes in inhospitable terrain, and in inclement weather, with limited transport facilities. It is desirable that the investigators are trained by a single team of instructors instead of through a chain of intermediaries. If such direct training is not possible, the number of intermediate stages of training should be kept at a minimum.

Inspection work should consist of different types of check on the work of the investigators, such as spot-checks, pre- and post-survey checks, field scrutiny of filled-in schedules, etc. Besides regular inspection, oriented towards improving those investigators whose work is of poor quality, it will also be useful to have *random inspection* by checking a sample of the investigation work selected in such a manner as to enable getting overall estimates of the quality of the data collected in the field. The supervisory staff should be in constant touch with the designing and processing staff, so that they can obtain necessary clarifications and guidance on unusual or unforeseen cases met with in the field and can apprise the latter on the difficulties encountered in implementing the instructions. The ratio of supervisory staff to the primary staff should be relatively high in those regions, where travelling and contacting the investigators would take up a lot of time on account of inadequate transport and communication facilities.

It is necessary to keep a record of the time spent by investigators on different job-items in the survey operations such as preparation of sampling frame, identifying and contacting sample units, enquiry, interview or observation, journey, etc. Besides serving as a means to control the survey operations, these records would be useful in arriving at time and cost estimates for planning future surveys.

14.11 PROCESSING OF SURVEY DATA

Analysis of data collected in a survey has three facets : (i) tabulation or summarization of data, (ii) subject analysis, and (iii) statistical analysis.

14.11a SUMMARIZATION OF DATA

The first task, which is of primary importance, is the reduction of the collected data into meaningful tables. The tables should be presented along with the background information such as the objectives of the survey, the sampling design adopted, the methods used for data collection and tabulation and margins of error applicable to the results. The purpose of this would be to permit proper appreciation of the summary figures by the users and to prevent any misuse of the statistics presented. Estimates of margins of error supplied by interpenetrating sub-samples might be useful, since these error estimates include both the sampling and non-sampling errors, and the sub-sample estimates provide a confidence interval with a known confidence coefficient.

14.11b SUBJECT ANALYSIS

Subject analysis, to be taken up after preparing summary tables, should include cross-tabulation of data by meaningful geographical, economic, demographic, or other breakdowns to study the relationships and trends among the various characteristics, and comparison of the survey results with data available from other sources. This is a detailed technical analysis and is likely to be time-consuming. Hence this part should not be tied up with the first part, as otherwise the publication of the summary results might get delayed.

14.11c STATISTICAL ANALYSIS

By statistical analysis here we mean analysis, which would lead to improvements in statistical survey techniques, and it should include developmental studies for improving operational techniques for future surveys. For instance, this analysis should include the study of

variance and cost functions and non sampling errors at different stages of the survey. This type of analysis is also likely to be time consuming and hence should be taken up separately.

14.11d PLANNING OF PROCESSING WORK

Processing of survey data in a large scale survey involves detailed quality scrutiny of the data sorting of schedules or questionnaires, transcription of information, manual and mechanical computation, preparation and scrutiny of tables, etc and these needs advance planning and proper deployment of staff. Built in checks and cross checks on the computations are necessary for reducing non sampling errors that might occur at this stage of the survey. Suitable sampling methods may be used to assess and control the errors involved in different stages of the tabulation work. A good deal of planning is necessary to permit a smooth flow of the work material through the various stages of tabulation operation to ensure a fair degree of accuracy in the final results.

The first task preparatory to tabulation of data is to scrutinize the material and find out recording mistakes, inconsistencies and other defects. Serious defects should be rectified in consultation with the investigator or with the help of cross checks wherever possible, and by re visits to the sample units if necessary. The next task is to finalize the lay outs of the tables that are intended to be included in the reports to meet the objectives of the survey, and the tables originally conceived should be reviewed in the light of the experience gained during data collection and the quality of the data collected. Besides these tables, it would be desirable to provide for some tables to indicate the sample sizes, margins of error, and volume and total cost of survey operations under suitable breakdowns.

The blank tables prepared would also determine the nature of computations to be carried out involving the use of mathematical formulae and arithmetical operations. In the analysis of large scale data, it is profitable to use electro mechanical or electronic equipments. In that case it is necessary to code the data appropriately.

and transfer them to punch-cards or other suitable forms of input before they are fed into the machines for automatic processing and computation. If the tabulation is to be done manually with the help of a batch of computers (or clerks) the system followed will be naturally different. Suitable computational forms providing for all the steps involved in obtaining the final results should be prepared, with clear instructions for manual work with or without the use of desk calculators as the case may be.

Wherever possible computational checks should be introduced. Accuracy in numerical computation must be ensured, as the presence of numerical errors will affect the inferences drawn from the survey results. This means that the computations should be checked at each stage before proceeding to the next stage. Whatever may be the system of checking adopted, it is to be ensured that the checking is done satisfactorily. There can be several devices to attain this objective. For instance, before starting the work of comparison or the checking of a computation, some dummy mistakes may be introduced, and the efficiency of the checking operation can be judged by the extent to which the dummy mistakes are detected.

A record of time spent on various processing operations along with the volume of work completed should be maintained in a large-scale processing operation, as this helps in (i) studying the relative merits of different methods of data processing, (ii) the relative efficiencies of the staff engaged in this work, (iii) obtaining the cost of tabulation, and (iv) constructing suitable cost functions which can be used in planning future surveys. A detailed discussion of the construction of cost functions based on time record analysis is given by Mahalanobis (1944).

14.11e FRACTILE GRAPHICAL ANALYSIS

In tabulating the data from a survey, it would be useful to adopt the method of *fractile graphical analysis*, wherever necessary and practicable. This method consists in dividing the sample observations into a number of groups after arranging them in increasing order of -

a suitable basic variable (such as per capita consumer expenditure in a household expenditure survey) in such a way that the groups, known as *fractile groups*, are of equal content with respect to the estimated number of units in them, and then obtaining the value of characteristic of interest such as an average or a ratio, for each of these groups. If the values for the different fractile groups are plotted and if the successive points are joined, then a graph of the relationship between the basic variable and the characteristic under study is obtained and this is termed a *fractile graph* (Mahalanobis, 1960, United Nations, 1964). The procedure, besides being helpful in studying inter relationships between variables, permits comparisons of such inter relationship patterns being made over regions and over time.

14.12 PREPARATION OF REPORTS

The lines along which a general report on the results of a sample survey should be prepared are given by the United Nations (1949). Some points which would serve as guide lines in the preparation of sample survey reports are given below:

- (i) *Objectives* A clear indication should be given of the purposes of the survey and of the ways in which the results are proposed to be utilized.
- (ii) *Scope* An exact description of the scope of the survey should be included by specifying unambiguously the domains of study and the geographical regions covered by the survey.
- (iii) *Subject Coverage* A detailed description should be given of the items of information collected in the survey and if some of these items are not tabulated, the reasons for this should be mentioned. It will be convenient if a specimen schedule and a copy of the instructions are given in an appendix to the report.
- (iv) *Method of data collection* The method adopted in collecting data should be clearly described. All difficulties faced during

data collection and how they were overcome should also be explained.

- (v) *Survey, reference and reporting periods* : The time period during which the survey was carried out should be mentioned together with the reference and reporting periods adopted for different groups of items of information.
- (vi) *Sampling design and estimation procedure* : A clear description should be given of the sampling unit, sampling frame and sampling method adopted in the survey stating the sample size used and the method by which the sample size was determined prior to the survey. The estimation procedure should be explicitly mentioned by giving the formulae used in estimation.
- (vii) *Tabulation procedure* : It should be mentioned whether the data were tabulated manually or through the use of mechanical and electronic computing machines. The stages involved in tabulation may be mentioned together with any problems met with and how they were overcome.
- (viii) *Presentation of results* : The tabulated data should be presented in the form of comprehensive tables with unambiguous titles and clear column headings. Suitable illustrative diagrams, charts and graphs may be given to enable quick and clear comprehension of the survey results.
- (ix) *Accuracy* : A general indication of the accuracy attained in the survey results should be given. Results of tests and comparisons made to assess the accuracy of data and the reliability of the inference drawn should be included. The nature and extent of non-response and the way in which it has been treated in obtaining the final results should also be mentioned.
- (x) *Cost structure* : An indication of the total cost of the survey is to be given together with its breakdown under broad heads like preliminary work, sample selection, field investigation, processing of data, etc. Whenever possible, the cost should

also be given by more detailed breakdowns such as journey, enquiry, scrutiny, coding, punching, machine tabulation, etc. The man power and other resources utilized should also be mentioned.

- (ix) *Responsibility* The organizations sponsoring and conducting the survey should be mentioned. A brief description of the organizational set up of the agency carrying out the survey will also be useful.
- (xu) *Reference* A list of relevant published papers and reports should be given for reference.

14.13 INTEGRATED MULTI-SUBJECT SURVEY

When data on more than one subject of enquiry is required, the question arises as to whether all the subjects should be canvassed together in one survey, termed a *multi subject survey*, or in a series of separate surveys each dealing with one subject, known as *uni subject surveys*. In a multi subject survey, data on two or more subjects, not necessarily very closely related are collected in a single joint survey operation. This is generally considered to be more economical and operationally convenient than a series of uni subject surveys, provided the subjects of enquiry are not so numerous and diversified as to affect the quality of data (Lahiri, 1963, United Nations, 1964). In many situations, the nature of available sampling frames, accessibility of the sampling units, need for direct physical observation or a personal interview approach, etc favour the use of multi subject surveys.

Integration of enquiries relating to different subjects is important as it results in economy of resources available for survey work. An *integrated survey* may be defined as a survey, in which data on several subjects are collected for the same set of ultimate sampling units for studying the relationship among items belonging to different subject fields or within the same set of sample area units or other types of units with a view to achieving economy and operational

convenience. Such surveys are of considerable help especially in studies on levels of living. A question that needs careful consideration relates to the nature and the extent of integration of the different subject fields, which in turn would depend on a number of factors, such as the nature of subjects, method of enquiry, type of ultimate sampling unit, sampling frame and sampling design. For instance, (i) data on population and employment may be collected from a common set of sample households, since the sampling unit is the same and the method of enquiry is the interview method for both these subjects; and (ii) when data on consumer expenditure and on land utilization are required, one may consider the possibility of having a common set of sample area units and collecting data on the former from a sample of households by enquiry and on the latter from a sample of fields or farms by physical observation. In the former case, there is *complete integration*, whereas there is only *partial integration* in the latter.

The economy in a multi-subject survey, where two or more subjects of enquiry are integrated, arises mainly from savings in overhead cost which consists of setting up a survey organization with adequate administrative and supervisory staff, and in survey cost which comprises travel to the sample units, camp-setting, and contacting and surveying the ultimate sample units, etc. Economy in overhead cost is important not only in terms of money but also in respect of trained personnel. However, the economy in survey cost may not be considerable if the survey is carried out using *ad hoc* local staff. But use of *ad hoc* local staff may not be desirable except in some special situations, in view of the advantages of having a permanent staff described in the next section.

Since in large-scale surveys one usually resorts to multi-stage and multi-phase sampling designs, there would be a saving in survey cost achieved by integrating two or more subjects to form a multi-subject survey when the survey is carried out by whole-time staff. The savings achieved through integration may be used for increasing the number of sample units in the first stage (and in the subsequent

stages upto the penultimate stage if necessary) thereby improving the efficiency of the survey estimates since the contribution of the variation between first stage units to the total variance is generally more important than that between the units in subsequent steps Alternatively for the same precision increasing the sample size at the first few stages would generally lead to a reduction in the total number of ultimate stage sample units which in its turn would reduce the cost of tabulation The increase in the sample first stage units may also facilitate deeper stratification taking into account more than one auxiliary characteristic when available with a view to increasing the overall efficiency of the sampling design

Integration of enquiries with respect to the penultimate stage sample units makes the work of data collection more worth while in terms of economy in survey cost since the investigator is able to stay in those units for a longer time for canvassing data on different subjects than is possible in a uni subject survey This helps in obtaining data of better quality as a longer stay enables the investigator to become familiar with the local conditions and to establish cordial relations with the local people thus ensuring their cooperation in the survey Further the idle time would be less owing to the flexibility the investigator has in adjusting the work of data collection from different sample units if some units are temporarily not available for the enquiry or observation

Apart from the economy complete integration makes it possible to tabulate the data in meaningful cross classifications using relevant items of information belonging to different subject fields Further in carrying out socio economic surveys in countries where the economic and domestic activities of a household are rather mixed up especially in rural areas collection of data on all the activities would help in reducing response errors arising due to unconsciously mixing up of different types of activities

The planning and carrying out of multi subject surveys requires considerable expertise necessitating the use of highly trained and experienced survey personnel who are generally not easily available

Even the formulation of the concept of optimum sampling design in the case of a multi-subject survey is quite difficult and at every stage of sample designing, such as stratification, allocation, selection, etc. one has to adopt compromises between the needs of the various subjects involved. It may be mentioned that in spite of these difficulties, wherever feasible, it is desirable to have integrated multi-subject surveys, as they generally result in greater overall efficiency per unit of cost than several single-subject surveys carried out separately. However, if there are different agencies, which can collect the required information in their fields of specialization as a by-product of their normal activities or at marginal additional effort, the surveys may be conducted by the respective agencies in the form of different uni-subject surveys or multi-subject surveys covering fewer subjects.

14.14 PERMANENT SURVEY ORGANIZATION

A pre-requisite to efficient planning and successful execution of a large-scale survey is the existence of a permanent survey organization with well-trained survey personnel to do the work of planning the survey, data collection and processing of data. As the work of designing and carrying out a large-scale survey efficiently requires considerable skill derived from experience in this field, one of the primary advantages of a permanent survey organization is that it would be able to build up a competent and experienced survey staff. This staff should be able not only to conduct large-scale surveys but also to evolve a suitable programme for training the primary staff on the principles of sampling, methods of data collection and processing work. One of the main advantages of conducting a multi-subject survey through a permanent survey organization with whole-time staff as compared to carrying out of a series of uni-subject surveys with part-time or temporary staff is that this requires fewer trained personnel, which is a very important consideration in a number of situations.

The preliminary work involved in conducting a sample survey (evolving a suitable sampling design, sample selection, planning the work programme, preparation of schedules and instructions, and training the field staff) is considerable, information has to be obtained on cost and variance functions for different sampling designs and on the question of feasibility of collecting the required statistical data. A permanent survey organization would be in a position to carry out both theoretical and applied research relating to sampling design, methods of data collection analysis and presentation of survey results as part of its normal activities with a view to evolving survey procedures best suited to the requirements. Moreover, in such an organization, the experience of the survey personnel is not lost and hence they can continue to gain expertise in survey work.

A point deserving special mention is that a permanent organization is in a better position to assess and control non sampling errors. Considerable experience in survey work is required even to recognize the existence of non sampling errors and then to realize how frequent and large they can be and in what manner they vitiate the survey results. In fact, it would be rewarding for every survey organization to set apart some of its resources for studying on a continuing basis, problems involved in measurement and control of the errors, since experience has shown that non sampling errors arising from various sources, such as defective sampling frames faulty methods of data collection, and compilation, can, under certain circumstances, be substantially greater than the sampling errors.

A permanent survey organization with a whole time field staff in the different parts of a country or region would obviously be in a position to meet sudden demands for urgently needed statistical information on a particular topic of current interest. It is not uncommon for the users to require statistical data on some topic or other at short notice. From this point of view, it would be desirable for the survey organization to make its survey plan flexible enough to permit addition of *ad hoc* enquiries as demands arise.

One of the main difficulties generally faced in designing a sample survey is the lack of a suitable sampling frame. For instance, though the decennial population census is expected to provide a sampling frame for certain types of surveys, this frame may prove to be inadequate for sampling purposes in some cases, owing to the difficulties of identifying the units after the lapse of some time and also owing to changes in the frame over time. Since it would have a direct interest in evolving and maintaining satisfactory sampling frames, a permanent survey organization, can be expected to prepare comprehensive and reliable frames of well-defined, identifiable and compact area units or other types of units, and to set apart a part of its resources for up-dating the frame periodically.

REFERENCES

- BROOKS, E. M. (1953) : Planning and operating sample surveys; *Estadistica*, 11, 63-71.
- HANSEN, M. H. and GURNEY, M. (1946) : Problems and methods of the sample survey of business; *J. Amer. Stat. Assn.*, 41, 173-189, 46 (1951), 529.
- INDIAN STATISTICAL INSTITUTE (1952) : *National Sample Survey, General Report No. 1*, issued by the Department of Economic Affairs, Ministry of Finance, Government of India.
- KEMSLEY, W. F. F. (1952) : *The Social Survey—Some Technical Problems in Planning Budget Surveys*; Central Office of Information, London.
- LAHIRI, D. B. (1963) : Some thoughts on multi-subject sample survey system; *Contributions to Statistics*, 175-220, Presented to Professor P. C. Mahalanobis on his 70th Birthday, Statistical Publishing Society, Calcutta.
- MAHALANOBIS, P. C. (1944) : On large-scale surveys; *Phil. Trans. Roy. Soc., (B)*, 231, 329-451.
- MAHALANOBIS, P. C. (1960) : A method of fractile graphical analysis; *Econometrica*, 28, (2), 325-351, reprinted in *Sankhyā*, 23, (A), (1961), 41-64.

- UNITED NATIONS (1949) *Recommendations Concerning the Preparation of Reports on Sampling Survey* Statistical Papers Series C, No 1, New York revised in 1964, Statistical Papers Series, C, No 1, rev 2
- UNITED NATIONS (1964) *Handbook of Household Surveys* Studies in Methods Series F, No 10, New York
- UNITED STATES BUREAU OF THE CENSUS (1963) *The Current Population Survey—A Report on Methodology* Technical Paper No 7 Department of Commerce Washington D C
- UNITED STATES BUREAU OF THE CENSUS (1966) *Censuses of Population and Housing—A Procedural History*, U S Department of Commerce Washington D C
- ZARKOVICH, S S (edited by) (1965) *Estimation of Areas in Agricultural Statistics* Food and Agricultural Organization of the United Nations Rome

National Sample Survey

15.1 SCOPE OF SURVEY

This chapter describes in detail the sample design of the fourteenth round of the Indian National Sample Survey (NSS), which was carried out during the period July 1958–June 1959¹.

15.1a HISTORICAL NOTE

The NSS was initiated in 1950 to conduct sampling enquiries with a view to providing the Government and other organizations with socio-economic data, which can be used for planning for national development and for various research purposes. It is a continuing survey being carried out in the form of *rounds*, each round covering some topics of current interest. The rounds of the NSS have so far been of varying duration ranging from 3 to 8 months. Uptill the eleventh round, the gap between two successive rounds varied from 1 to 4 months, which period was utilized to re-train the field staff and to conduct some *ad hoc* or pilot surveys. Since the eleventh round, attempts were made to reduce the gap between two consecutive rounds so as to enable collection of data without appreciably missing any part of the year. The NSS completed 13 rounds of sampling

¹ This Chapter, prepared by the author and originally published as NSS report number 70 in 1962, is being given here with the permission of the Government of India, and the Government is not responsible for the views expressed herein.

enquiries by the end of May 1958 and the fourteenth round was conducted during the period July 1958 to June 1959²

15.1b SUBJECTS OF ENQUIRY

In the first few rounds, the emphasis was given on getting statistical information needed for the computation of national income and this related to statistics of consumer expenditure and small scale household enterprise. In the eighth round of the survey, the emphasis was shifted to the study of distribution of land holdings. The primary objective in the ninth round was the study of employment and unemployment situation in the country. In the tenth round of the survey, an exploratory study relating to the estimation of acreage and yield rates of cereal crops was undertaken along with the socio-economic enquiries. During the eleventh and the twelfth rounds the emphasis was on the study of economic conditions of agricultural labourers, such as their consumption, employment, wages and indebtedness. Stress was again laid on the crop survey in the thirteenth round.

15.1c METHODS OF ENQUIRY

In the NSS the work of data collection is done by specially trained quasi permanent full time investigators. The data on socio-economic characteristics are collected by investigators by personally interviewing the sample households or persons. In the case of crop survey, the acreage data are obtained for the selected plots (parcels of land) by direct physical observation and the yield rate is obtained by actually harvesting crop standing in randomly located circular cuts in sample plots. The reference periods for the different enquiries may be a day, week, month or a year depending on the characteristics under consideration.

² Since 1958-59 the rounds have been of one year duration and till July 1967, the NSS had completed 21 rounds of survey.

15.1d RESPONSIBILITY

The Central Statistical Organization, set up in 1951 by the Government of India to coordinate statistical activities of the different Ministries, State Governments and other statistical organizations in the country, is responsible for deciding the subject coverage and the methodology to be used in the NSS. The technical work relating to planning of the survey, formulation of the sample design, designing of schedules, writing of instructions and providing technical guidance to the field workers, processing and tabulation of the data and preparation of the final reports is done at the Indian Statistical Institute (ISI). The major portion of the field work for this large-scale sample survey is done by the Directorate of the NSS which is under the jurisdiction of the Cabinet Secretariat. The field work in West Bengal and Bombay City is being done by the field branches of the ISI.

15.1e MULTI-SUBJECT SURVEY

The NSS is a *multi-subject survey*³. Multi-subject surveys are generally recognised to be more economical than a series of uni-subject surveys. The main advantage of such surveys is that there is better utilization of the available resources especially when the time taken for the journey and camp-setting accounts for a considerable portion of the total time spent on the survey. Further, grouping of different subjects of enquiry in the same sample first stage units (sample villages and urban blocks) helps in increasing the sample size at the first stage. This means greater precision of the estimates than what would have been possible by conducting separate uni-subject surveys for each of the characteristics within the same total budget, because the variation between first stage units usually counts more than the variation within first-stage units for a number of characteristics. In other words, the precision attainable by separate uni-subject surveys may be achieved by incurring a much smaller expenditure in a multi-subject survey. It should, however, be mentioned that though

³ Section 9.11 of Chapter 9 (p. 347) and Section 14.13 of Chapter 14 p. 504).

multi subject surveys are generally more economical and more efficient than uni subject surveys the enquiries to be included in one survey should not be made so numerous and diversified as to over burden the investigators

15.1f REPORTING PERIOD

Indian economy being mostly dependent on agriculture is subject to pronounced seasonal fluctuation especially in the rural sector To take this seasonal factor into account it is desirable to make the survey period one complete year The survey periods in the rounds previous to the fourteenth round varied from 3 months to 8 months The survey period was made a complete year for the first time in the fourteenth round It may, however, be mentioned that the eleventh and the twelfth rounds together accounted for one complete year with little gap between the rounds

In the NSS, the practice has been to collect data from households on the basis of a moving *reporting period*, which is usually a week, month or year preceding the date of survey Thus the data collected do not usually refer to the fixed time period but refer to overlapping time periods of equal length This mode of survey will help in obtaining estimates of the averages of the characteristics over the period of the survey This is in contrast to a point survey where the estimates obtained refer to a point of time In an agricultural economy which is subject to considerable seasonal fluctuations, estimates based on a moving reporting period are likely to be more meaningful than those relating to a fixed point of time

For some characteristics, however, it may be desirable to collect the data for a fixed reporting period but this is not possible in a survey like the NSS due to the employment of a permanent moving field staff Because of this, the survey has to be spread over a period of time In such a situation it is not possible to have a fixed reporting period, for this would introduce recall bias particularly for the units surveyed during the end of the survey period long after the fixed reporting period has elapsed

15.1g INTERPENETRATING SUB-SAMPLES

One of the important features of the sample design of the NSS has been the use of independent interpenetrating sub-samples for studying the effect of sampling and non-sampling variation in the estimates. Usually the sample for any round is drawn in the form of two or more independent interpenetrating sub-samples and are usually surveyed by different investigators. Further, the data collected are also analysed by sub-samples by two different agencies (cf. Sub-section 15.1h). This procedure helps in analysing the total variation into its different components such as sampling variation, variation due to investigators, interaction between samples and the investigators, etc., (cf. Sub-section 13.10g of Chapter 13, p. 474).

15.1h PARTICIPATION OF STATES

In the eighth round all the State Statistical Bureaus participated in the field work of the NSS by doing the field work for two-thirds of the total sample for their respective States mainly with a view to increasing the sample size for the land holding survey. This State participation in the work of the NSS also helped in providing an overall check on the survey results. Both the central and the State agencies use the same set of schedules and instructions for data collection. Since the eighth round, the participation of the States in the work of the NSS has become a regular feature and at present almost all the States are participating both in the work of the NSS on a full-matching basis with the centre in the sense of doing the field and tabulation work in respect of as many units as covered by the central agency in their respective States. The total sample for each State is usually drawn in the form of 4 (or multiples of 4) interpenetrating sub-samples. Of these, 2 (or more) sub-samples are allotted to the State agency and the other sub-samples to the central agency. Within an agency the sub-samples are surveyed by two different parties of investigators.

15 II FOURTEENTH ROUND

Suggestions regarding the subjects to be undertaken in the fourteenth round were invited from the different Ministries of the Government of India and the States participating in the work of the NSS. A number of suggestions were received and these were carefully examined with a view to accommodating as many requirements of the requisitioning agencies as possible with the existing resources. Based on this a draft programme for the fourteenth round was prepared by the ISI and this was circulated to the members of the NSS Programme Advisory Committee. The draft proposals were considered in detail by a Working Group in its meeting held at Calcutta on 8th and 9th May 1958 and later by the Programme Advisory Committee in its meeting held on 12th May 1958 and they were accepted after certain modifications in the proposed schedules.

The subjects covered in this round were population crop income and expenditure small scale manufacture and handicrafts employment and unemployment in both the rural and urban sectors and population births and deaths village statistics crop statistics and retail prices only in the rural sector.

The geographical coverage for the survey was all India excluding Andaman and Nicobar Islands Amindive Laccadive and Minicoy Islands the North East Frontier Agency and the rural areas of Ladakh district in Jammu and Kashmir. The exclusion of these areas was necessitated due to cost and operational considerations.

In this round the States of Assam Bihar Bombay Kerala Orissa Punjab and Uttar Pradesh participated in the work of the NSS on full matching basis and the State of Andhra Pradesh participated on half matching basis. The sampling design and the programme of work for the State samples both in the rural and urban sectors were exactly similar to those for the central samples. Thus participation helped considerably in increasing the sample size for these States.

TABLE 15.1. ALLOCATION OF SAMPLE VILLAGES AND BLOCKS AND INVESTIGATOR REQUIREMENTS FOR THE CENTRAL SAMPLE.

sr. no.	State	number of sample		number of investigators (net)		
		villages	blocks	rural	urban	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1.	Andhra Pradesh	216	154	36	5	41
2.	Assam*	96	24	16	2	18
3.	Bihar	228	68	38	2	40
4.	Bombay ⁴	312	431	52	14	66
5.	Jammu & Kashmir	324	216	54	6	60
6.	Kerala	72	63	12	2	14
7.	Madhya Pradesh	252	88	42	3	45
8.	Madras	180	233	30	7	37
9.	Mysore	120	125	20	4	24
10.	Orissa	120	16	20	1	21
11.	Punjab ^{4**}	96	180	16	5	21
12.	Rajasthan	120	82	20	3	23
13.	Uttar Pradesh	312	264	52	8	60
14.	West Bengal	168	284	28	8	36
15.	total ⁵	2616	2228	436	70	506

* includes Manipur and Tripura ; ** includes Delhi and Himachal Pradesh.

TABLE 15.2. ALLOCATION OF SAMPLE VILLAGES AND BLOCKS AND INVESTIGATOR REQUIREMENTS FOR THE PARTICIPATING STATES.

sr. no.	State	number of sample		number of investigator (net)		
		villages	blocks	rural	urban	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1.	Andhra Pradesh†	108	77	18	3	21
2.	Assam	84	20	14	1	15
3.	Bihar	228	68	38	2	40
4.	Bombay ⁴	312	428	52	14	66
5.	Kerala	72	61	12	2	14
6.	Orissa	120	16	20	1	21
7.	Punjab ⁴	84	104	14	3	17
8.	Uttar Pradesh	312	264	52	8	60

† participation on half-matching basis.

⁴ Bombay has since been divided into the States of Maharashtra and Gujarat and Punjab into the States of Punjab and Haryana.

⁵ The sample size has subsequently been increased in the latter rounds to about 8500 villages and 4500 blocks in the central sample with a field staff of about 750 investigators (net).

TABLE 15.3 ALLOTTED SAMPLE SIZES FOR DIFFERENT ENQUIRIES
(CENTRAL SAMPLE)

sr no	schedule number	description	number of		households/plots*		
			villages	blocks	rural	urban	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1 1	income and expenditure	2616	2228	7848	2228	10076
2	2 2	small scale manufacture and handicrafts	2616	2228	27032	6684	33716
3	3 0	village statistics	2616	—	—	—	—
4	3 01	retail prices of selected commodities	872†	—	—	—	—
5	5 0	land utilization survey	2616	—	156960‡	—	156960
6	5 1	crop cutting experiments	872	—	5232‡	—	5232
7	10	employment and unemployment	2616	2228	10464†	8912	19376
8	12 1	population births and deaths enumeration	872	—	104640	—	104640
9	12 1 1	population births and deaths re enumeration	872	—	104640	—	104640
10	12 2	births and deaths enumeration	2616	—	313920	—	313920

* Household applies to schedules 1 1 2 2 10 12 1 12 1 1 and 12 2 and plot applies to schedules 5 0 and 5 1 † surveyed once in every two months, ‡ surveyed once in every crop season

15.2 RURAL SECTOR

In this section various aspects of the sample design of the rural sector are described

15.2a SUBJECT COVERAGE

The Working Group on Vital and Health Statistics of the Planning Commission, Government of India, desired that some exploratory work relating to the estimation of *rate of increase of population* and of *birth and death rates* should be taken up during the fourteenth

round of the NSS. It may be mentioned that no firm estimates of these rates, which are essential for the developmental planning of the national economy, were then available. To fill this gap in the Indian population statistics, it was decided to collect data on births and deaths in this round on an intensive scale.

The rate of natural increase in the population is obtained by taking the difference between the birth-rate and the death-rate. The rate of growth of population may be obtained by estimating the population at two points of time. For getting fairly good estimates for these rates, it was felt that complete enumeration of all the households in the sample villages would be desirable. For adopting the latter method of estimating the growth rate, it was necessary to survey the sample villages at two points of time with a sufficiently wide interval of time.

Information regarding the unorganized sector of economic activity, namely the small-scale and cottage industry, is of vital importance in formulating developmental plans. Hence, data on *household manufacturing enterprises* were collected with a view to throwing some light on this sector. As it is likely that there will be seasonal fluctuations, this information was collected over the seasons. It may be mentioned that this enquiry was taken up in some of the earlier rounds, the last round being the tenth round (December 1955–May 1956).

During the ninth round of the NSS (May–September 1955) data on *employment and unemployment* were collected. Since it would be of interest to compare the employment and unemployment situation with that observed three years ago, it was decided to have this enquiry in the fourteenth round. As there is seasonal fluctuation in the employment pattern especially in the rural areas, the data were collected over the seasons in this round. The employment data were collected from the same set of sample households once in every two months as this would make the study of changes in employment pattern over the seasons more effective.

An enquiry into the family budgets and employment pattern of agricultural labour households by the Ministry of Labour was undertaken by NSS during its eleventh and twelfth rounds (August 1956-July 1957). The weights to be used in the construction of the cost of living index for the rural agricultural labour population were obtained from this enquiry. To build up these indices the Ministry of Labour and Employment was in need of *retail price data* in rural areas. Hence, the rural retail price collection was continued in this round. Price data were collected from the same set of sample villages once in every two months in this round since it would make the comparison of the indices over time more effective.

The NSS has been collecting data on *household income and expenditure* since its inception. In this round also it was decided to take up this enquiry with a view to throwing some light on the seasonal variation in the expenditure pattern in the rural areas and to keep up the time series of estimates relating to the consumer expenditure pattern available from the first round onwards.

Since the tenth round of the NSS, intensive exploratory work has been undertaken with a view to estimating *crop acreage and yield rates* on the basis of an all India sample by the method of direct physical observation. A crop survey was undertaken in this round also for providing estimates of the acreage and production of all the cereal crops taken together at the all India level.

Besides the above enquiries, statistics relating to the availability of educational, medical and other social amenities in the sample villages were also collected. The subjects taken up for enquiry are shown in Table 15.4 along with their schedule numbers. Besides these schedules there were four other schedules 0 1, 0 11, 0 12 and 5 01 for listing and selection of households or plots for different enquiries. There was also a schedule (4 0) for keeping record of time spent on different survey operations by the investigators.

TABLE 15.4. SUBJECTS TAKEN UP FOR ENQUIRY (RURAL).

sr. no.	description	schedule number
(1)	(2)	(3)
1.	income and expenditure	1.1
2.	small-scale manufacture and handicrafts	2.2
3.	village statistics	3.0
4.	retail prices of selected commodities	3.01
5.	land utilization survey	5.0
6.	crop-cutting experiments	5.1
7.	employment and unemployment	10
8.	population, births and deaths enumeration	12.1
9.	population, births and deaths re-enumeration	12.1.1
10.	births and deaths enumeration	12.2

15.2b SURVEY PERIOD

The period of the survey was taken as one complete year, since the seasonal fluctuations were to be taken into account and studied in case of enquiries like employment, household enterprises and income and expenditure. As it is desirable to collect data from the same set of households, or at least from the same set of villages, over the seasons for estimating seasonal fluctuations, the sample villages had to be visited periodically. The period was taken as two months, since this would enable us to cover all the seasons which are generally of 3 or 4 months duration. Again, as there are very few crops which stand for less than two months, there was little chance of missing some crop if the sample villages were visited once in two months. Hence, this round consisted of six *sub-rounds* of two months duration and all the sample villages were visited in each of the six sub-rounds⁶.

⁶ Since 1960-61 (sixteenth round) the sample villages are generally visited only once for socio-economic enquiries and more often for crop survey.

15.2c FIXATION OF WORK-LOAD

The field staff available for the central sample was of the order of about 400 investigators together with the necessary supervisory staff. While planning this survey the strength of the existing field staff had to be borne in mind, as it would be difficult from the operational view-point to recruit and train the additional field staff in the short time that was available. Another consideration that was taken into account was the sample size required for giving fairly reliable estimates for the different characteristics that had been proposed to be surveyed. An idea regarding the sample sizes needed for the different enquiries was obtained on the basis of the results of the previous rounds of the NSS.

Fixing up of the work load for an investigator had to be done taking into consideration the sample sizes needed for the different enquiries, the operational difficulties involved and the time requirement for canvassing the various schedules. If one had choice regarding the number of investigators, then the procedure of fixing up the work load would have been done taking into consideration the period and the scope of the survey. In this way we would have determined the number of investigators required for this survey. Since this choice was not available it was the scope of the enquiry which had to be adjusted.

The problem regarding the fixation of the work load for investigators might be stated as follows. The number of investigators available for the rural survey was fixed (of the order of 400). The period of the survey was given to be one year. The number of villages⁷ to be surveyed for the most important enquiry in this round, namely population enquiry, was of the order of 2500⁸. The time requirements

⁷ Village is a well defined socio economic area unit consisting of households and plots (parcels of land). There are about 622000 villages in India.

⁸ This figure is based on empirical studies conducted on the basis of the data collected in the schedule used for listing households during the eleventh round (August 1956-January 1957).

for canvassing the different schedules were fixed on the basis of past experience. Taking into account all the above constraints, one had to determine how to achieve the maximum utilization of the available resources. In other words, how the scope of different enquiries was to be adjusted so as to get the maximum amount of data necessary for the purposes of national planning using the available field staff. It was not desirable to fix the work-load for the investigators on the basis of purely theoretical considerations and the approach had to be, of necessity, empirical—one of trial and success.

TABLE 15.5. AVERAGE TIME REQUIREMENTS FOR DIFFERENT SCHEDULES AND FOR JOURNEY BETWEEN VILLAGES.

sr. no.	schedule number	enquiry	time standard
(1)	(2)	(3)	(4)
1.	0.1	list of households	2 days/village
2.	1.1	income and expenditure	3 households/2 days
3.	2.2	small-scale manufacture and handicrafts	2 households/day
4.	3.0	village statistics	$\frac{1}{2}$ day/village
5.	3.01	retail prices of selected commodities	1 day/village
6.	5.0	land utilization survey	3 clusters of 10 plots/day
7.	5.1	crop-cutting experiments	2 cuts/day
8.	10	employment and unemployment	6 households/day
9.	12.1	population, births and deaths enumeration	10 days/village
10.	12.1.1	population, births and deaths re-enumeration	5 days/village
11.	12.2	births and deaths enumeration	5 days/village
12.		journey	2 days/village

As there were about 400 investigators and about 2500 villages to be surveyed, each investigator had to survey six villages. Then, the problem reduced to that of fixing up work load for only one investigator in these six villages. The number of working days in a year for an investigator was 280 days (number of Sundays 52, number of public holidays 18, casual leave 15 days)⁹. Privilege or earned leave was not considered here because there is a reserve of 10% investigators in the field to take account of such eventualities. The time requirements for the various schedules based on the experience of the field staff during the thirteenth round in the different States are given below. In case of population schedules (12 series), the time requirements were based on a try out of these schedules in the field.

It would have been desirable to fix the work load on the basis of the time requirements in the different States and if possible even at lower levels since there is considerable variation in time requirements from region to region. But fixation of work load was done on the basis of the average all India time requirements, because there were a number of other considerations influencing the work load of the investigator such as communication facilities, weather, etc which could not be taken into account particularly owing to lack of factual information. It may be pointed out that the work load for a particular investigator depends much on the actual sample villages, households and plots.

It was considered desirable to integrate the socio-economic and crop surveys, that is, to conduct both the enquiries in a common set of villages. This helped in reducing the time taken for journey and camp setting and hence the number of sample villages could be increased for all the enquiries. This meant greater precision for estimates since usually the contribution to the total variation from sampling villages is large for many characteristics.

* In the subsequent rounds the number of working days is taken as 270.

To study seasonal fluctuation, it would have been desirable to collect the information on consumer expenditure and household enterprise from the same set of sample households in each of the sub-rounds. But this could not be done since the field staff had experienced some resistance on the part of informants in giving the detailed information required in this schedule more than once in a short period of time.

Considering the sample size desired for different enquiries, initially it was proposed to have the following programme of work : collection of data on population, births and deaths from all the sample villages in the first and the last sub-rounds, on prices in one-third of the sample villages in each sub-round, on village statistics in all the villages in any one sub-round, and on land utilization in all the sample villages and yield survey in one-third of the sample villages in each crop season, and canvassing of schedule 1.1 for 1 household, schedule 2.2 for 2 households and schedule 10 for 4 households per sample village in each sub-round.

The total number of working days required for the above programme of work was 330 as against the 280 working days available for an investigator. Hence this programme was found to be impracticable. Further the work-load was not evenly spread out over the six sub-rounds. The work-load in the first and the last sub-rounds was very heavy compared with that in the other sub-rounds. The above programme of work, therefore, needed modification in two directions. The overall work-load had to be reduced and it had to be spread evenly over the six sub-rounds. Even-spreading of the work-load over the sub-rounds was necessary to ensure equal duration of the sub-rounds and a fairly constant time interval between two successive investigations of the same village.

As a large portion of the time was taken up by schedule 12.1 (population), the work-programme relating to this schedule was examined carefully with a view to achieving substantial reduction in the overall work-load and even-spread of the work over sub-rounds.

After this examination it was decided to canvass schedule 12.1 in only one sample village per investigator in each of the first two sub rounds and to take up schedule 12.1 for these two villages in the last two sub rounds one in each sub round. Schedule 12.1 was simplified by omitting the detailed person by person enumeration to form schedule 12.2. This schedule was meant to be canvassed in one sample village in each of the last four sub rounds. These changes in effect meant that we would be getting data on births and deaths from all the six villages allotted to an investigator at the rate of one village per sub round and that the population count at two points of time could be done only in one third of the sample villages the gap between the two points being about eight months. This modification of the work programme helped considerably in solving the problem of work load mentioned above.

The following were some of the other modifications in the work programme. The sample size for schedule 1.1 was reduced by half and that for schedule 2.2 was reduced to 10 from 12 households for an investigator in each sub round except the first. It was decided to collect village statistics in all the sample villages in the third sub round instead of the first sub round which had a heavy work load even after this modification. It may be pointed out that the work load for crop survey has been calculated on the basis of three crop seasons whereas in many of the States there would perhaps be only two seasons. Hence it was expected that the time requirement taken at the planning stage would be adequate.

The final programme of work for an investigator in the rural sector in the different sub rounds of this round was arrived at by adopting the method of trial and success with a view to getting the maximum amount of the data with the available resources. Tables 15.6 and 15.7 give the finalized work load for an investigator and the time requirements for the different schedules. The six villages allotted to an investigator were numbered from 1 to 6 according to the order of selection.

TABLE 15.6. SAMPLE VILLAGES TO BE SURVEYED FOR DIFFERENT SOCIO-ECONOMIC ENQUIRIES BY AN INVESTIGATOR.

sr. no.	sche- dule num- ber	sub-round											
		1		2		3		4		5		6	
		s. v.	days	s. v.	days	s. v.	days	s. v.	days	s. v.	days	s. v.	days
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1.	0.1	1-6	10†	—	—	—	—	—	—	—	—	—	—
2.	1.1	1,3,5 (3)	2	1-6 (3)	2								
3.	2.2	1-6 (12)	6	1-6 (10)	5								
4.	3.0	—	—	—	—	1-6	3	—	—	—	—	—	—
5.	3.01	2,5	2	2,5	2	2,5	2	2,5	2	2,5	2	2,5	2
6.	10	1-6 (24)	4	1-6 (24)	4	1-6 (24)	4	1-6 (24)	4	1-6 (24)	4	1-6 (24)	4
7.	12.1	1	10	4	10	—	—	—	—	—	—	—	—
8.	12.1.1	—	—	—	—	—	—	—	—	1	5	4	5
9.	12.2	—	—	—	—	2	5	5	5	3	5	6	5
10.	journey	1-6	12	1-6	12	1-6	12	1-6	12	1-6	12	1-6	12
11.	total	—	46	—	35	—	33	—	30	—	35	—	35

s. v. : Sample Village.

† In sample village 1, the information in schedule 0.1 was obtained along with the data required for schedule 12.1.

(The figures in brackets denote the total number of sample households to be investigated in the case of household schedules).

Total number of working days required for socio-economic enquiries : 214.

TABLE 15.7 SAMPLE VILLAGES TO BE SURVEYED
FOR CROP SURVEY BY AN INVESTIGATOR

sr no	schedule number	sample villages*	number of days*
(1)	(2)	(3)	(4)
1	5 0	1-6 (360)	12
2	5 1	3 6 (12)	6
3	journey (revisits)	3 6	4
4	total	—	22

*In each of three crop seasons. The figures in the brackets denote the total number of plots to be surveyed

Total number of working days required for crop survey 66

Total number of working days required for the modified programme $280 = (214 + 66)$

15.2d PROGRAMME OF WORK

In each stratum the same two investigators were working throughout the survey period each investigating a sub sample of six villages in each of the sub rounds. The six sample villages allotted to an investigator were numbered from 1 to 6 in the sample list according to their order of selection and were visited in that order in each sub round. It was expected that the sample villages 1, 2, 3 would be investigated in the first month and the sample villages 4, 5, 6 in the second month of each sub round. The work specified for one sub round had to be finished in that sub round itself as far as possible. In each agricultural season the land utilization survey was carried out in all the sample villages and the crop cutting experiments in sample villages 3 and 6.

If the work of the first sub round could not be completed in the first sub round period (first two months), then attempts were made to complete the work specified for the first and the second sub round.

by the end of the second sub-round period (in the first four months). Similarly if the work specified for the first two sub-rounds could not be finished in time, then attempts were made to complete the work of the first three sub-rounds by the end of the third sub-round period (in the first six months) and so on. The work of any particular sub-round had to start immediately after the completion of work of the previous sub-round, but not earlier than ten days preceding the starting date of that sub-round period.

15.2e SAMPLE DESIGN

Complete integration of the socio-economic and crop surveys was achieved by selecting the villages circular systematically with equal probability after proper stratification and arrangement. In other words, the land utilization and yield surveys as well as the various socio-economic enquiries were undertaken in a common set of villages. This integration helped in getting a considerably larger sample size for both the surveys than that would have been possible otherwise with the necessary sub-round restrictions discussed in Sub-section 15.2b. It is expected that the loss of efficiency in having equal probability sampling as compared to varying probability selection could be offset, at least partially, by using the method of ratio estimation at the tabulation stage.

The general sampling design was stratified two-stage one, in which the villages were the first stage units and households and clusters of plots formed the second stage units for socio-economic enquiries and crop survey respectively. In the case of yield survey, crop-plot¹⁰ and circular cuts in them formed the third and fourth stage units respectively. The strata were formed by grouping contiguous tehsils¹¹, which were homogeneous with respect to 1951 census population density, altitude above sea-level and food crops, and equalizing strata populations as far as possible within each State. From each stratum

¹⁰ Plots growing one or more of the specified cereal crops.

¹¹ Tehsil is an administrative unit consisting of villages and a few towns. There are about 2500 such units in India.

2 circular systematic samples of 6 villages were selected with independent random starts after arranging the tehsils according to geographical contiguity to allow for interpenetration of investigators at stratum level. Such interpenetration helped in obtaining a quick estimate of the total error of the estimate including the differential non sampling errors. Within each selected village, the required number of households were selected from all the households in it systematically with a random start for the different socio-economic enquiries after some suitable arrangement of the households. For the land utilization survey the required number of clusters of plots were selected systematically from the selected villages. In one third of the villages, crop cutting experiments were conducted for the cereal crops.

15.2f ALLOCATION OF SAMPLE VILLAGES

The sample size for this round in the rural sector was about 2600 villages. These were allocated to the States on a joint consideration of their population, geographical area, crop acreage and the number of persons engaged in household enterprise obtained on the basis of the 1951 census. Special weight was given to the State of Jammu and Kashmir in the final allocation as separate estimates were required on the basis of the central sample for that State.

As different languages are being spoken in different regions, it was not feasible to transfer the investigators from one region to another. Recruitment or discharging of investigators at short notice was also difficult. Hence in making the allocation, the existing investigating strength in the different administrative blocks of the NSS was also taken into consideration. The allocations were finally rounded off to multiples of 12 with a view to allowing 2 investigators to work in each stratum.

Within a State, the allocation of the sample size to the strata was in proportion to the stratum populations. This, together with the need for equal work load in each stratum, made it necessary to form strata, each having approximately the same population.

TABLE 15.8. ALLOCATION OF SAMPLE VILLAGES TO THE STATES ON THE BASIS OF DIFFERENT CRITERIA.

sr. no.	State	allocation of villages proportional to						
		rural popu- lation	geogra- phical area G	cultivat- ed area GP	$G\sqrt{PQ}$	persons in ind- ustry	no. of investi- gators	final alloca- tion
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1.	Andhra Pradesh	208	201	218	220	317	181	216
2.	Assam*	80	186	129	118	31	85	96
3.	Bihar	292	128	205	140	101	204	228
4.	Bombay	279	362	339	406	252	318	312
5.	Jammu & Kashmir	20	176	63	65	†	108	324‡
6.	Kerala	95	28	55	31	139	57	72
7.	Madhya Pradesh	185	325	261	346	158	255	252
8.	Madras	183	95	141	106	204	181	180
9.	Mysore	121	142	139	157	108	159	120
10.	Orissa	113	115	121	123	113	114	120
11.	Punjab**	116	111	119	114	132	114	96
12.	Rajasthan	105	251	173	262	134	136	120
13.	Uttar Pradesh	441	216	328	239	545	312	312
14.	West Bengal	162	64	109	73	166	176	168
15.	total	2400	2400	2400	2400	2400	2400	2616

*includes Manipur and Tripura; **includes Delhi and Himachal Pradesh.

† Data on number of persons engaged in household industry were not available.

‡ This is due to subsequent increase in field strength for providing better estimates for this State.

(Allocation in proportion to $G\sqrt{PQ}$ is considered, since this has been empirically found to be a good approximation to the optimum allocation for estimating cultivated area).

15.2g STRATIFICATION

In 1957 the Ministry of Education and Scientific Research published the *Indian National Atlas* giving a variety of maps showing population density, relief, food crops, etc. In each State, strata were formed by grouping contiguous tehsils, which are homogeneous with respect to population density, altitude above sea-level and cultivation of food crops. At the same time the population of each stratum

was made approximately the same. The number of strata in a State was taken as one twelfth of the allocation of sample villages for that State, since two investigators had to be posted in every stratum, each surveying a sub sample of six villages, to take account of the differential investigator bias also in the calculation and analysis of the total error in the estimates.

The above system of stratification was adopted as it was felt that strata so formed would be more or less homogeneous for a number of subjects of enquiry that were undertaken in this round. This survey, being a multi subject one had to take a number of criteria into account in stratification and consequently this system of stratification may not necessarily be optimum for certain specific enquiries. The operational procedure followed at the time of stratification is given below.

From the National Atlas State district, and tehsil boundaries of every State were traced out. This outline of each State was superimposed on the map showing altitude above sea level and the tehsils in the State were classified into three strata on the basis of altitude above sea level. The sketched out map of that State was then superimposed on the map showing the population density and the tehsils were classified into three density groups, formed in such a way that approximately one third of the tehsils fell in each group. For big States like Uttar Pradesh the map showing food crops was also used to form 2 or 3 groups of the tehsils on the basis of the main crops grown there. Thus all the tehsils in each State were classified into 9, 18 or 27 classes. A specimen of a table showing the classification of the tehsils in a State is given in Table 15.9.

After getting this two or three way table the tehsils were arranged in a *serpentine order* with bonds at as short intervals as possible in such a way that any two consecutive tehsils in the list were contiguous on the map and were roughly homogeneous with respect to population density, altitude above sea level and food crops. This type of serpentine arrangement was used so that, as far as possible, compact strata could be formed at the time of demarcation of strata. At the

TABLE 15.9. SPECIMEN WORKING SHEET USED FOR CLASSIFYING THE TEHSILS IN UTTAR PRADESH INTO DIFFERENT CLASSES.

persons per sq. km.	food crop	altitude above sea-level in metres		
		less than 150	150-300	above 300
(1)	(2)	(3)	(4)	(5)
below 80	rice	k 3, 4.	—	—
	wheat	—	Q 1, 2, 3, 4; V 2; x 2, 5, 6.	B 3; K 2, 3; Q 5, 6; T 1, 2; x 4.
	others	—	U 4.	B 1, 2, 4; K 1; R 1, 2, 3, 4; U 2, 3.
80-240	rice	F 2, 6, 8, e 1, 3; k 1, 2; s, 4.	—	—
	wheat	E 3, 4; G 3, 4; I 1, 2, 3, 4, 5, 6; J 2, 3; O 1, 3; X 1, 2, 3; Y 1, 3, 4; f 1, 2, 3, 4, 5; r 2; t 4; v 1, 2, 3; x 1, 3; y 1, 2, 3, 4.	A 1, 4; C 1, 2, 5, 6, 7; E 1, 2, H 1, 2, 3; J 1; O 2, 4; V 1, 3; Y 2; a 1, 4, 5; b 1, 3, 5, h 1, 2, 3, 4; i 1; j 1, 2, 3, 1 2, 4; m 2, 3, 4; o 1, 2, 3, 4, 5; p 1, 2, 3, 4, 6; t 1, 2, 3; u 2, 4.	u 1.
	others	—	U 1.	—
above 240	rice	D 1, 2, 3, 4, 5, 6; F 1, 3, 4, 5, 7; L 1, 2, 3; M 1, 2, 3; N 1, 2, 3, 4; P 1, 2, 3, 4, 5; S 1, 2, 3, 4; W 1, 2, 3; Z 1, 2, 3, 4; c 1, 2, 3; d 1, 2, 3, 4, 5; e 2; q 1, 2, 3, 4; s 1, 2, 3; w 1, 2, 3, 4.	—	—
	wheat	G 1, 2; g 1, 2, 3, 4; r 1, 3; v 4.	A 2, 3, 5, 6; C 3, 4; H 4; a 2, 3; b 2, 4; i 2, 3, 4; j 4, 1 1, 3; m 1, 5, 6; n 1, 2, 3, 4, 5, 6; p 5; u 3.	—
	others	—	—	—

(Alphabet codes stand for districts and numeral codes as given in National Atlas for tehsils within districts).

A : Aligarh; B : Almora; C : Agra; D : Azamgarh; E : Etawah; F : Allahabad;
G : Unnao; H : Etah; I : Kannur; J : Kheri; K : Garhwal; L : Ghazipur;
M : Gonda; N : Gorakhpur; O : Jalaun; P : Jaunpur; Q : Jhansi; R : Tehri-
garhwal; S : Deoria; T : Dehradun; U : Nainital; V : Pilibhit; W : Pratapgarh;
X : Fatehpur; Y : Farukhabad; Z : Faizabad; a : Budaun; b : Bareilly; c : Ballia;
d : Basti; e : Bahraich; f : Banda; g : Barabanki; h : Bijnor; i : Bulandshahr;
j : Mathura; k : Mirzapur; l : Muzaffarnagar; m : Moradabad; n : Meerut;
o : Mainpuri; p : Rampur; q : Raebareli; r : Lucknow; s : Varanasi; t : Shah-
jahanpur; u : Saharanpur; v : Sitapur; w : Sultanpur; x : Hamirpur; y : Hardoi.

time of arranging the tehsils top priority was given to contiguity, next priority to population density, third priority to the altitude above sea level with food crops getting the last priority. This order of priority was given with a view to having compact investigation zones (Sub section 15 2h) and because of the importance attached to getting good estimates of population growth rate. It is expected that arrangement with respect to altitude and population density would help in having a good stratification for at least major crops. After arranging tehsils required number of strata were formed by grouping consecutive tehsils and equalizing the strata populations¹².

TABLE 15 10 DENSITY AND ALTITUDE CLASSES USED FOR CLASSIFYING THE TEHSILS IN EACH STATE FOR STRATIFICATION

sr no	State	density class population per square kilometre			relief class altitude above sea level in metres		
		1	2	3	1	2	3
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	Andhra Pradesh	— 40	40—100	160—	— 75	75—300	300—
2	Assam*	— 40	40—100	100—	— 300	300—	—
3	Bihar	—160	160—400	400—	— 75	75—300	300—
4	Bombay	— 60	60—120	120—	—150	150—600	600—
5	Madhya Pradesh	— 40	40—80	80—	—300	300—600	600—
6	Madras	— 40	40—160	160—	— 75	75—300	300—
7	Mysore	— 80	80—120	120—	—300	300—600	600—
8	Orissa	— 80	80—160	160—	—300	300—600	600—
9	Punjab**	— 80	80—160	160—	—300	300—	
10	Rajasthan	— 20	20—60	60—	—300	300—600	600—
11	Uttar Pradesh	— 80	80—240	240—	—150	150—300	300—
12	West Bengal	—200	200—300	300—	—300	300—900	900—

* includes Manipur and Tripura

** includes Delhi and Himachal Pradesh

Since the sample size for Jammu and Kashmir was large, it was possible to consider each tehsil or a group of two contiguous tehsils as a stratum. It may be mentioned that the list of tehsils given in

¹² Since 1960 (the sixteenth round) information on transport and communication facilities between adjacent tehsils is also taken into account at the time of stratification.

the National Atlas for Kerala did not tally with that available in the NSS frame, which is based on the district census handbooks. Hence, the stratification in Kerala was done on the basis of the maps given in the census volumes. Even in some of the other States it was not possible to identify a few tehsils listed in the census records on the map. In such cases the list of tehsils as given in the census records was adopted, since the details regarding the villages in the tehsils, included in the National Atlas but excluded from the census records, were not available.

15.2h INVESTIGATION ZONES

Since in this round the strata were formed by grouping tehsils which were homogeneous with respect to population density, the area of some strata happened to be large. In such cases the strata were further sub-divided into compact *investigation zones*, each of

TABLE 15.11. DISTRIBUTION OF STRATA BY NUMBER OF INVESTIGATION ZONES.

sr. no.	State	total no. of strata	number of strata with investigation zones						no. of zones ¹³	
			1	2	3	4	5	6	total	n.s.†
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1.	Andhra Pradesh	18	10	7	1	—	—	—	27	3
2.	Assam*	8	5	—	2	—	1	—	16	5
3.	Bihar	19	15	4	—	—	—	—	23	1
4.	Bombay	26	10	14	2	—	—	—	44	10
5.	Jammu & Kashmir	27	27	—	—	—	—	—	27	—
6.	Kerala	6	6	—	—	—	—	—	6	—
7.	Madhya Pradesh	21	5	13	3	—	—	—	40	10
8.	Madras	15	15	—	—	—	—	—	15	—
9.	Mysore	10	3	7	—	—	—	—	17	2
10.	Orissa	10	5	4	1	—	—	—	16	3
11.	Punjab**	8	4	3	1	—	—	—	13	1
12.	Rajasthan	10	1	6	1	1	—	1	26	10
13.	Uttar Pradesh	26	23	2	—	1	—	—	31	3
14.	West Bengal	14	14	—	—	—	—	—	14	—
15.	total	218	143	60	11	2	1	1	315	48

* includes Manipur and Tripura; ** includes Delhi and Himachal Pradesh;

† not selected.

¹³ With increase in sample size and number of strata in the subsequent rounds, the number of investigation zones not selected has become very small.

which had an area not greater than 6000 square miles with a view providing approximately the same geographical coverage for each investigator. In all other cases the stratum itself was taken as investigation zone. Thus each stratum consisted of one or more investigation zones.

In strata which formed single investigation zones two sub samples of six villages each were selected systematically with independent random starts. In cases of strata, which consisted of two or more investigation zones, two investigation zones were selected with probability proportional to the number of villages in them and with replacement, and from each selected investigation zone, a sample of six villages was chosen systematically with a random start. Then in each stratum two independent sub samples of six villages each were selected and these were surveyed by two different investigators.

15.2i SELECTION OF VILLAGES

The sampling frame being used in the NSS is the 1951 census list of villages. In the first few rounds, the NSS field staff had compiled the frame from the census records available at tehsil or district level. This frame was in manuscript form and using of this frame had become more difficult with the passage of time. This frame was being replaced by that given in the district census handbooks as soon as they became available. For this round, sampling was done using the district census handbooks in about 90% of the districts and for the remaining districts the manuscript frame was utilized.

After having formed the investigation zones, the work of selecting of villages was taken up. For this purpose the number of villages in each of the tehsils was found out and these figures were posted against the respective tehsils. While finding the number of villages in each tehsil from the district census handbook or from the manuscript records, it was not possible to take it as the highest serial number, since the list included certain duplications, areas declar-

¹⁴ Since 1962-63 (eighteenth round) the 1961 census frame is used in the N.S.S.

urban, etc. The number of duplications, blank serial numbers and villages declared urban had to be subtracted from the highest serial number to get the total number of villages. Adjustments were also made in the case of villages transferred from one tehsil to another.

Within an investigation zone, the arrangement of tehsils adopted for demarcation of strata was retained for selecting the villages. Let the corrected number of villages in the tehsils so arranged in a particular stratum (investigation zone) be

$$N_1, N_2, \dots, N_k.$$

For these tehsils, the cumulative totals

$$T_1 (= N_1), T_2 (= T_1 + N_2), \dots, T_k (= T_{k-1} + N_k)$$

were found. A random start was taken from 1 to T_k (say R_1). With this random start, a circular systematic sample of six villages was selected with the interval I , the integral part of $T_k/6$.¹⁵ This constituted sub-sample 1 of the central sample. A similar sample, drawn with an independent random start (R_2), constituted sub-sample 2 of the central sample. The second sub-sample was selected from the same or another investigation zone. The actual procedure of selection adopted is illustrated by the following example. The actual serial number in the census handbooks, which was to be selected, was the serial number shown in column (7) of Table 15.12 increased by the number of serial numbers less than or equal to that number which have been excluded from the frame.

For the State sample, the same procedure was adopted except that the random start used was $R_1+I/2$ for sub-sample 1, and $R_2+I/2$ for sub-sample 2. Corresponding central and State sub-samples are not independent because of this procedure of linking. This linking ensures better representation of the geographic spread of villages in each stratum of the participating States. The two sub-samples are independent for both the central and State samples.

¹⁵ In the later rounds the six villages are selected pps systematically, size being related to the 1961 census population.

Ordinarily the boundary of the census village coincides with its revenue boundary but there may be instances, although infrequent, where the census boundary of a village does not tally with its revenue boundary. It may be noted that the unit of survey for the socio-economic survey was the census village whereas it was the revenue village for the crop survey¹⁸.

TABLE 15.12 PROCEDURE OF SELECTING A SAMPLE OF SIX VILLAGES FROM A STRATUM

zone North India			State Punjab			Stratum 2	
sr no	district	tehsil	no of villages	cumulated number of villages	selected numbers	selected village sr no	order of selection
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	Jullunder	Nakodal	285	285	87	87	2
2	Kapurthala	Kapurthala	431	716	667	382	3
3	Jullunder	Jullunder	382	1098			
4	Hoshiarpur	Hoshiarpur	488	1586	1247	149	4
5	Hoshiarpur	Una	519	2105	1827	241	5
6	Kangra	Hamirpur	64	2169			
7	Kangra	Desagohupur	143	2312			
8	Kangra	Nurpur	190	2502	2407	95	6
9	Gurdaspur	Pathankot	392	2894			
10	Hoshiarpur	Dasuya	588	3482	2989	95	1

(sampling interval $I = 580$ random start R from 1 to 3482 = 2989)

15.2j INTERPENETRATING SUB-SAMPLES

As mentioned earlier, the sample of villages was drawn in the form of two independent and interpenetrating sub samples to be surveyed by two different sets of investigators. In each stratum, the same 2 investigators carried out the survey, each surveying a

¹⁸ Since 1961-62 (seventeenth round) the revenue village is taken as the unit of survey for socio-economic survey also.

sub-sample of 6 villages. The State sub-samples were linked with the central sub-samples in a systematic manner to get better spread of the sample and also to make the study of the differential agency bias more effective. Thus in each stratum of the participating States, there were 4 interpenetrating sub-samples, 2 of which were surveyed by the central agency and the other 2 sub-samples by the State agency¹⁷. Further, the round was divided into 6 sub-rounds of 2 months each. Almost all the socio-economic enquiries were spread over the 6 sub-rounds in a random manner so as to enable the study of the differential effect of time on the characteristics under consideration. It may be noted that this procedure of interpenetrating of the sample over agencies, investigators and time would enable us to analyse the total variation of the estimates into its components such as variation due to (i) agencies, (ii) investigators, (iii) interaction between (i) and (ii).

15.2k HAMLET-GROUP SELECTION

The number of households varies widely from village to village. The following procedure was adopted to ensure uniformity in work-load within a village. In some of the big villages consisting of well-defined hamlets, the hamlets were grouped to form the required number of groups having approximately the same population and one such group was selected at random with equal probability. The number of hamlet-groups to be formed in a particular sample village was worked out such that an investigator would not have to list more than 800 households taking together all the six villages assigned to him. For this purpose, the population figures of the sample villages in 1951 were used, as current population figures were not available at the time of planning this survey. If the total 1951 census population in the 6 villages in a sub-sample of a stratum was

¹⁷ Since 1962-63 the overall sample is drawn in the form of 8 interpenetrating sub-samples, out of which 4 sub-samples are surveyed by the central agency and the other 4 sub-samples by the State agency.

less than 3500 hamlet-group selection was not allowed in any of the 6 villages

The number of hamlet groups to be formed in a sample village was given in the list of sample villages. If the investigator found on local enquiry that the population of the village had changed considerably since the 1951 census he was allowed to change the number of hamlet groups to be formed on the following lines

- (i) If on visiting a sample village the investigator found that its present population was more than or equal to k times and not more than $k+1$ times of what it was in 1951, the number of hamlet groups to be formed was taken as k times the number specified in the sample list,
- (ii) If the present population of a sample village for which hamlet group selection has been indicated in the sample list was less than half of what it was in 1951 the number of hamlet groups to be formed was the quotient obtained on dividing the specified number of groups by two

In a village where hamlet group selection was to be resorted to the hamlets were listed in the alphabetical order of their names. If the number of hamlet groups to be formed was L then the groups were formed by taking consecutive hamlets in the list such that the percentage of total population covered in each group is $100/k$ as far as possible

In case the investigator found it convenient to form hamlet groups by grouping contiguous hamlets he was allowed to proceed in that manner. After forming such groups the hamlets within each group were arranged in the alphabetical order of their names. These groups were then arranged in the alphabetical order of the names of the hamlets coming first in the within group arrangement.

If there were no hamlets in a village where hamlet group selection was found necessary, the investigator was allowed to use the blocks

(groups of households) formed, if any, by the census authorities or to form the required number of area units himself, by sub-dividing the village in a suitable manner.

After the formation of the required number of hamlet-groups, one such group was selected with equal probability. The socio-economic survey was to be confined only to the selected hamlet-group.

15.2l DIVISION OF SAMPLE VILLAGE

As mentioned earlier, the revenue village corresponding to the selected census village was taken as the sample unit for the crop survey. In case the selected census village consisted of more than one revenue village, one of the revenue villages was selected with equal probability. If the census village was a part of a revenue village, the entire revenue village was taken up for crop survey.

If the area of a sample village was greater than 16 square miles and consisted of a number of hamlets having separate survey maps (or village records), the hamlets were grouped to form hamlet-groups of not less than 2 square miles in area and one such hamlet-group was selected with equal probability. In case hamlet-wise village maps were not available, but the village area was mapped in two or more map sheets, one of the map sheets (or a group of the map sheets each with area not less than 2 square miles) was selected with equal probability. In case the village map was available on one sheet, a sub-sheet having an area of at least 2 square miles formed by folding the map sheet into different parts was selected with equal probability. In the above-mentioned cases, the crop survey was confined to the selected hamlet or hamlet group or part of the village covered by the selected map sheet or sub-sheet.

In villages not cadastrally surveyed, where village maps or alternative records were not available, plots (parcels of land) possessed by a sample of households were surveyed. In such cases, crop survey

TABLE 1613 DISTRIBUTION OF SAMPLE VILLAGES BY NUMBER OF HAMLET GROUPS

sr no	State	number of hamlet groups																		surv eyed vill	casu alty vill	allo cated vill
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	(17)	(18)	(19)	(20)	
1	Andhra Pradesh	126	41	13	10	8	4	6	3	1	1	-	1	1	-	1	1	-	216	-	216	
2	Assam*	89	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	91	6	96	
3	Bihar	195	18	4	6	2	-	2	-	-	1	-	-	-	-	-	-	-	228	-	228	
4	Bombay	251	40	4	8	1	2	2	1	-	-	-	-	-	-	-	-	-	309	3	312	
5	Jammu & Kashmir	310	6	-	2	1	-	-	-	-	1	-	-	-	-	-	-	-	319	5	324	
6	Kerala	20	14	10	3	8	4	4	3	1	2	1	-	1	-	1	1	72	-	72		
7	Madhya Pradesh	240	4	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	252	-	252	
8	Madras	86	40	17	10	10	3	3	5	2	2	-	1	1	-	-	-	180	-	180		
9	Mysore	109	8	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	119	1	120	
10	Orissa	118	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	118	2	120	
11	Punjab**	86	8	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	96	-	96	
12	Rajasthan	107	10	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	120	-	120	
13	Uttar Pradesh	276	25	5	1	1	2	-	-	-	-	-	-	-	-	-	-	-	310	2	312	
14	West Bengal	114	12	0	3	1	-	-	2	-	-	-	-	-	-	-	-	-	109	-	108	
15	total	2163	227	60	45	32	15	17	14	4	6	3	1	1	1	1	1	2599	18	2610		

* includes Mewatpur and Tripura ** includes Delhi and Himachal Pradesh, vill : villages

was confined to the hamlet-group selected for socio-economic survey in that village.

15.2m SELECTION OF HOUSEHOLDS

A list of all the households residing in the village/hamlet-group was prepared for sampling households starting from that point where the 1951 census enumeration began. If it was not known where the census numbering started, the listing of households was taken up from approximately the north-west corner of the village. If hamlet-group selection was done in a sample village, the hamlets in the selected group were taken up in the alphabetical order of their names for the purpose of listing households. At the time of listing the households in the sample village, information was collected on

- (i) household size,
- (ii) self-employment in manufacturing enterprise, and
- (iii) principal means of livelihood.

This information was utilized for selection of households for the different enquiries.

To avoid any possible bias on the part of the investigator in selecting the random number, the column of the random number table to be referred to was linked up with the last two digits of the serial number of the sample village given in the sample list. The order in which the random numbers should be selected for different enquiries was also prespecified. Each investigator was provided with a random number table having 104 columns with 50 four-digited random numbers in each column.

Income and Expenditure (Schedule 1.1)

In the first sub-round one household was selected with equal probability from each of the sample villages 1, 3 and 5 for canvassing

this schedule. The sample villages 1, 3 and 5 were considered to ensure geographical spread of the samples over the stratum as the sample of 6 villages assigned to an investigator was selected systematically with a random start after arranging the tehsils in a serpentine order. A sample of 15 households to be surveyed for this schedule in the second and subsequent sub rounds was selected circularly systematically from all the households in the 6 selected villages (or hamlet groups) taken together. This could not be done in the first sub round since the household lists for all the villages were not available at the beginning of the sub-round. This procedure of selection ensures proportional allocation of sample households to the selected villages.

Before selection the households in each of the sample villages 1, 3 and 5 were arranged in such a way that the households with less than 5 persons came first and those with 5 and more persons came next. In villages 2, 4 and 6 this arrangement by household size classes was reversed. This was done to ensure proper representation from these two classes of households. It would have been desirable to have the above pattern of arrangement for all the households in the six villages taken together, but this was not done so as to avoid the renumbering of households, since the households were selected separately from each village in the first sub round. The orders of selection of sample households surveyed in the second to sixth sub rounds were as follows:

sub round	orders of selection of sample households
2	1, 6, 11
3	2, 7, 12
4	3, 8, 13
5	4, 9, 14
6	5, 10, 15

Small-scale Manufacture and Handicrafts (Schedule 2.2)

For this enquiry the frame consisted of all the households self-employed in a non-registered manufacturing enterprise. In the first sub-round a sample of 2 households was selected circular systematically after arranging the households so that those having manufacturing enterprise as their principal means of livelihood came first and the other households came next in villages 1, 3 and 5. In villages 2, 4 and 6 this arrangement was reversed. A circular systematic sample of 50 households to be surveyed in the second and subsequent sub-rounds was selected with a random start from all the self-employed households in the 6 villages (or hamlet-groups) assigned to an investigator with the arrangement of the households effected in the first sub-round as indicated above. The sample households with orders of selection 1, 6, 11, 16, ..., 46 were surveyed in the second sub-round, those with orders of selection 2, 7, 12, 17, ..., 47 in the third sub-round and so on.

Employment and Unemployment (Schedule 10)

The number of households to be selected from each sample village was specified in the sample list. This number was arrived at by allocating 24 households to the 6 sample villages in proportion to their 1951 census population. In the case of villages which were *uninhabited* in 1951, one household was allotted in anticipation of their having become inhabited since 1951. From each sample village a circular systematic sample of the required number of households was selected with a random start after arranging the households according to the household size classes 1-4 and 5 & above. The same set of sample households were surveyed in each of the six sub-rounds.

15.2n SELECTION OF PLOTS

The unit of observation for crop survey was a plot which was defined as a distinct piece of land having a *survey number*¹⁸. If a plot had no survey number given to it, then it was associated

¹⁸ A *survey number* is the serial number given to a plot in the cadastral map of a village after land survey.

with the adjoining plot having the lowest survey number among the plots adjoining the unnumbered plot. The plots in the sample village were considered to be grouped into mutually exclusive clusters of 10 consecutive survey numbers or sampling serial numbers such as 1-10, 11-20, 21-30, 31-40 and so on.

From each sample village (or selected division) 6 clusters of 10 consecutive plots each were selected systematically with a random start from 1 to I , the sampling interval, which was taken as the integer next to the quotient obtained by dividing the highest survey number by 6. This selection procedure was adopted in sample villages where cadastral maps were available (method 1). In the case of villages where only a list of plots was available, the above selection procedure was adopted after giving sampling serial numbers to the plots (method 2). In the absence of a cadastral map or a list of plots, a linear systematic sample of 6 households was selected with a random start (method 3). For each crop season, data on land utilization were collected for all the 60 sample plots in the case of methods 1 and 2 and for all the plots possessed by a sample of 6 households in the case of method 3¹⁹.

In order to relieve the strain on the investigators in some specified hilly tracts, desert areas and inaccessible regions, the work load was reduced by allowing selection of 6 clusters of 5 plots each in villages where methods 1 and 2 were adopted and 3 households in villages where method 3 was adopted.

In the case of sample villages selected for crop cutting work, the crop plots in the sample clusters were arranged according to the crops grown in them in the following order, viz., paddy, jowar, bajra, ragi, maize, wheat and barley, the plots with mixed crops occurring two or more times in the appropriate places. A sample of 6 crop plots was selected with probability proportional to allocated crop area systematically with a random start. Crop cutting experiments

¹⁹ In the later rounds 6 clusters of 10 plots are selected only in crop cutting sample villages whereas 4 clusters of 5 plots are selected in the other sample villages.

were carried out in concentric circular cuts of radii 2'3" and 4' located at random in each of the selected crop-plots. In the case of partial cuts, a full cut was also obtained and the data for both the partial cut and the full cut were reported. The crop-cutting survey was carried out in the same sample villages in each of the crop seasons.

15.2o ESTIMATION PROCEDURE

Notation

- s subscript for s -th stratum;
- i subscript for i -th village or selected part in i -th village;
- j subscript for j -th household/cluster;
- K number of strata;
- N total number of villages;
- n number of sample villages surveyed in the sub-sample (including uninhabited villages and excluding casualties not substituted) in a particular sub-round;
- n' number of villages reporting price for a commodity;
- D number of hamlet-groups for socio-economic survey/divisions for crop survey formed within the village ($D = 1$ in case no such division was made);
- H total number of households/highest survey number/highest sampling serial number of the plots;
- h number of sample households for the schedule/plots surveyed in the round/sub-round/season (excluding casualties not substituted);
- y value of the study variable (in the case of dichotomy, this value is 1 if the unit belongs to the class, otherwise 0);
- G total geographical area of stratum;
- g geographical area of sample village/cluster;
- p price of the commodity;
- r proportion of area under particular type of land utilization.

Socio economic Survey

Table 15.14 gives unbiased estimators and the corresponding multipliers for the total value of any characteristic based on any particular sub sample by nature of enquiry and by sub rounds

TABLE 15.14 UNBIASED ESTIMATORS OF POPULATION TOTALS AND MULTIPLIERS FOR DIFFERENT ENQUIRIES

sr no	schedule †	sub round	unbiased estimator ‡	multiplier ²⁰
(1)	(2)	(3)	(4)	(5)
I income and expenditure (1.1)	1	$\sum_{s=1}^K \frac{N_s}{n_s} \sum_{i=1}^{n_s} D_{si} \frac{H_{si}}{h_{si}} \sum_{j=1}^{h_{si}} y_{sij}$	$\frac{N_s}{n_s} D_{si} \frac{H_{si}}{h_{si}}$	
2 small scale manufacture and handicrafts (2.2)	2 to 6	$\sum_{s=1}^K \frac{N_s}{n_s} \frac{\sum_{i=1}^{n_s} H_{si}}{\sum_{i=1}^{n_s} h_{si}} \sum_{i=1}^{n_s} D_{si} \frac{h_{si}}{N_s} y_{sij}$	$\frac{N_s}{n_s} \sum_{i=1}^{n_s} D_{si}$	
3 village statistics (3.0)	3	$\sum_{s=1}^K \frac{N_s}{n_s} \sum_{i=1}^{n_s} y_{sij}$	$\frac{N_s}{n_s}$	
4 retail prices of selected commodities (3.01)	1 to 6	$\sum_{s=1}^K \sum_{i=1}^{n_s} p_{si} / \sum_{s=1}^K n_s$	$1 / \sum_{s=1}^K n_s$	
5 employment and unemployment	1 to 6	$\sum_{s=1}^K \frac{N_s}{n_s} \sum_{i=1}^{n_s} D_{si} \frac{H_{si}}{h_{si}} \sum_{j=1}^{h_{si}} y_{sij}$	$\frac{N_s}{n_s} D_{si} \frac{H_{si}}{h_{si}}$	
6 population births and deaths (12.1 12.2)	1 to 6	$\sum_{s=1}^K \frac{N_s}{n_s} \sum_{i=1}^{n_s} D_{si} \sum_{j=1}^{h_{si}} y_{sij}$	$\frac{N_s}{n_s} D_{si}$	

† The figures in brackets denote the schedule numbers

‡ except for schedule 3.01 where a biased estimator of the average price was used for operational convenience

²⁰ In the later rounds the rural sample design is made self weighting for all household enquiries with one common inflation factor for each State. Also in these rounds the population enquiry is confined to a sample of households in the village instead of complete enumeration.

Crop Survey

An estimator of the area under a given type of utilization for a particular season based on a sub-sample or on the sample as a whole is given by

$$\hat{A} = \sum_{s=1}^K \hat{A}_s, \quad \dots \quad (15.1)$$

where for a hilly stratum²¹

$$\hat{A}_s = \frac{N_s}{n_s} \sum_{i=1}^{n_s} f_{si} D_{si} \frac{H_{si}}{h_{si}} \sum_{j=1}^{h_{si}} g_{sij} r_{sij}$$

and for a plains stratum

$$\hat{A}_s = G_s \left\{ \sum_{i=1}^{n_s} \hat{A}_{si} \right\} \left/ \sum_{i=1}^{n_s} g_{si} \right\}, \quad \hat{A}_{si} = g_{si} \left\{ \sum_{j=1}^{h_{si}} g_{sij} r_{sij} \right\} \left/ \sum_{j=1}^{h_{si}} g_{sij} \right\},$$

where $f_{si} = 1$ if the surveyed village coincided with the selected census village,

= number of revenue villages contained wholly or partly in the selected census village, or

= inverse of the number of census villages contained wholly or partly in the surveyed revenue village.

An estimator of the yield rate for a particular crop in a season was obtained as follows from sample villages taken up for crop-cutting experiments separately for pure and mixed crops and within these separately for hilly strata and plains strata :

$$\hat{R}_y = \frac{\sum_s \hat{A}_s \bar{y}_s}{\sum_s \hat{A}_s}, \quad \dots \quad (15.2)$$

where \bar{y}_s = simple average of yield rates over the cuts taken for the crop in the s -th stratum,

\hat{A}_s = estimate of area under the crop obtained from the villages where land utilization survey was conducted,

Σ' denotes summation over strata reporting crop-cutting experiments for the crop.

²¹Since the agricultural year 1962-63, this estimator is used for the plains strata also.

An estimator of production of crop was also obtained separately for pure and mixed crops and for hilly and plains strata separately, as product of the yield rate obtained as shown above from the reporting strata and the estimate of the area under the crop based on all the sample villages in all the strata, that is,

$$\hat{P} = \hat{R}_y \hat{A} \quad (153)$$

The final estimate was the sum of the four production estimates thus obtained

The above estimates are for the green weight of the crop. The estimate for the dry weight was obtained by multiplying the final estimate for each State by a drage factor. This factor was the ratio of the total dry weight to the total green weight of the crop (pure and mixed) obtained from the circular cuts of $2\frac{3}{4}$ " radius for the whole State.

Variance Estimator

If \hat{Y}_i ($i = 1, 2$) is the i th sub sample estimate (unbiased) of the total value Y , then a combined estimate \hat{Y} is given by

$$\hat{Y} = \frac{1}{2}(\hat{Y}_1 + \hat{Y}_2) = \frac{1}{2} \sum_{s=1}^K (\hat{Y}_{s1} + \hat{Y}_{s2}), \quad (154)$$

where \hat{Y}_{si} ($i = 1, 2$) is the i th sub sample estimate for the total in the s th stratum. An unbiased estimator of the variance of \hat{Y} is given by

$$v(\hat{Y}) = \frac{1}{4} \sum_{s=1}^K (\hat{Y}_{s1} - \hat{Y}_{s2})^2 \quad (155)$$

Another estimate $v(\hat{Y}) = \frac{1}{4}(\hat{Y}_1 - \hat{Y}_2)^2$ can be given, but this is less efficient than the former one.

An estimator of the ratio between two totals $R = Y/X$ is given by

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\hat{Y}_1 + \hat{Y}_2}{\hat{X}_1 + \hat{X}_2} \quad (156)$$

An estimator of the variance of \hat{R} is given by

$$v(\hat{R}) = \frac{1}{4\hat{X}^2} \sum_{s=1}^K \{ (\hat{Y}_{s1} - \hat{Y}_{s2})^2 - 2\hat{R}(\hat{Y}_{s1} - \hat{Y}_{s2})(\hat{X}_{s1} - \hat{X}_{s2}) + \hat{R}^2(\hat{X}_{s1} - \hat{X}_{s2})^2 \}. \quad \dots \quad (15.7)$$

A less efficient estimator of $V(\hat{R})$ but easier to compute is given by

$$v(\hat{R}) = \frac{1}{4} \left(\frac{\hat{Y}_1}{\hat{X}_1} - \frac{\hat{Y}_2}{\hat{X}_2} \right)^2 \quad \dots \quad (15.8)$$

An estimator of the variance of \hat{P} , the production estimate, is given by

$$v(\hat{P}) = \frac{1}{4} \left\{ \frac{\hat{P}'_1}{\hat{A}'_1} \hat{A}_1 - \frac{\hat{P}'_2}{\hat{A}'_2} \hat{A}_2 \right\}^2, \quad \dots \quad (15.9)$$

where \hat{P}' and \hat{A}' denote production and crop acreage estimates based on the strata reporting crop-cutting for that crop.

15.3 URBAN SECTOR

In this section different aspects of the sample design of the urban sector are described

15.3a SUBJECT COVERAGE

To get a complete picture for the country as a whole, the survey was undertaken in the urban sector also. The subjects taken up for survey in this round in the urban sector were (i) consumer expenditure, (ii) small-scale household manufacturing enterprise and (iii) employment and unemployment. These subjects were included with a view to comparing the employment and unemployment situation with that obtained in the ninth round (May–November 1955) and the information on household manufacturing enterprise with that obtained in the tenth round (December 1955–May 1956) and to continuing the time series of data on consumer expenditure.

A list of schedules canvassed in this round is given in Table 15.15. Besides these schedules, there was a schedule (0.2) for listing and selection of households for these enquiries. There was also a schedule (4.0) for recording time spent on different survey operations by the investigators.

TABLE 15.15 SCHEDULES OF ENQUIRY (URBAN)

sr no	schedule number	description
(1)	(2)	(3)
1	1.1	income and expenditure
2	2.2	small-scale manufacture and handicrafts
3	10	employment and unemployment

15.3b SURVEY PERIOD

For reasons similar to those mentioned for the rural survey, the period for the urban survey was also taken as one complete year. This period was divided into two sub rounds of 6 months each as against 6 sub rounds of 2 months each in the rural sector, since it was felt that there would not be substantial seasonal variation in the urban sector in the characteristics to be studied.

15.3c FIXATION OF WORK-LOAD

For the urban sector, it was decided to re-survey the sample of 2108 blocks taken up for survey in the ninth round so as to make the comparison of employment and unemployment situation with that obtained in the earlier round more effective. Since the number of investigators in Jammu and Kashmir was increased, the sample size of 108 blocks surveyed in the ninth round was increased to 216 blocks in that State for the fourteenth round. Thus the total number of blocks selected for this round came to 2228 blocks including 12 sample blocks selected afresh for Chandigarh.

On the basis of the available investigation strength of 70 investigators for the urban survey, it was found that an investigator should cover about 36 blocks. As the survey had to be completed in a period of one year, the average number of net working days available for surveying one block came to about 8 days. After considering the sample size desired for the different enquiries and the time requirements in the urban areas, the work-load in each sample block was fixed as in Table 15.16.

TABLE 15.16. WORK-LOAD AND TIME REQUIREMENTS FOR A. BLOCK.

sr. no.	schedule number	description of work and time requirement	no. of days (net)
(1)	(2)	(3)	(4)
1.	0.2	listing of households (about 150 hhs; 50 hhs/day)	3.0
2.	1.1	income and expenditure (1 hh; 1 hh/day)	1.0
3.	2.2	household manufacturing enterprise (3 hhs; 2 hhs/day)	1.5
4.	10	employment and unemployment (4 hhs; 4 hhs/day)	1.0
5.	—	journey (1 block; 1½ days/block)	1.5
6.		total number of days per block	8.0

(hhs : households)

15.3d INTERPENETRATING SUB-SAMPLES

The sample of blocks selected for survey in this round had been drawn in the form of 4 independent interpenetrating sub-samples. Sub-samples 1 and 3 were surveyed by one party of investigators and sub-samples 2 and 4 by a different party of investigators. The survey in sub-samples 1 and 2 was conducted first during the first sub-round of six months and sub-samples 3 and 4 were surveyed in the second sub-round period. In the case of participating States, a similar procedure was followed for the 4 independent interpenetrating sub-samples allotted to them for being surveyed.

15.3e SAMPLE DESIGN

As recorded earlier, all the sample blocks of the ninth round were surveyed in this round, except for the sample blocks in Madras and Jammu and Kashmir. In these cases, as fresh sampling lines were available, samples were drawn from them to minimize the difficulties experienced in the field in identifying the 1951 census blocks, sampling design being the same as was adopted in the ninth round. The general sampling design was a stratified two stage one, where 1951 census blocks were first stage units and households were second stage units. The details of the sampling design adopted for sampling blocks in the ninth round is given here.

In the case of the State sample, the sample blocks of the ninth round were retained for those States and parts of States which had been covered by the State in that round, and fresh selection was started to in the case of other States or parts of States.

15.3f STRATIFICATION

Each city with a population of 300000 and above as well as each taluk town of the former²² part A and part B States except Shillong, capital of Assam, was taken as a separate stratum. This was done because the unemployment situation was likely to be more acute in such places. In the case of Greater Calcutta 8 strata were formed because of greater heterogeneity of the population. In Jammu and Kashmir, Jammu town was considered as a separate stratum besides Sagar. In the rest of the urban area towns within *natural regions*²³ formed separate strata. There were altogether 94 strata for the whole of the Indian Union.

²² Prior to the reorganization of States on 1st November 1956 the States of the Indian Union had been divided into three categories parts A, B and C on political and administrative considerations.

²³ In the 1951 census 15 natural regions were formed on the basis of geographical and climatic conditions and the intersection of these regions with the State boundaries resulted in 52 natural divisions.

15.3g ALLOCATION OF SAMPLE BLOCKS

The first stage unit within each stratum was the 1951 census enumeration block consisting of roughly 100 to 200 households. Out of the 2228 blocks sampled throughout India for this round, 216 were from Jammu and Kashmir and 12 from Chandigarh. The rest of the blocks were allocated to the different strata in proportion to their respective non-agricultural population in 1951. However, as the acuteness of the unemployment problem in densely populated areas demanded special consideration, preferential weights were given to the cities and towns treated as separate strata. This was achieved by first allocating 1600 blocks to all the strata on the basis of their respective non-agricultural population and then the remaining 400 blocks to the cities and towns treated as separate strata on the same basis. In all cases the stratum allocations were rounded off to the nearest multiples of 4 in view of the requirement of 4 independent sub-samples. Since the sample blocks of the ninth round were retained for this round except in a few cases and since these blocks had been selected prior to the reorganization of States, the number of sample blocks in some of the reorganized States is not a multiple of 4. The allocations to the different States are shown in Table 15.1.

15.3h SELECTION OF URBAN BLOCKS

Since the socio-economic characteristics are likely to be related to the means of livelihood pattern of the region, all the towns (excepting those treated as separate strata) in each natural division were arranged according to their means of livelihood pattern. Within each town the census blocks were arranged according to their geographical nearness. With this arrangement of towns and blocks, 4 circular systematic samples were drawn with independent random starts.²⁴

²⁴In the later rounds sampling frames of area units specially prepared for this purpose in bigger towns and cities are used together with the 1961 census frame for the smaller towns. The sample blocks are selected pps systematically, size being related to population.

In the 1951 population census, the means of livelihood were classified into the following 8 broad classes, and the distribution of population in these 8 classes was given for each town in the country

<i>class</i>	<i>means of livelihood</i>
I	cultivators of land wholly or mainly owned and their dependants
II	cultivators of land wholly or mainly unowned and their dependants
III	cultivating labourers and their dependants
IV	non cultivating owners of land, agricultural rent receivers and their dependants
V	production other than cultivation
VI	commerce
VII	transport
VIII	other services and miscellaneous sources

The following dichotomies were considered

- A—towns and cities having 25% or more of population dependent on means of livelihood classes I, II, III and IV, A'—others,
- B—towns and cities having 25% or more of population dependent on means of livelihood class V, B—others,
- C—towns and cities having 25% or more of population dependent on means of livelihood class VI, C—others,
- D—towns and cities having 25% or more of population dependent on means of livelihood class VIII, D'—others

Classification (Order of Arrangement)

sr no	classification					sr no	classification				
1	A	B	C	D		9	A'	B	C	D	
2	A	B	C	D		10	A	B	C	D	
3	A	B	C	D		11	A	B	C	D	
4	A	B	C	D		12	A	B	C	D	
5	A	B	C	D		13	A	B	C	D	
6	A	B	C	D		14	A	B	C	D	
7	A	B	C	D		15	A	B	C'	D	
8	A'	B	C	D		16	A	B	C	D	

There were 16 joint means of livelihood classes within which the towns and cities could be classified, since four dichotomies were considered. All towns belonging to the joint class A'BC'D', namely, those having 25% or more of their population dependent on the means of livelihood class V and having less than 25% of the population dependent on each of the means of livelihood classes I to IV, VI and VIII, were put together first in the arrangement. The towns and cities belonging to the joint class ABC'D' came next in the arrangement, and so on, as shown above. This particular arrangement was adopted with a view to ensuring proper representation in the sample for the different classes with special emphasis on classes B and C. This can be seen from the arrangement where all the B's, B''s, C's and C''s come together if the arrangement is taken to be circular. Emphasis was given to classes B and C since they represent production other than cultivation and commerce respectively, both of which are important from the view-point of the unemployment situation in the urban sector.

The towns and cities within each of the above classes were further arranged according to geographical nearness as far as possible. The blocks were then arranged within each town or city according to their geographical nearness. The blocks were made as nearly equal in population content as possible by merging two or more adjacent blocks and sometimes by splitting the original blocks in terms of census house numbers. After the arrangement, 4 circular systematic samples of blocks were selected with independent random starts with a view to obtaining 4 independent sub-samples.

If a selected block was found to be very large, the investigator divided it into a number of sub-blocks each consisting of about 100-200 households. One of these sub-blocks was then selected with equal probability and the survey was confined to the households in it. This was done with a view to equalizing the work-load between different sample blocks and to avoiding the spending of much time in listing households for sample selection. In such cases, the inflation factor was suitably changed.

15.31 SELF-WEIGHTING DESIGN

To facilitate the work at the tabulation stage and to increase the efficiency of the sample, a self weighting design was adopted for the employment and unemployment enquiry²⁵. This was done by fixing the sampling intervals and random starts to be used in selecting households for the employment survey in the selected blocks (or sub blocks) in such a way that the inflation factor (multiplier) remains the same for all sample households. The constant inflation factor was so chosen as to get 4 sample households per block on the average for this survey. On the basis of the figures for the number of households in these sample blocks during the ninth round, it was found that the work load would vary much from block to block if one inflation factor was aimed at. To reduce the variation in work load in the different blocks, 6 different inflation factors were used. In the case of areas where fresh selection was resorted to, the 1951 census population of the sample blocks were used in determining the inflation factor and fixing the sampling interval in the sample blocks.

If B_s and b_s are the number of blocks in the population and in the sample respectively for the s th stratum and H_{si} and h_{si} the number of households in the population and in the sample respectively for the i th selected block of the s th stratum, the inflation factor is given by $B_s H_{si} / b_s h_{si}$. If the design were self weighting, the value of the constant inflation factor k would be the ratio of the total number of households in the urban sector to the number of sample households desired for the survey. The total number of households in the urban sector estimated on the basis of the ninth round survey (13.53 million) was inflated by 8% to allow for the increase since 1955. Hence the value of k is 1733.12 (= 14.614 million - 2108 × 4)²⁶. The sampling interval I_{si}

²⁵ In the later rounds the urban sample design is made self weighting for all household enquiries with one or two common inflation factors for each State.

²⁶ The number of blocks was later raised to 2298 due to an increase of 108 sample blocks in Jammu and Kashmir and allotment of 12 blocks to Chandigarh.

to be taken for the i -th selected block of s -th stratum was obtained as $k b_s / B_s$.

The ratio of H_{si} as obtained in the ninth round (or in the 1951 census for cases where fresh selection was adopted) to I_{si} was used to obtain an idea of the expected number of sample households in each sample block. The interval was changed using the following scheme with a view to equalizing the work-load between sample blocks. Only in Jammu and Kashmir, where the sizes of the blocks were very small, $I_{si}/4$ was used uniformly as the sampling interval for every block.

It may be noted that for the same sampling interval, the range of variation of H_{si} / I_{si} specified in Table 15.17 is less for the State sample than for the central sample. This was done because the figures for H_{si} used for the State samples were obtained from the 1951 census and the actual number of households was expected to have increased in the intervening period.

TABLE 15.17. SCHEME FOR CHANGING THE INTERVAL
TO EQUALIZE THE WORK-LOAD IN SAMPLE BLOCKS.

multiplier code (c)	central sample		State sample	
	$\frac{H_{si}}{I_{si}}$	sampling interval	$\frac{H_{si}}{I_{si}}$	sampling interval
(1)	(2)	(3)	(4)	(5)
1	< 2	$\frac{1}{2}I_{st}$	< 2	$\frac{1}{2}I_{st}$
2	2 — 5	I_{st}	2 — 4	I_{st}
3	5 — 10	$2I_{st}$	4 — 8	$2I_{st}$
4	10 — 20	$4I_{st}$	8 — 16	$4I_{st}$
5	20 — 40	$8I_{st}$	16 — 32	$8I_{st}$

TABLE 15.18 DISTRIBUTION OF SAMPLE BLOCKS BY INFLATION
FACTORS IN DIFFERENT STATES

sr no	State	inflation factor (k)						
		$\frac{1}{k}$	$\frac{1}{k}$	k	$2k$	$4k$	$8k$	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Andhra Pradesh	—	47	53	50	4	—	154
2	Assam*	—	3	17	3	1	—	24
3	Bihar	—	10	42	15	1	—	68
4	Bombay	—	62	259	103	6	1	431
5	Jammu and Kashmir	216	—	—	—	—	—	216
6	Kerala	—	16	44	3	—	—	63
7	Madhya Pradesh	—	9	54	19	6	—	88
8	Madras	—	52	116	60	5	—	233
9	Mysore	—	19	70	33	3	—	125
10	Orissa	—	1	10	4	1	—	16
11	Punjab**	—	56	96	13	1	2	168
12	Rajasthan	—	4	70	8	—	—	82
13	Uttar Pradesh	—	45	147	65	7	—	264
14	West Bengal	—	59	197	26	1	1	284
15	total	216	383	1175	402	36	4	2216

*includes Manipur and Tripura, **includes Himachal Pradesh and Delhi and excludes Chandigarh

15.3j SELECTION OF HOUSEHOLDS

After properly identifying the sample block with the help of the block boundaries given in the sample list (and after formation and selection of a sub block, if necessary), a list of all the households in that block (or sub block) was prepared starting from the point where the census enumeration began. At the time of listing the households, information on household size, self-employment in manufacturing enterprise and principal means of livelihood was collected to be used for sampling households for different enquiries. As explained in Sub section 15.2m, the column of the random number table to be referred to sample selection in a sample urban block was linked up with the last two digits of the serial number of the sample block given in the sample list.

Income and Expenditure (Schedule 1.1)

One household was selected with equal probability from all the households in each sample block (or sub-block).

Household Manufacturing Enterprise (Schedule 2.2)

From each sample block (or selected sub-block) a circular systematic sample of 3 households was selected with a random start for his schedule from all households self-employed in unregistered manufacturing enterprises, after arranging them such that those with manufacturing enterprise as their principal means of livelihood came first and the other households came next.

Employment and Unemployment (Schedule 10)

All the households in the selected block (or sub-block) were arranged such that those with at least one unemployed person came first and the other households came next. Using this arrangement, a linear systematic sample of households was selected using the random start and interval specified in the sample list.

15.3k ESTIMATION PROCEDURE

Notation :

s subscript for the *s*-th stratum

K number of strata

i subscript for the *i*-th block in the *s*-th stratum

j subscript for *j*-th household in the *i*-th block of the *s*-th stratum

c subscript for the *c*-th multiplier code for schedule 10

b number of blocks surveyed in the sub-sample

D number of sub-blocks formed in a sample block

H number of households in the block (or sub-block)

h number of sample households surveyed

M multiplier corresponding to the code

y value of the characteristic.

An unbiased estimator of total value of any characteristic in Schedules 1.1 and 2.2 based on a particular sub-sample is given by

$$\hat{Y} = \sum_{s=1}^K \frac{B_s}{b_s} \sum_{i=1}^{b_s} D_{si} \frac{H_{si}}{h_{si}} \sum_{j=1}^{h_{si}} y_{sij} \quad \dots \quad (15.10)$$

and an unbiased estimator of total value of any characteristic in schedule 10 based on a particular sub sample is given by

$$\hat{Y} = \sum_c M_c \sum_{i,j} y_{sij}, \quad (15.11)$$

where the second summation is taken over all sample blocks with the multiplier code c

39

If \hat{Y}_i is the estimate obtained from the i th sub-sample ($i = 1, 2, 3, 4$), the combined estimate is given by $\hat{Y} = \frac{1}{4} \sum_{i=1}^4 \hat{Y}_i$. An unbiased estimator of the variance of \hat{Y} is given by

$$v(\hat{Y}) = \frac{1}{12} \sum_{s=1}^K \sum_{i=1}^4 (\hat{Y}_{si} - \hat{f}_s)^2, \quad (15.12)$$

where \hat{Y}_{si} is the estimate of s th stratum total from the i th sub sample and \hat{f}_s is the combined estimate of the s th stratum total. It may be noted that another estimator of the variance of \hat{Y} is given by

$$v(\hat{Y}) = \frac{1}{12} \sum_{i=1}^4 (\hat{Y}_i - \hat{Y})^2 \quad (15.13)$$

But this is less efficient than the former estimate

An estimator of the ratio $R (= Y/X)$ of two totals Y and Z is $\hat{R} = \hat{Y}/\hat{X}$. An estimator of the variance of \hat{R} is given by

$$v(\hat{R}) = \frac{1}{12 \hat{X}^2} \sum_{s=1}^K \left\{ \sum_{i=1}^4 (\hat{Y}_{si} - \hat{Y}_s)^2 - 2\hat{R} \sum_{i=1}^4 (\hat{Y}_{si} - \hat{Y}_s)(\hat{X}_{si} - \hat{X}_s) + \hat{R}^2 \sum_{i=1}^4 (\hat{X}_{si} - \hat{X}_s)^2 \right\} \quad (15.14)$$

An alternative estimator which is less efficient is given by

$$v(\hat{R}) = \frac{1}{12} \sum_{i=1}^4 (\hat{R}_i - \hat{R})^2, \quad (15.15)$$

where \hat{R}_i is the estimate of the population ratio from the i th sub-sample and $\hat{R} = \frac{1}{4} \sum_{i=1}^4 \hat{R}_i$

Family Living Surveys

16.1 INTRODUCTION

The consumer price *index number* for a given class of people is an index built to indicate the change in the price structure of items consumed by them over a certain period of time. Usually it is computed by finding the weighted average of the ratios of the prices (*price-relatives*) of different items at the two points of time between which the change is to be measured, the weights being the respective proportions of expenditure of an average family on the corresponding items of consumption as observed in the earlier period (*base period*). The index is given by $I_{to} = \sum W_{io} r_{ito}$, where $i = 1, 2, \dots, N$ is over the different items of consumption, r_{ito} is the ratio of the price of that item at time point t to the price at time o , W_{io} is the proportion of expenditure of an average family on item i at time o .

In this chapter the sample designs adopted for working class and middle class family living surveys carried out by the Indian National Sample Survey (NSS) in several centres in India during the period 1958-59 are described*.

16.1a HISTORICAL NOTE

The earlier series of consumer price index numbers in India were compiled by various agencies in different areas. They originated from specific needs over a period of time in those areas and there is

* This chapter is based on a paper by Mrs. Nanamma Chinnappa, published in *Sankhyā*, (25, B, 1963, 359-418).

considerable lack of uniformity in the concepts and procedure of data collection and methods of constructing the index numbers. Many of them had a pre-war base and consequently did not reflect the current living conditions of the population classes they were meant for.

Their methods of construction and maintenance, particularly those of the older series, left much to be desired in view of the subsequent recent developments in the techniques of statistical surveys.

Consumer price index numbers for the working class were being maintained for more than 50 centres all over India by the State Governments and the Labour Bureau. In some cases, two different series were being maintained for the same centre by two agencies. The weights for these series were based on family budget enquiries conducted at those centres during periods varying between 1927 and 1952.

Very little had been done as regards similar enquiries for the middle class. In 1945-46, the Economic Adviser to the Government of India undertook an enquiry into the budgets of families of Central Government employees. Later, similar surveys were undertaken by the Reserve Bank of India in Bombay, the Gujarat Chamber of Commerce in Ahmedabad and the West Bengal State Statistical Bureau in some cities and towns of that State.

16.1b ORIGIN OF THE SURVEYS

The workers' and employers' organizations, the Central Pay Commission and other such bodies who use these index numbers, felt that fresh family budget enquiries should be conducted all over India on a uniform basis. In response to the growing demand for a new set of index numbers, the Second Joint Conference of Central and State Statisticians held in October 1953 recommended that fresh budget enquiries should be carried out on a uniform basis in all the States during 1955-56. A project for conducting fresh family budget enquiries among working classes at important industrial centres was included in the Second Five Year Plan. Some of the State Governments were also very keen on conducting these enquiries so

as to enable them to revise the earlier index numbers. For various reasons, however, the project could not be taken up in 1955-56. After consultations with various authorities, the Government of India undertook the scheme for conducting the Working Class and Middle Class Family Living Surveys in 1958-59.

16.1c ORGANIZATIONAL SET-UP

The working class family living survey was sponsored by the Labour Bureau and the middle class family living survey by the Central Statistical Organization (CSO). These agencies were mainly responsible for the overall implementation of respective enquiries, general supervision of their progress, preparation and printing of reports and construction and maintenance of the corresponding consumer price index numbers. The task of compiling the middle class consumer price index numbers has since been transferred from the CSO to the Labour Bureau.

In conformity with the requirements of the users, the Indian Statistical Institute (ISI) prepared the sampling design and the schedules to be canvassed after conducting the necessary preliminary enquiries.

The agencies for the field work for the family living surveys were the NSS Directorate and the ISI field branch (the latter in West Bengal and Bombay city only). The ISI was responsible for processing and tabulation of data.

As regards price quotations for the working class survey, the Labour Bureau used its price agency for collection of prices, whereas for the middle class survey, the NSS field staff collected the prices.

16.1d SCOPE OF THE SURVEYS

The object of these surveys was two-fold :

(i) To estimate the weighting diagrams for the working class and middle class populations in different important centres all over India. These were to be based on up-to-date consumption patterns for being used in compiling consumer price index numbers for these classes

in each centre (It was also envisaged that the centre wise consumer price index numbers so compiled would be utilized in the compilation of a representative composite series of all India consumer price index numbers for the working class and middle class separately)

(ii) To collect data for a general study of the living conditions of these classes in the different centres

For the first purpose, information on all types of income and expenditure of the sample family was collected in Schedule A. For the second purpose Schedule B was used. This schedule covered various aspects of the family such as health, employment, housing conditions, satisfaction of social, cultural and educational needs, etc.

Barring a few exceptions each of the surveys was conducted for a period of one full year in the allotted centres. The period for the middle class surveys was July 1958-June 1959, and for the working class surveys it was August 1958-July 1959 in each centre. The sample size at any centre was spread evenly over twelve months to eliminate seasonal variations.

16.2 WORKING CLASS SURVEY

As stated in Section 16 Id, it was decided to collect data on consumer expenditure and general level of living of the working class in certain specified centres distributed all over India. These centres were towns or cities which were important for their concentration of working class population. The survey was planned separately in each centre, taking into consideration the local conditions obtained there and a suitably selected sample of working class families was contacted for collection of data.

For the purposes of this survey, the working class was defined broadly to include those who did manual work in registered factories, mines or plantations (registered under the Factory, Mine and Plantation Acts of 1948, 1952 and 1951 respectively), and information was

collected only from working class families in each centre. A working class family was defined as one which derived 50% or more of its income through manual work in a registered factory, mine or plantation in the calendar month preceding the date of contacting the family for the survey. It may be noted that only families of factory workers were surveyed in a centre which was specified as a factory centre, only families of mine workers in a mining centre and only families of plantation workers in a plantation centre.

16.2a SELECTION OF CENTRES

Initially it was decided to conduct the survey in about forty important factory, mining and plantation centres representing the industrial areas of India. The numbers of centres that were to represent these three industry sectors were first determined in proportion to the all-India employment in them. The number of centres for each sector was distributed to the States in proportion to their employments in that sector. A State which had substantial employment in a particular sector but did not receive any allocation by the above procedure (because its employment in that sector was comparatively low judged at the all-India level), was allocated one centre to represent that sector. The State Governments were then requested to name the allocated number of important centres in each sector in their respective States. Later, after mutual consultations between the Labour Bureau and the State Governments some changes were made in the list so as to ensure better representation for different types of industries, mines or plantations and better geographical coverage within each State.

There were 50 centres in all, of which 32 were factory centres, 8 were mining centres and 10 were plantation centres. The complete list of centres is given in Table 16.1.

TABLE 16.1. LIST OF CENTRES AND SAMPLE SIZES FOR
THE WORKING CLASS SURVEY (1958-59)

sr. no.	State	centre	sampling method	no of investi- gators	no. of families for schedule A	no. of families for schedule B
(1)	(2)	(3)	(4)	(5)	(6)	(7)
(a) factory centres						
1.	Andhra Pradesh	Guntur	T	1	180	60
2.	"	Hyderabad	P	2	480	120
3.	Assam	Digboi	T	1	240	60
4.	Bihar	Jamshedpur	P	4	720	240
5.	"	Monghyr-Jamalpur	P	2	480	120
6.	Gujerat	Ahmedabad	T	4	720	240
7.	"	Bhavanagar	P	1	240	60
8.	Maharashtra	Bombay	T	8	1440	480
9.	"	Nagpur	P	3	720	180
10.	"	Sholapur	T	3	540	180
11.	Jammu & Kashmir	Srinagar	P	2	480	120
12.	Kerala	Alleppey	T	2	360	120
13.	"	Alwaye	P	1	240	60
14.	Madhya Pradesh	Bhopal	P	1	240	60
15.	"	Gwalior	T	2	360	120
16.	"	Indore	P	2	480	120
17.	Madras	Coimbatore	P	3	720	180
18.	"	Madras	P	4	960	240
19.	"	Madurai	T	2	360	120
20.	Mysore	Bangalore	T	4	720	240
21.	Orissa	Sambalpur	P	1	240	60
22.	Punjab	Amritsar	T	3	540	180
23.	"	Jamunanager	P	1	240	60
24.	Rajasthan	Ajmer	T	2	360	120
25.	"	Jaipur	P	1	240	60
26.	Uttar Pradesh	Banaras	P	1	240	60
27.	"	Kanpur	T	4	720	240
28.	"	Saharanpur	T	2	360	120
29.	West Bengal	Asansol	P	3	720	180
30.	"	Calcutta	T	4	720	240
31.	"	Howrah	P	5	1200	300
32.	Delhi	Delhi	T	4	720	240
sub-total I to 32				83	16980	4980

TABLE 16.1. (contd.) LIST OF CENTRES AND SAMPLE SIZES
FOR THE WORKING CLASS SURVEY (1958-59).

sr. no.	State	centre	sampling method	no. of investi- gators	no. of families for schedule A	no. of families for schedule B
(1)	(2)	(3)	(4)	(5)	(6)	(7)
(b) mining centres						
33.	Andhra Pradesh	Gudur	P	1	240	60
34.	Bihar	Jharia	P	4	960	240
35.	"	Kodarma	P	1	240	60
36.	"	Noamundi	P	1	240	60
37.	Madhya Pradesh	Balaghat	P	1	240	60
38.	Mysore	Kolar G. F.	T	1	180	60
39.	Orissa	Barbil	P	1	240	60
40.	West Bengal	Raniganj	P	2	480	120
sub-total 33 to 40				12	2820	720
(c) plantation centres						
41.	Assam	Doom Dooma	P	2	480	120
42.	"	Labac	P	1	240	60
43.	"	Mariani	P	1	240	60
44.	"	Rangapara	P	2	480	120
45.	Kerala	Mundakkayam	P	1	240	60
46.	Madras	Coonoor	P	1	240	60
47.	Mysore	Ammathi	P	1	240	60
48.	"	Chikmagalur	P	2	480	120
49.	West Bengal	Darjeeling	P	2	480	120
50.	"	Jalpaiguri	P	2	480	120
sub-total 41 to 50				15	3600	900
grand total				110	23400	6600

(T = tenement sampling method; P = payroll sampling method).

16.2b FIXATION OF SAMPLE SIZE

Evidently the precision of the consumer price index number depends very much on the price relatives used in building it up. In fact, it depends more on the magnitude and variances of the price

relatives and their covariances with the weights than on the variances of the weights used. The problem can be posed thus—if the price relatives remained constant, what sample size would be required to provide estimates of consumer price index numbers with a margin of error not exceeding 3 or 4%? It was later decided that the maximum error permissible should be 2%. The error of the index number would under these restrictions depend on the variation in the relative magnitudes of the price relatives used and on the sampling variation in the estimated weights. A study was made using the weights obtained from the Jagaddal Labour Enquiry conducted by the ISI in 1941 and two sets of price relatives, one for 1942 and one for 1945 with 1941 as base. Studies were also made on the basis of the NSS data for Bombay, Ahmedabad and Madras to have an idea of the error of the food index for different sets of price relatives. These studies led to the conclusion that a sample of 300 to 1000 families per centre (depending on the type and size of centre) would possibly satisfy the requirements of precision aimed at for the consumer price index number, and for obtaining sufficiently precise estimates of the weights for the major groups of items. In centres with considerable variation in the consumption patterns of the working class, the sample size had to be larger, and in small centres, particularly in mining and plantation centres with less variation in the consumption patterns of the working class, a smaller sample size was considered sufficient.

Another important consideration in fixing the final sample size in any centre was the work load manageable by an investigator. From the NSS experience it was decided that a two stage sampling design, in which blocks of houses (in a *tenement† sampling centre*) or groups of establishments (in a *payroll† sampling centre*) were selected in the first stage and working class families or workers in the establishments were

[†] In a *tenement sampling centre* a sample of working class families is drawn from a frame of residential blocks and families, whereas in *payroll sampling* a sample of workers is drawn from the payrolls of the factories, mines or plantations in the centre and the families to which they belong are contacted for obtaining the required information, (cf Sub section 16 2c for further details)

selected in the second stage, would be suitable for this survey. It was necessary that the number of blocks or establishments thus covered was fairly large to get representation of different types of the working class population living in different areas. After field testing of the time taken for listing blocks (of the size of 1951 census enumerator blocks), for collecting payrolls and for canvassing the proposed schedules, it was found that an investigator could list about 3 blocks and survey about 20 families per month in a tenement centre and collect payrolls of 3 or 4 establishments and survey about 25 families in a payroll centre (larger in the latter case, because of the lesser time taken for collecting payrolls). Thus the minimum sample size was fixed at 240 families for a tenement sampling centre and 300 for a payroll sampling centre. Of this one-fourth and one-fifth of the sample size was set apart for the study on conditions of living (Schedule B) in tenement sampling and payroll sampling centres respectively. The suggestion to canvass both the schedules for the same sample households was dropped because of the fatigue that such a scheme would cause both to the informant and the investigator. The effective sample size for Schedule A, that is, for estimating weights for the index number therefore had to be 180 or its multiple in a tenement sampling centre and 240 or its multiple in a payroll sampling centre.

With the budget limitation of about 100 investigators for this survey, the sample size for each centre was fixed as a multiple of either of the sizes arrived at above depending on the method of sampling adopted. This was done after considering the size (measured by the labour force employed in that centre), importance of the centre and expected variation in the consumption pattern of working class population living in the centre.

One important feature of the survey was the provision for having at least two independent sub-samples in each centre. Since it was desirable to maintain this independence at the stage of data collection itself by allotting one sub-sample to each investigator, this meant a minimum sample size of 360 for a tenement sampling centre and 480 for a payroll centre for estimating weights,

With the budget limitations imposed, however, it was found that this could not be achieved without substantial decrease in the sample size for the larger centres. Further, a sample size of 360 or 480 was not required in the very small centres. In 20 of the smallest centres, therefore, the sample size was kept at 180, if it was a tenement centre and 240 if it was a payroll centre. In each of them two independent sub-samples of 90 or 120 families each were allotted to two six-monthly periods, each period consisting of six alternate months. In these centres, independence at the investigation stage was achieved by pairing two nearby centres and interchanging the investigators between them in every successive month. The final allocations decided for the centres are given in Table 16.1.

16.2c PRELIMINARY SURVEY

The different conditions prevailing in the various centres required a study of these local conditions for efficient centre-wise planning of the details of the survey. To this end, a preliminary survey was taken up in each centre. The main purposes of this survey were (i) to fix the boundaries of the centre, (ii) to decide on the most suitable sampling method for the centre, (iii) to prepare a suitable sampling frame and (iv) to test the schedules drafted for the survey and study the operational difficulties that might arise during the main survey. This survey was conducted in the period December 1958–February 1959 and was of one to two months' duration in any one centre. The field officer in charge of this preliminary survey consulted the labour, trade union, municipal and other knowledgeable authorities of each centre and collected the required information in the questionnaires and schedules drafted for this purpose.

The working class population in any centre could be contacted by two methods. The *tenement sampling method* consists in approaching them through their places of residence. This is the method of sampling adopted in the NSS socio-economic surveys. For practical convenience, a number of blocks are sampled in each town, and within these blocks, the families living in them are listed and sampled.

The other method, called the *payroll sampling method*, is to approach them through their places of work. A number of establishments employing the working class population are sampled and workers are selected using the payrolls of these establishments and their families are contacted. The suitability of either of these methods for a centre depended on the local conditions. If the working class population is concentrated in some areas of the centre, tenement sampling would be operationally effective, whereas if they were dispersed over a large area, the payroll method would be more effective.

For the tenement sampling method, the NSS had already the list of blocks (based on the 1951 census blocks) for every town/city, which was used for the NSS surveys. In some centres these lists were *outdated and the blocks were difficult to identify or had to be subdivided because they were too large*. For the payroll sampling method, the latest list of establishments for the centre had to be collected from the Inspector of Factories, Mines or Plantations, or other concerned authorities in the centre. If this method was to be adopted, it had to be ensured that the payrolls of the establishments were accessible and complete, and that it would be feasible to use them as a frame for sampling workers and later to contact the families of the sampled workers for detailed investigation.

During the preliminary survey at each centre, the officers of the NSS acquainted themselves with the local conditions in the centre by visiting the working class areas and discussing with the local authorities. The decision as to which sampling method was to be adopted in any centre was based on the reports of these preliminary enquiries. In some large centres, tenement sampling method was abandoned in favour of payroll sampling, because the preliminary enquiry for picking out areas of labour concentration, etc. would have been time-consuming. The payroll method was adopted in 17 factory centres and in all the plantation and mining centres except in Kolar Gold Fields, where tenement sampling was found to be more practicable. In the remaining centres tenement sampling method was used.

Since each of the selected centres was to represent the working class population of the area in which it was situated, it was evident that the municipal or corporation limits of the centre would not necessarily define the boundaries of the centre for the purposes of this survey. Very often there was a large proportion of workers coming from outside the administrative boundaries of the centre to work in the industrial establishments within it, or a large number of such establishments were situated outside the administrative limits of the centre and employed labour from that centre. The first task of the NSS officers in charge of the preliminary survey was to consult the local authorities of the centre and arrive at the boundaries of the centre.

The boundaries were fixed so that if it was a tenement sampling centre, a sizable majority of the working class population of that area resided within these limits, and alternatively if it was a payroll sampling centre, the industrial establishments which employed a sizable majority of the working class population of that area were situated within the prescribed boundaries. (One restriction imposed was that the geographical size of the centre should not be so big as to be unmanageable in a localized survey of this kind). In the latter case, this amounted to defining the centre by a list of factories, mines or plantations which commonly went by the name of the selected centre and covering the families of all the workers employed in these establishments. In some cases the Labour Bureau, with its knowledge of the area after consultation with the State Government concerned, defined the boundaries of the centre, and the local authorities were asked to suggest modifications to this. All these suggestions at various levels were pooled to decide upon the final boundaries.

16.2d TENEMENT SAMPLING METHOD

Frame Collection

In a centre where the tenement sampling method was adopted, the boundary fixation and collection of the basic frame were done during the preliminary survey. Having fixed the boundaries of the centre those areas of the centre which could be safely omitted

from the survey coverage (because they contained a meagre proportion of working class population) were demarcated with the help of the local authorities. In the rest of the area, the blocks were classified into two types—those which had working class population residing in them in large concentrations and were therefore important for this survey, and those which did not have such concentrations. This was done by marking out all such well-known important concentrations on a map or determining their boundaries and matching them with the NSS frame of blocks. Sometimes, important wards were picked out, and in these wards and other smaller pockets of working class concentration, the blocks constituting them were classified as important or not according to the proportion of working class population in them. Wherever sufficiently reliable information was available, information was also obtained on the industry-wise and State-wise distribution of the working class population in each such concentration. The State to which a person belongs, and sometimes the industry in which he is employed influences his consumption pattern; this information therefore proved useful in improving the sampling procedure. The actual size of each block, that is, its population, which was relevant for controlling the listing time, was obtained either by local enquiry or by using 1951 census figures. This list of blocks was the sampling frame.

In a few centres, where the 1951 census blocks could not be identified, blocks of the same size as in 1951 census were formed specially for this survey and were used as the sampling frame in the same manner as given above.

Sampling of Blocks

The number of investigators to be allotted to a centre was obtained by dividing the sample size for Schedule A for a centre by 180 (which was the prescribed work-load for an investigator for a year). Because of the work-load restrictions, it was decided to sample 36 blocks per investigator in any centre, and distribute them over the months of enquiry at the rate of 3 blocks per month. The samples for different

investigators were drawn independently. The first stage unit of sampling was a group of blocks.

In a centre, if reliable auxiliary information was available and it was observed that there was substantial variation in the type of working class concentration among the blocks, clusters of 3 blocks each were formed so as to include blocks having different concentrations and types of workers in a cluster. Attempts were also made to equalize the cluster sizes with respect to population. If sufficiently reliable auxiliary information was not obtained on the nature of working class composition in the blocks, they were grouped into clusters of 3 blocks each as follows while grouping attempts were made to include blocks from different areas of the centre in a cluster so as to ensure some measure of heterogeneity in the cluster and to make each cluster comprise about 450 households. In either case, a systematic sample of 12 such clusters was drawn independently for each sub sample and allotted to an investigator.

If the number of blocks in a centre was too large for forming such clusters before selection, the clusters were formed after selection in the following manner. The blocks were arranged using the criteria for clustering and a systematic sample of 36 blocks was drawn from this list and the selected blocks were then grouped to form 12 clusters of 3 blocks each, the grouping being done in a systematic manner using the same arrangement. In cases where reliable information was also recorded on the expected number of workers or the working class population in each block this was used as a measure of size for the blocks, and clusters were selected systematically with probability proportional to the total of this size for the blocks constituting them.

In some cases where a marked difference was noted in the degree of concentration of the working class in the blocks, they were separated into two strata one containing blocks with a high concentration of the working class and the other containing the remaining blocks. The quota of 3 blocks per month per investigator was distributed to these two strata in the ratio 2 : 1 or 1 : 2 depending on

the total strength of the working class population in them. Within each stratum, sampling of the blocks or clusters of blocks (if the allocation was 2 per month) was done utilizing the relevant information available as in those centres where no stratification was adopted.

Selection of Sub-blocks

In centres other than those where pps sampling was adopted as described earlier, if it was found that the expected population in a block was very large according to the information collected at the preliminary survey, sub-block selection was adopted. In such cases, the number of divisions into which the block was to be divided was specified at the stage of sampling such that each such division had roughly the size of an average block (150 households). Selection of these blocks was done systematically with pps, size being the number of divisions in them. That is, these blocks were given as many serial numbers as the specified number of sub-blocks before systematic selection. On the field, the investigator demarcated the block into the required number of sub-blocks which were of roughly equal size and selected one of them at random for the actual survey. In any case, whether sub-block selection was specified or not, if it was found that a sample block was unmanageable in size, the investigator was allowed to decide the number of sub-blocks himself, according to the above criteria and to resort to sub-block division as described above.

Order of Visit to Blocks

In the larger cities, after a few months of survey it was found that because of the size of the blocks and the difficulties in contacting the heads of the households, listing took so much time that keeping to the programme of work was becoming impossible. Sub-block division did not appreciably reduce the listing time because forming them in itself was time-consuming. For the remaining months, therefore, the following scheme, designed to cut down the listing time, was adopted. The 3 blocks to be surveyed in a month were given a random order of visit. The investigator listed the blocks in that order. If on completely listing the first block in this order, he found that he had listed 450 or

more families and had obtained 25 or more working class families among them, he could stop listing and select the families from that block rejecting the blocks not listed. If not, he completely listed the second block also. And if the prescribed quota of families (450 or more listed with 25 or more working class families among them) was obtained in these 2 blocks put together, he stopped listing and selected the families from these 2 blocks rejecting the third block. Of course, the investigator had to list the block with order of visit 3 also, if the first two blocks together did not yield the quota of 450 and 25 families for listing and survey respectively. This procedure helped in reducing the listing time and stabilizing the work programme.

Allotment of Clusters to Months

Each of the 12 clusters sampled for an investigator was assigned to a particular month for enquiry by a random process.

Substitution of Blocks

As mentioned earlier since the sampling frame used was usually the 1951 census list of blocks, many difficulties were faced in identifying the sampled blocks. In a few cases when a block could not be identified at all, or was demolished, or could not be listed for security reasons (if it fell within military area), this was substituted or sampling was done from the remaining blocks in the cluster.

In a few centres due to unavoidable practical difficulties, such as an investigator resigning and no substitute being appointed immediately to take his place or due to the excessive time taken in listing (this happened in the beginning of the survey), it was found that some clusters had to be treated as casualties. Wherever feasible, the investigator was allowed to sample the families for that month from the families listed for a cluster surveyed earlier so as to save the listing time and avoid a total casualty for that month.

Sampling of Families

The second stage unit for selection was a working class family. Each month the investigator listed all the families in the cluster allotted to that month by a house to house visit and classified them

as working class families and others. While listing, information was also collected on the family size, the expenditure class to which it belonged and the state of origin of the head of the family. The two expenditure classes were those with expenditure less than Rs. 60 per month and others. Families belonging to the State of origin of the majority of families in the cluster were put in one class and the rest formed the other class. This information was used to arrange the working class families in the cluster, first by family size (single member families, and others) and within these classes by expenditure class and within these by the State of origin class. A systematic sample of 20 working class families was drawn from this arranged list. Every fourth family in this sample was contacted for filling in Schedule B (level of living) and the remaining were for Schedule A (consumer expenditure). In centres where stratified sampling was adopted, the number of families to be sampled separately from each stratum in any month was determined by allocating 20 to the strata in proportion to the expected size of the working class population in them.

Substitution of Families

In case a particular family could not be contacted due to various reasons, or refused to give information, (even after trying all the means of persuasion available with the field officers), it was substituted by the next working class family in the arranged list of that cluster. It may be noted that in case the main earners of a family were employed in manual labour in registered establishments but were on strike or lay-off for a period of not more than 45 days preceding the date of enquiry, that family was treated as a working class family even if 50 per cent or more of its income in the last calendar month was not from manual labour in registered establishments (as is required by the general definition).

If in a cluster sampled for a particular month the required number of working class families (20) was not available, the deficiency was made up from the sampling frame for the next month's cluster for the same stratum (if any) and sub-sample so as to keep the total sample size unaltered.

16.2e PAYROLL SAMPLING METHOD

Frame Collection

In each centre where this method was adopted, the latest available complete list of registered establishments (factories, mines or plantations) that could serve as the first stage sampling frame was compiled during the preliminary survey. Information, wherever available, was also obtained on the number of persons employed, the industry and management type, the seasonal nature of work and any other relevant classificatory characters applicable to the establishment. To have an idea of the geographical distribution of the establishments in mining and plantation centres (where travelling facilities were often poor), a map showing their location was prepared, or the distance and sometimes the direction of each establishment from the headquarters of the centre was obtained.

Sampling of Establishments

The establishments were clustered into groups of 3, 4 or 5 depending on their size (number employed) so as to have clusters of roughly equal size. Attempts were made to group within a cluster the establishments which were as heterogeneous as possible with respect to the various relevant auxiliary characters on which information was obtained. In mining and plantation centres, it was ensured as far as possible that a cluster was composed of units which were not very far apart. 12 clusters per sub sample were sampled systematically with probability proportional to the total size of the clusters, after arranging them in increasing order of size, and these were allotted to an investigator. In cases where the sizes were not available or where it was reported that the sizes recorded were not reliable, the sample of clusters was drawn systematically with equal probability.

In some centres where the sizes of a few establishments were very large compared to the average size of the others, the large establishments were grouped to form a stratum and the remaining establishments constituted another stratum. Clustering and sampling were done separately within each stratum. The number of factories to

be sampled each month from the two strata was decided on the basis of the stratum sizes. Sometimes it happened that there was one very big establishment which employed a substantial majority of the workers in the centre. In that case this establishment was repeated in the sample every month and a cluster of factories was sampled from among the remaining. The sampling within strata was similar to that in centres where there was no stratification.

Allotment of Clusters to Months

The month to which any cluster was allotted for enquiry was decided by a random process as in tenement sampling centres.

Substitution of Establishments

If while listing, a sample establishment was found to have been closed down permanently or temporarily for a sufficiently long or indeterminate period, it was substituted by some other establishment in the same stratum (if any) which was similar to it in respect of industry, number employed, etc.

If a seasonal establishment was allotted for survey in a month outside the season in which it was expected to be working, it was exchanged with a sample establishment which was expected to be working then. Sometimes such exchanging was done after actually visiting the establishment for enquiry, because the seasonality was known only then. If an establishment was inaccessible in the month of enquiry because of a strike or lay-off, it was exchanged with a similar establishment allotted to a month when it was expected to be working.

Sampling of Workers

In payroll sampling centres, the ultimate sampling unit, a working class family was approached through the payrolls of the establishments. The payroll of each sample establishment was used, after ensuring that it was complete, without duplications and up to date. Within each establishment, any available arrangement by section, pay-scale, or type of work was retained and from the pooled

payrolls of the establishments in a cluster a systematic sample of 25 workers was drawn. Of this a simple random sample of 5 workers was selected for Schedule B (level of living) and the remaining 20 were taken for Schedule A (consumer expenditure). In centres where stratified sampling was adopted the number of workers to be sampled from a stratum was specified in proportion to the size (total number of workers) of the stratum. Selection was done as described above within each stratum.

The families to which the sample workers belonged were contacted for the actual enquiry. If the same family was to be contacted more than once in the same month for the same schedule because two or more of the workers in it were sampled in that month the schedule was duplicated as many times as the family was sampled for the purposes of final tabulation.

Substitution of Families

As in the case of tenement sampling substitution by the family of the worker listed next was resorted to when the sample family could not be contacted or refused to give information (even after all efforts were made by the field officers). If the required number of working class families (25) was not available from a sample cluster of a particular month the deficiency was made up from the sampling frame of the next month's cluster for the same stratum (if any) and sub sample so as to keep the total sample size unaltered.

16.3 MIDDLE CLASS SURVEY

In this survey data on consumer expenditure and level of living were collected from families sampled in some towns and cities distributed all over India. The middle class was defined as the class comprising those who were gainfully occupied as employees (in both the public and private sectors) doing non manual work in the non agricultural sector. This excluded the working class population (as defined for the working class survey) those employed in the agricultural sector, employers and self employed persons. A middle

class family was defined as one which derived 50% or more of its income during the calendar month preceding the date of contacting the family from the occupation of one or more of its members as employees doing non-manual work in the non-agricultural sector. No income limits were proposed at the stage of collecting the information; such limits were to be imposed, centre-wise, at the stage of tabulating the data so as to exclude the very rich and the very poor who do not actually belong to this class.

16.3a SELECTION OF CENTRES

At first it was suggested that the survey should cover the four big cities, Delhi, Bombay, Calcutta and Madras and the urban areas in the six census zones, to be treated, possibly, as separate strata so that separate consumer price index numbers could be constructed for each of them as well as for all-India. Later, after discussion with the users, it was decided that the survey should be conducted in about 45 centres to be chosen in consultation with the State Governments.

The centres consisted of the State capitals and other large cities and towns which were important business centres and where there was considerable middle class concentration, because the index number for each such individual centre would be required for the adjustment of pay-scales and allowances for Government employees and employees of banking, insurance and other commercial concerns.

Initially, the State capitals were selected and their number was 16. The allotment of the remaining centres to the States was in proportion to their urban middle class population (excluding the population of the capitals). Each State Government named the required number of important centres to fulfil the quota allotted. Minor changes were made later to allow for better regional representation within a State. The final list of 45 centres selected is shown in Table 16.2. It may be noted that 18 of these centres were selected for the working class family living survey also.

TABLE 16.2 LIST OF CENTRES AND SAMPLE SIZES FOR
THE MIDDLE CLASS SURVEY (1958-59)

sr no	State	centre	no of investigators	no of families for	
				schedule A	schedule B
(1)	(2)	(3)	(4)	(5)	(6)
1	Andhra Pradesh	Hyderabad Secunderabad	4	720	240
2	"	Walair Vishakapatnam	3	540	180
3	"	Vijayawada	3	540	180
4	"	Kurnool	2	360	120
5	Assam	Shillong	2	360	120
6	"	Gauhati	2	360	120
7	Bihar	Patna	3	540	180
8	"	Ranchi	3	540	180
9	"	Muzaffarpur	2	360	120
10	Gujerat	Ahmedabad	4	720	240
11	"	Rajkot	3	540	180
12	Maharashtra	Bombay	8	1440	480
13	"	Poona	4	720	240
14.	"	Nagpur	4	720	240
15.	Jammu & Kashmir	Srinagar	2	360	120
16	"	Jammu	2	360	120
17.	Kerala	Trivandrum	3	540	180
18	"	Kozhikode	3	540	180
19	Madhya Pradesh	Bhopal	3	540	180
20.	"	Indore	3	540	180
21	"	Gwalior	3	540	180
22	"	Jabalpur	3	540	180
23.	Madras	Madras	6	1080	360
24	"	Madurai	3	540	180
25	"	Tiruchirapalli	3	540	180
26	Mysore	Bangalore	4	720	240
27	"	Mangalore	3	540	180
28	"	Hubli Dharwar	3	540	180
29	"	Gulbarga	2	360	120
30	Orissa	Cuttack Bhubaneswar	3	540	180
31.	"	Sambalpur	2	360	120
32	Punjab	Chandigarh	3	540	180
33	"	Amritsar	3	540	180
34	Himachal Pradesh	Simla	2	360	120

TABLE 16.2. (*contd.*) LIST OF CENTRES AND SAMPLE SIZES
FOR THE MIDDLE CLASS SURVEY (1958-59).

sr. no.	State	centre	number of investi- gators	number of families for	
				schedule A	schedule B
(1)	(2)	(3)	(4)	(5)	(6)
35.	Rajasthan	Jaipur	3	540	180
36.	"	Jodhpur	3	540	180
37.	"	Ajmer	3	540	180
38.	Uttar Pradesh	Lucknow	4	720	240
39.	"	Kanpur	4	720	240
40.	"	Allahabad	3	540	180
41.	"	Agra	3	540	180
42.	"	Meerut	3	540	180
43.	West Bengal	Calcutta	8	1440	480
44.	"	Kharagpur	3	540	180
45.	Delhi	Delhi-New Delhi	6	1080	360
total			149	26820	8940

Note : Tenement sampling method was used in all the centres except in stratum 1 of Kharagpur where payroll sampling was adopted.

16.3b FIXATION OF SAMPLE SIZE

The requirement, as in the case of the working class survey, was to fix the sample size that would ensure less than 2% error in the consumer price index number for each centre. Again, it was difficult to arrive at such a sample size without knowing the pattern of price relatives. However, after studying the NSS urban data, the results of the Faridabad township survey and the family budget surveys conducted in West Bengal, it was felt that a sample of 400 to 500 families would be required in the smaller centres and about 1000 families in the bigger cities. The larger heterogeneity in the consumption pattern and living habits of the middle class within the bigger centres warranted a larger scale of sampling in them as compared to the smaller centres.

The budget sanction expected for the survey allowed a field staff of about 150 investigators. It was decided that the easiest and possibly the only way of contacting the middle class families was

through tenement sampling. The work load for an investigator with the type of two stage design that was considered suitable for this survey involving listing of about 3 blocks per month was put at 20 families per month, or 240 a year. One fourth of the sample size was set apart for the study of conditions of living.

The allocation of the total sample size to the centres was made on a joint consideration of their 1951 census population and the expected middle class population as judged from census figures in the relevant occupation classes and the NSS estimates wherever available. It was found feasible to have at least two investigators to provide independent sub samples of 240 families each for each of the smaller centres also. The final allocation of sample size to each centre is shown in Table 16 2.

16 3c PRELIMINARY SURVEY

A preliminary enquiry was taken up for a few weeks in each centre for the purpose of studying the middle class concentrations in the centre fixing up the boundaries of the centre and collecting the frame required for sampling.

The tenement method of sampling was adopted in each centre since sampling middle class families through the payrolls of establishments was not found feasible. A block was the first stage unit and a middle class family was the second stage unit.

To cover the middle class population in each centre it was decided to go by the municipal or corporation limits of the centre. If well known middle class colonies had sprung up in the immediate neighbourhood of above mentioned limits, they were also included in the centre for the purpose of this survey.

The task of frame collection was therefore reduced to that of procuring a complete list of blocks covering the whole area of the centre falling within its boundaries defined as above. Very often this frame was the 1951 census list of blocks. For new middle class residential areas which had sprung up after the 1951 census, blocks were formed by the investigator along lines similar to that adopted

in the 1951 census, to make the frame complete. In some centres, it was found that the 1951 census list of blocks was wholly inadequate for this survey because of their obsolescence, unidentifiability and the large size of blocks. So an entirely new list of blocks was prepared by the field staff, sometimes with the help of the municipal divisions in the centre which were more up to date.

In centres common for the middle class and working class surveys, the information on the working class concentrations in different areas was used, as far as possible, to exclude the highly concentrated working class areas from the middle class survey. In some other centres also, where reliable information was readily available about some areas having very little middle class population, such areas were excluded from the sampling frame.

16.3d SAMPLING OF BLOCKS

In each centre, a systematic sample of 36 blocks was selected per sub-sample, the arrangement of the blocks in the frame being kept unaltered. Then the sample blocks were clustered into 12 clusters of 3 blocks each by grouping 1st, 13th and 25th sample blocks into one cluster, the 2nd, 14th and 26th sample blocks into another cluster, and so on. This method of clustering was adopted to ensure some measure of geographical spread in each cluster. The 12 clusters were then assigned to 12 months in a given order to ensure better representation for groups of months taken together. An independent sample was drawn for each sub-sample and was allotted to an investigator.

In cases where the blocks were very large, and too much time was taken for listing them or they were unidentifiable or were casualties because of lack of field staff, etc., the devices of sub-block division, random order of visit and substitution were adopted as in the case of tenement sampling for the working class survey (cf. Sub-section 16.2d).

16.3e SAMPLING OF FAMILIES

The investigator listed all the families in the blocks assigned for survey in a month and picked out the middle class families among them. These families were arranged by their size (single member families coming together and then the others) and within the size groups by monthly expenditure classes. A systematic sample of 20 families was selected from this arranged list and every fourth family among them was investigated for Schedule B (level of living) and the others were investigated for Schedule A (pattern of consumer expenditure).

In case a family could not be contacted due to various reasons or refused to give information in spite of all the means of persuasion at the disposal of the investigating officer, it was substituted by the next middle class family in the arranged list for that cluster. If in a sample cluster of a particular month the required number of middle class families (20) was not available, the deficiency was made up from the sampling frame for the next month's cluster for the same stratum (if any) and sub sample so as to keep the total sample size unaltered.

16.4 ESTIMATION PROCEDURES

This section gives the estimation procedures used for the working class and middle class family living surveys in the tenement and the payroll sampling centres.

16.4a TENEMENT SAMPLING CENTRES

Sampling schemes

Scheme 1 36 blocks per sub sample were selected systematically with equal probability and grouped into 12 clusters of 3 blocks each. Each sub sample was allotted to an investigator and each cluster in it was to be surveyed in a particular month by him. The sub samples were drawn independently for the different investigators. If some blocks were very large, the number of sub blocks to be formed in them was specified and the blocks were selected systematically with probability proportional to the number of such sub blocks. In any

sample block the number of sub-blocks actually formed for the survey depended on its population as observed in the field, and this could differ from the number specified earlier. In a block where such division was adopted, one of the sub-blocks was selected at random with equal probability for survey.

All the working/middle class families in the blocks or selected sub-blocks of a sample cluster were listed together and a sample of families was drawn from this list, systematically with equal probability for detailed enquiry. In some centres the sampling scheme was modified as follows. A random order of visit was assigned to the 3 blocks within a cluster and only 1, 2 or 3 of these blocks were surveyed depending on the total number of families listed and the number of working/middle class families obtained when the listing of each block (visited in the random order given) was completed.

Scheme 2 : A pps systematic sample of 12 clusters of 3 blocks each was drawn in some centres, size being the estimated number of working/middle class families and the procedure of random order of visit, sub-block selection and sampling of families within a cluster was the same as in scheme 1.

Scheme 3 : In some centres the blocks were grouped into 2 strata and selection of the allotted number of blocks or clusters was done independently in each stratum according to one of the schemes 1 and 2.

Notation

i : denotes a sample cluster of a sub-sample; j : denotes a block in the sample cluster; k : denotes a sampled family in a block of the sampled cluster; B : total number of blocks/specifyed sub-blocks; T : size adopted for pps sampling; F : total number of working/middle class families in the sample cluster (as obtained from the listing schedules); f : number of sample working/middle class families selected for a schedule in the sampled cluster; b : number of blocks surveyed in the sampled cluster; D : number of sub-blocks specified in the sample list for a block; D' : number of sub-blocks actually formed in a block; y : value of the characteristic under study.

Estimation Procedure

The estimate of the total of y for each stratum/centre for a month and sub sample for the different sampling schemes is given by $S S M_{ijk} y_{ijk}$ where M_{ijk} is the multiplier for the k th family in the j th block of the i th cluster (sub sample) and is obtained as shown below

$$\text{Scheme 1} \quad M_{ijk} = \frac{B}{b_i} \frac{F_i}{f_i} \frac{D_{ij}}{D_{ij}} \quad (16\ 1)$$

$$\text{Scheme 2} \quad M_{ijk} = \frac{T}{T_i} \frac{F_i}{f_i} \frac{B_i}{b_i} D_{ij} \quad (16\ 2)$$

Scheme 3 The estimate is the sum of the estimates in the two strata arrived at using the multipliers given for scheme 1 or 2 depending on the method of sampling adopted

Note (i) The sub sample wise estimate of the monthly average for a characteristic and that of the total number of families/persons in the centre were obtained by dividing the sum of the monthly estimates for that sub sample by 12 (by 6 for single investigator centre)

(ii) A simple average of the sub sample estimates was the combined estimate for the centre

(iii) The sub sample (or combined) estimate for the ratio of two characteristics is obtained by taking the ratio of the corresponding sub sample (or combined) estimates

16.4b PAYROLL SAMPLING CENTRES

Sampling schemes

Scheme 1 The establishments (factories/mines/plantations) in the centre were grouped to form clusters of 3, 4 or 5 establishments according to a suitable criterion and a sample of 12 such clusters per sub sample was selected pps systematically, size being the number of workers given in the sampling frame. Each sub sample was allotted to an investigator and the clusters were surveyed by him in the specified months. The sub samples were drawn independently for each investigator.

All the workers in a sample cluster were listed together and a systematic sample of workers was drawn from this list. The families to which these workers belonged were contacted for the detailed enquiry. (If two or more sample workers belonged to the same family, the information for that family was repeated as many times as it was selected).

Scheme 2 : In some centres clusters were not formed; the individual establishments were selected pps systematically and the number of workers to be sampled from each establishment was specified in the sample list. Sampling within each establishment was as within a cluster in scheme 1.

Scheme 3 : In some centres where the number of workers was not available in the frame for most of the establishments, 12 clusters of 3 or 4 establishments were selected systematically with equal probability. Sampling within a cluster was as in scheme 1.

Scheme 4 : In some centres the establishments were grouped to form 2 strata and the selection was done as in scheme 1, 2 or 3 in each stratum.

Notation

i : denotes a sample cluster of a sub-sample; *j* : denotes an establishment in the sample cluster; *k* : denotes a sample family in an establishment of the sample cluster; *W* : total number of workers as in the sampling frame; *W'* : total number of workers as obtained in the listing schedule; *E* : total number of establishments in the stratum/centre; *e* : number of establishments in a sample cluster; *f* : number of families surveyed; *w* : number of workers in a sample family; *y* : value of the characteristic under study.

Estimation Procedure

The estimate of the total *Y* for each stratum/centre for a month and sub-sample for the different sampling schemes is given by $\sum_{j} \sum_{k} M_{ijk} y_{ijk}$, where M_{ijk} is the multiplier for the *k*-th family in the

j-th establishment of the *i*-th cluster and is obtained as shown below.

$$\text{Scheme 1 : } M_{ijk} = \frac{W}{W_i} \cdot \frac{W'_i}{f_i} \cdot \frac{1}{w_{ijk}}. \quad \dots \quad (16.3)$$

$$\text{Scheme 2 : } M_{ijk} = \frac{1}{e_i} \cdot \frac{W}{W_{ij}} \cdot \frac{W'_{ij}}{f_{ij}} \cdot \frac{1}{w_{ijk}}. \quad \dots \quad (16.4)$$

$$\text{Scheme 3 : } M_{ijk} = \frac{E}{e_i} \cdot \frac{W_i}{f_i} \cdot \frac{1}{w_{ijk}} \quad (16.5)$$

Scheme 4 The estimate is the sum of the estimates in the two strata arrived at using the multipliers given for schemes 1, 2 or 3 depending on the method of sampling adopted.

The note given on page 590 for tenement sampling centres applied to the payroll sampling centres also.

Bibliography

An attempt has been made to give in this *Bibliography* a consolidated list of papers and books on sampling theory and methods. Articles and reports giving mainly results or analysis of surveys without any direct contribution to sampling methodology have been excluded. A list of journals with their abbreviations used here is given at the end. All the papers have been broadly classified by the following categories of the field of application and the method of sampling.

<i>subject</i>	<i>code</i>	<i>method</i>	<i>code</i>
methodology	M	general	1
population and demography	P	historical review	2
level of living surveys	L	simple random sampling	3
socio-economic enquiries (other than P and L)	S	determination of sample size	4
agricultural economics	A	systematic sampling	5
crop statistics	C	varying probability sampling	6
forest surveys	F	stratified sampling	7
industrial and other establishment surveys	E	cluster sampling	8
opinion and attitude surveys	O	multi-stage sampling	9
wild and other populations	W	ratio method of estimation	10
		multi-phase sampling	11
		regression estimator	
		self-weighting design	12
		non-sampling errors	13
		other sampling designs	14

For instance, the code M-10 against a paper indicates that the paper deals primarily with the theoretical and methodological aspects of ratio method of estimation and the code F-5 denotes that the paper deals mainly with systematic sampling as applied to forest surveys. Though efforts have been made to include most of the relevant papers and to classify them properly, it may be pointed out that there is still the possibility of omission or misclassification of some papers in a work of this nature.

This bibliography originally issued as a mimeographed publication in 1962 was compiled by the author in collaboration with Mrs. B. N. Chinnappa, Dr. D. N. Ghosh, Mr. A. S. Roy and Mr. G. Parthasarathy, and Mr. M. P. Singh has helped in up-dating it from 1962 to 1965-66.

- ACKOFF, R. L. and PRITZKER, L. (1951) The methodology of survey research, *IJOAR*, 5, 313-334 O- 1
- ADAMS, J. S. (1956) An experiment on question and response bias, *POQ*, 20, 593-598 S-13
- AGARWAL, O. P. (1959) Bayes and minimax procedures in sampling from finite and infinite populations, *AMS*, 30, 206-218 M- 1
- AIRTH, J. M. (1958) See Fleischer, J.
- ALLEN, R. G. D. (1964) Sampling for current economic statistics, *JRSS*, (A), 127, 76-88 M- 1
- ALPERT, H. (1952) Some observations on the sociology of sampling, *Social Forces*, 31, 30-33 S- 1
- ANDERSON, P. H. (1942) Distributions in stratified sampling, *AMS*, 13, 42-52 M- 7
- ANDERSON, R. L. (1954) See Sen, A. R.
- ANGSTROM, K. H. (1958) An asymptotic expansion of bias in a non linear function of a set of unbiased characteristics from a finite sample, *Skand Alt.*, 41, 40-46 M- 1
- ANSCOMBE, F. J. (1948) On estimating the population of aphids in a potato field, *AAB*, 35, 567-571 W- 1
- ANSCOMBE, F. J. (1950) Soil sampling for potato root eel worm cysts, *AAB*, 37, 286-295 W- 1
- ANTLE, C. E. (1965) See Folks, J. L.
- Aoyama, H. (1951) On practical systematic sampling, *AISM*, 3, 57-64 M- 5
- Aoyama, H. (1954a) A study of stratified random sampling, *AISM*, 6, 1-36 M- 7
- Aoyama, H. (1954b) On the interviewing bias, *AISM*, 5, 73-76 M-13
- Aoyama, H. (1963) Stratified random sampling with optimum allocation for multivariate population, *AISM*, 14, 251-258 M- 7
- ARMITAGE, P. (1947) A comparison of stratified with unrestricted random sampling from a finite population, *Biometrika*, 34, 273-280 M- 7
- AROLAN, L. A. (1944) Some methods for the evaluation of a sum, *JASA*, 39, 511-515 M- 3
- ASHFORD, J. R. (1958) The design of a long term sampling programme to measure the hazard associated with an industrial environment, *JRSS*, (A), 121, 333-347 S- 1
- AVADHANI, M. S. and SUKHATME, B. V. (1965) Controlled simple random sampling, *JISAS*, 17, 34-42 M- 3
- AVADHANI, M. S. and SUKHATME, B. V. (1966) A note on the ratio and regression methods of estimation in controlled simple random sampling, *JISAS*, 18, 17-20 M-10
- BAILEY, N. T. J. (1951) On estimating the size of mobile populations from recapture data, *Biometrika*, 38, 293-306 W-14
- BAKER, E. F. and FLEMING, W. E. (1936) A method for estimating populations of larvae of the Japanese beetle in the fields, *JAR*, 53, 319- W- 1
- BAKER, E. F. (1949) The variance of the proportions of samples falling within a fixed interval for a normal population, *AMS*, 20, 123-126 M- 3

- BAKER, R. L. and BYLUND, H. B. (1957): Consumer survey versus store data for determining egg production; *JFE*, 39, 770-777. E-13
- BALABAN, V. (1961): See Macura, M.
- BALAKRISHNAN, T. R. (1966): See Hess, I.
- BANCROFT, G. (1954): Special uses of the current population survey mechanism; *Estadistica*, 12, 198-206. S- 1
- BANCROFT, T. A. (1963): See Clyde, R. W.
- BANERJEE, K. S. and ROY, S. N. (1940): On hierarchical sampling, hierarchical variances and connection with other aspects of statistical theory; *Science and Culture*, 6, 189. M- 1
- BANERJEE, K. S. (1955): A note on successive sampling; *BCSA*, 6, 35-39. M-11
- BANERJEE, K. S. (1958): Probability selection of the constituent items of a composite item in the construction of cost of living index numbers; *BCSA*, 8, 104-109. L- 1
- BANERJEE, K. S. (1959): Precision in the construction of cost of living index numbers; *Sankhyā*, 21, 393-400. S- 1
- BANERJEE, K. S. (1960): Calculation of sampling errors for index numbers; *Sankhyā*, 22, 119-130. S- 1
- BANERJEE, P. K. (1952): Error of one-stage and two-stage sampling; *BCSA*, 4, 74-78. M- 9
- BANERJEE, S. K. (1934): See Mahalanobis, P. C.
- BANKS, S. (1948): See Meier, N. C.
- BARBEBI, B. (1958): Some application of the sampling method in Italian official statistics; *BISI*, 36, (3), 107-112. M- 1
- BARNARD, M. M. (1936): An examination of the sampling observations on wheat of the Crop-weather Scheme; *JAS*, 26, 456-487. C- 1
- BARTHOLOMEW, D. S. (1961): A method of allowing for 'not-at-home' bias in sample surveys; *Appl. Stat.*, 10, 52-59. M-13
- BARTLETT, M. S. (1937): Sub-sampling for attributes; *JRSS Supplement*, 4, 131-135. M- 9
- BASAVARAJAPPA, K. G. and RAMACHANDRAN, K. V. (1963): A note on testing interviewer difference in sampling design; *JISA*, 1, 32-39. M-13
- BASU, D. (1954): On the optimum character of some estimators used in multi-stage sampling problems; *Sankhyā*, 13, 363-368. M- 9
- BASU, D. (1958): On sampling with and without replacement; *Sankhyā*, 20, 287-294. M- 1
- BAUR, E. J. (1947): Response bias in a mail survey; *POQ*, 11, 594-600. O-13
- BAXTER, R. (1964): An inquiry into the misuse of the survey technique by sales solicitors; *POQ*, 20, 124-134. O- 1
- BEALL, G. (1939): Methods of estimating the population of insects in a field; *Biometrika*, 30, 422-439. W- 1
- BECKER, J. A. and HARLAN, C. L. (1939): Developments in crop and livestock reporting since 1920; *JFE*, 21, 799-827. C- 2
- BECKER, M. E. (1950): Forest survey procedures for area and volume determination; *J. Forestry*, 48, 465-469. F- 1

- BEIERMANN, H I (1954) Sampling technique in an economic survey of sugar-cane production, *South African J Economics*, 22, 326-336 A- 1
- BELLOC, N B (1954) Validation of morbidity survey data by comparison with hospital records, *JASA*, 49, 832-846 S-13
- BEVE, L (1958) Complete enumeration and sampling surveys in the population censuses, *Demografia*, 1, 161-181 P- 1
- BENJAMIN, B (1960) Statistical problems connected with the 1961 population census, *JRSS, (A)*, 123, 413-426 P- 1
- BENJAMIN, B (1961) The 1961 census of population, *Ind Stat*, 11, 130-143 P- 1
- BENNETT, B M (1957) Note on the method of inverse sampling; *Trab Est*, 8, 29-31 M- 1
- BENSON, L E (1946) Mail surveys can be valuable, *POQ*, 10, 234-241 O- 1
- BENSON, P H and BENTLEY, E (1957) Sources of sampling bias in sex studies, *POQ*, 21, 388-394 S-13
- BENTLEY, E (1957) See Benson, P H
- BERNERT, E H (1945) See Hagoood, M J
- BERSHAD, M A (1961) See Hansen, M H
- BEVILLE, H M (Jr.) (1949) Surveying radio listeners by use of a probability sample, *J Marketing*, 14, 373-384 O- 1
- BEYLEVELD, A J (1929) Determination of a precise indication of change in crop acreage, *JASA*, 24, 405-411 C- 1
- BHARGAVA, R P (1951) See Nair, K R
- BHATTACHARYA N (1958) See Sengupta, J M
- BICKERSTAFF A (1947a) The measurement of growth on forest areas by means of recurrent line plot surveys *Forest Chronicle*, 23, 36-43 F- 1
- BICKERSTAFF, A (1947b) One fifth acre versus one tenth acre plots in sampling immature stands, Dominion Forest Service, Canada, *Silviculture Research Note* 83 F- 1
- BICKERSTAFF, A (1947c) Sampling efficiency of line plot survey on Riding Mountain research area, Dominion Forest Service Canada, *Silviculture Research Note*, 84 F- 1
- BICKFORD, C A (1952) The sampling design used in the forest survey of the North East, *J Forestry*, 50, 290-293 F- 1
- BILLETER, E P (1956) Optimum design in mixed sampling plan *RISI*, 24, (1-3) 73-76 M- 7
- BIRNBAUM, Z W and SIRKEN, M G (1950a) Bias due to non availability in sampling surveys, *JASA*, 45, 98-111 M-13
- BIRNBAUM, Z W and SIRKEN, M G (1950b) On the total error due to random sampling, *IJOAR*, 4, 179-191 M-13
- BIRNBAUM, Z W and HEALY, W C (1960) Estimates with prescribed variance based on two stage sampling, *AMS*, 31, 662-676 M- 9
- BIRNBAUM, Z W and SIRKEN, M G (1965) Design of sample surveys to estimate the prevalence of rare diseases—three unbiased estimates, *U S Public Health Service, Series 2*, No 11 S- 1

- BLASER, R. E. (1948): *See* Rigney, J. A.
- BLOCK, E. (1958): Numerical considerations for the stratification of variables following a logarithmic normal distribution; *Skand. Akt.*, 4, 185-200. M- 7
- BLYTHE, R. H. (Jr.) (1945): The economics of sample size applied to the scaling of sawlogs; *Biometrics*, 1, 67-70. F- 1
- BLYTHE, R. H. (1947): *See* Jessen, R. J. (1947b).
- BLYTHE, R. H. (1951): *See* Rosander, A. C.
- BOLAS, B. D. (1956): *See* Freeman, G. H.
- BOOKER, H. S. and DAVID, S. T. (1952): Differences in results obtained by experienced and inexperienced interviews; *JRSS*, (A), 115, 232-257. S-13
- BORUS, M. E. (1966): Response error in survey reports of earnings information; *JASA*, 61, 729-738. S-13
- BOSE, C. (1943): Note on the sampling error in the method of double sampling; *Sankhyā*, 6, 329-330. M-11
- BOSE, C. (1951): Some further results on errors in double sampling technique; *Sankhyā*, 11, 191-194. M-11
- BOSE, S. S. (1934): *See* Mahalanobis, P. C.
- BOSE, S. S. (1936): *See* Mahalanobis, P. C. (1936b).
- BOWLEY, A. L. (1926): Measurement of the precision attained in sampling; *BISI*, 22, (1), 1-62. M- 1
- BOWLEY, A. L. (1936): The application of sampling to economic and socio-logical problems; *JASA*, 31, 474-480. S- 1
- BOX, K. and THOMAS, G. (1944): The war time social surveys; *JRSS*, 107, 151-189. S- 1
- BOYARSKY, A. Y. (1958): An experiment in the theory of a census with control rounds; *RISI*, 26, (1/3), 48-55. P- 1
- BOYD, H. W. (Jr.) and WESTFALL, R. (1955): Interviewers as a source of error in surveys; *J. Marketing*, 19, 311-324. M-13
- BRANKO, B. (1958): A formula for the upper bound of relative variance in a simple case of two-stage sampling; *Statistical Review*, Belgrade, 9, 298-301. M- 9
- BRAYER, E. F. (1957): Calculating the standard error of a proportion; *Appl. Stat.*, 6, 67-68. M- 3
- BREWBAKER, H. E. and BUSH, H. L. (1942): Pre-harvest estimate of yield and sugar percentage based on random sampling technique; *Annals of American Society of Sug. & Technology*, 1-13. C- 1
- BREWER, K. R. W. and UNDY, G. C. (1962): Samples of two units drawn with unequal probabilities without replacement; *AJS*, 4, 89-100. M- 6
- BREWER, K. R. W. (1963a): A model of systematic sampling with unequal probabilities; *AJS*, 5, 5-13. M- 6
- BREWER, K. R. W. (1963b): Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process; *AJS*, 5, 93-105. M-10
- BREYER, R. F. (1946): Some preliminary problems of sample design for a survey of retail trade flow; *J. Marketing*, 11, 343-353. E- 1

- BRILLINGER, D R (1963) A note on re use of samples, *AMS*, 34, 341-342 M- 1
- BROOKS, E M (1953) Planning and operating sample surveys, *Estadística*, 11, 63-71 M- 1
- BROOKS, S H (1955) The estimation of an optimum sub sampling number, *JASA*, 50, 398-415 M- 9
- BROWN, G H (1947) A comparison of sampling methods, *J Marketing*, 12, 331-337 M- 1
- BROWN, T H (1942) Scientific sampling in business, *Harvard Business Review*, 358-368 E- 1
- BROWN, T H (1951) *See* Leavens, D H
- BROWNLEE, K A (1957) A note on the effects of non response on surveys, *JASA*, 52, 29-32 M-13
- BRUNK, M E and FEDERER, W T (1953) Experimental designs and probability sampling in marketing research, *JASA*, 48, 440-452 O- 1
- BRUNSMEN, H G (1944) The sample census of congested production areas, *JASA*, 39, 303-310 P- 1
- BRYANT, E C, HARTLEY, H O and JESSEN, R J (1960) Design and estimation in two way stratification, *JASA*, 55, 105-124 M- 7
- BRYSON, M R (1961) Physical inventory using sampling methods, *Appl Stat*, 9, 178-188 E- 1
- BRYSON, M R (1965) Errors of classification in Binomial population, *JASA*, 60, 217-224 M-13
- BUCKLAND, W R (1951) A review of the literature of systematic sampling, *JRSS, (B)*, 13, 208-215 M-2
- BULL, H (1932) *See* Schumacher F X
- BURKE, C J (1947) *See* Meier, N C
- BURKE, C J (1948) *See* Meier, N C
- BURROWS, W D (1953) The problem of exceptional samples in agricultural sample surveys, *JRSS, (A)*, 116, 175-176 C- 1
- BURROWS, W D (1957) The problem of exceptional samples in agricultural sample surveys, *Estadística*, 15, 601-604 C-1
- BUSH, H L (1942) *See* Brewbaker, H E
- BYLUND, H B (1957) *See* Baker, R L
- CALLANDER, W F and SABLE, C F (1947) The Bureau of the Agricultural Economics programmes in enumerative sampling, *JFE*, 29, 233-236 A- 1
- CAMERON, J M (1951) The use of components of variance in preparing schedules for the sampling of baled wool, *Biometrics*, 7, 83-96 E- 1
- CANSADO, E (1952) Expectations and variances in multi stage sampling, *Trab Est*, 3, 27-41 M- 9
- CANSADO, E (1957) Sampling without replacement from finite populations, *Trab Est*, 8, 3-12 M- 6
- CANTRIL, H (1945) Do different polls get the same results? *POQ*, 9, 61-69 O- 1
- CARTER, R E (JR), TROLDAHL, V C and SCHUNEMAN, R S (1963) Interviewer bias in selecting households, *J. Marketing*, 27, (2), 27-34 M-13

- CARTWRIGHT, A. (1957): The effect of obtaining information from different informants on a family morbidity enquiry; *Appl. Stat.*, 6, 18-25. S-13
- CARTWRIGHT, A. (1959a): Some problems in the collection and analysis of morbidity data from sample surveys; *Milbanks Memorial Foundation Quarterly*, 37, 33-48. S- 1
- CARTWRIGHT, A. (1959b): The families and individuals who did not co-operate in a sample survey; *Milbanks Memorial Foundation Quarterly*, 37, 347-368. M-13
- CARVER, H. C. (1930): Fundamentals of the theory of sampling; *AMS*, 1, 101-121, 260-274. M- 1
- CASSADY, R. (Jr.) (1945): Statistical sampling techniques and marketing research; *J. Marketing*, 9, 317-341. O- 1
- CATTON, W. R. (Jr.) (1959): See Larson, R. F.
- CHAKRABARTY, R. P. (1964): See Sen, A. R.
- CHAKRABORTY, P. N. (1963): On a method of estimating birth and death rates from several agencies; *BCSA*, 12, 106-112. P-14
- CHAKRAVARTY, I. M. (1951): See Sengupta, J. M. (1951a).
- CHAKRAVARTY, I. M. (1952): Use of analysis of covariance in two-stage sampling; *BCSA*, 4, 127-129. M- 9
- CHAKRAVARTY, I. M. (1954): On the problem of planning a multi-stage survey for multiple correlated characters; *Sankhyā*, 14, 211-216. M- 9
- CHAKRAVARTY, I. M. (1960): See Roy, J.
- CHAMBERS, M. L. and JARRATT, P. (1964): Use of double sampling for selecting best population; *Biometrika*, 51, 49-64. M-11
- CHANDA, K. (1952): A note on the comparative efficiencies of selection of sampling units with and without replacement; *Science and Culture*, 18, 288-289. M- 4
- CHANDLER, K. N. and TANNER, J. C. (1958): Estimates of the total miles run by road vehicles in Great Britain; *JRSS*, (A), 121, 420-437. E- 1
- CHAPMAN, D. G. (1951): Some properties of the hyper-geometric distribution with applications to Zoological sample censuses; *University of California Publication on Statistics*, 1, 131-159. W- 1
- CHAPMAN, D. G. (1952): Inverse, multiple and sequential sample censuses; *Biometrics*, 8, 286-306. M-14
- CHAPMAN, D. G. (1956): See Junge, C. D. (Jr.).
- CHAPMAN, R. A. and SCHUMACHER, F. X. (1948): Sampling methods in forest and range management; *N. C. Duke University School of Forestry Bulletin*, No. 7. F- 1
- CHAWLA, H. K. (1956): See Rao, J. N. K.
- CHEVRY, G. (1949): Control of a general census by means of an area sampling method, *JASA*, 44, 373-379. E-13
- CHIAN, C. L. (1951): On design of mass medical surveys; *Human Biology*, 23, 242-271. S- 1
- CHIANG, C. A. (1949): Using the Pao as the primary sampling unit : some notes and reflections on the possibilities of a census of China by sampling; *Population Studies*, 2, 444-453. P- 1

- CHIKKAGOUDE, M S (1966a) A note on inverse sampling with equal probabilities, *Sanjhyā*, 28, (A), 93-96 M-14
- CHIKKAGOUDE, M S (1966b) A note on sampling with varying probabilities, *JISAS*, 18, 86-92 M- 6
- CHITTY, D (1951) See Leslie, P H
- CHITTY, D (1953) See Leslie, P H
- CHITTY, H (1953) See Leslie, P H
- CHURCH, B M (1954) Problems of sample allocation and estimation in an agricultural survey, *JRSS*, (B), 16, 223-235 C- 1
- CLAPHAM, A R (1929) The estimation of yield in cereal crops by sampling methods, *JAS*, 19, 214-235 C- 1
- CLAPHAM, A R (1929) See Wishart, J
- CLAPHAM, A R (1931a) Studies in sampling technique—Cereal experiments I, Field technique, *JAS*, 21, 366-371 C- 1
- CLAPHAM, A R (1931b) Studies in sampling technique—Cereal experiments III, Results and discussion, *JAS*, 21, 376-390 C- 1
- CLAUSEN, J A and FORD, R N (1947) Controlling bias in mail questionnaires, *JASA*, 42, 497-511 M-13
- CLEMENT, D V P (1956) See Taylor, W B
- CLYDE, R W, HEMMARLÉ, W J and BANCROFT, T A (1963) An application of 'post stratification' technique in local TV election predictions, *POQ*, 27, 467-472 O- 7
- COALE, A J (1955) The population of the United States in 1950 classified by age, sex and color—a revision of census figures, *JASA*, 50, 16-54, 1331 P-13
- COATS, R H (1931) Enumeration and sampling in the field of the census, *JASA*, 26, 270-284 S- 1
- COCHRAN, W G and WATSON, D J (1936) An experiment on observer's bias in the selection of shoot heights, *Empire Journal of Experimental Agriculture*, 4, 69-76 C-13
- COCHRAN, W G (1938a) See Irwin, J O
- COCHRAN, W G (1938b) The information supplied by the sampling results, *AAB*, 25, 383-389 M- 1
- COCHRAN, W G (1939) The use of analysis of variance in enumeration by sampling, *JASA*, 34, 492-510 M- 1
- COCHRAN, W G (1940) The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce, *JAS*, 30, 262-275 C- 1
- COCHRAN, W G (1942) Sampling theory when the sampling units are of unequal sizes, *JASA*, 37, 199-212 M-11
- COCHRAN, W G (1946) Relative accuracy of systematic and stratified random samples for a certain class of population, *AMS*, 17, 164-177 M- 5
- COCHRAN, W G (1947) Recent developments in sampling theory in the United States, *Proceedings of the International Statistical Conference*, 3, (A), 40-66 M- 1

BIBLIOGRAPHY

601

- COCHRAN, W. G. (1951) : Modern methods in the sampling of human population: general principles in the selection of a sample; *AJPH*, 41, 647-653. M- 1
- COCHRAN, W. G., MOSTELLER, F. and TUKEY, J. W. (1953a) : Statistical problems of the Kinsey Report; *JASA*, 48, 673-716. M- 1
- COCHRAN, W. G. and CONROLL, S. P. (1953b) : A sampling investigation of the efficiency of weighting inversely as the estimated variance; *Biometrics*, 9, 447-459. M- 1
- COCHRAN, W. G., MOSTELLER, F. and TUKEY, J. W. (1954) : Principles of sampling; *JASA*, 49, 13-35. M- 1
- COCHRAN, W. G. (1961) : Comparison of methods for determining stratum boundaries; *BISI*, 38, (2), 345-358. M- 7
- COCHRAN, W. G. (1962) : See Rao, J. N. K. (1962b).
- CODY, D. D. (1948) : Sampling errors in mortality and other statistics in life insurance; *JASA*, 43, 442-450. S- 1
- COHEN, S. E. and LIPSTEIN, B. (1954) : Response errors in the collection of wage statistics by mail questionnaire; *JASA*, 49, 240-250. S-13
- COLE, D. (1953) : See Utting, J. E. G.
- COLE, D. (1954) : See Utting, J. E. G.
- COLE, D. (1956) : See Utting, J. E. G.
- CONROLL, S. P. (1953) : See Cochran, W. G. (1953b).
- CORDRILL, W. N. (1949) : The commercial use of probability samples; *J. Marketing*, 14, 447-449. O- 1
- CORLETT, T. (1950) : See Gray, P. G.
- CORLETT, T. (1952) : A use for the jury qualification in sample design; *Appl. Stat.*, 1, 34-36. M- 1
- CORLETT, T. (1963) : Rapid methods of estimating standard errors of stratified multi-stage samples—a preliminary investigation; *The Statistician*, 13, 5-16. M- 9
- CORMICK, T. C. (1937) : Sampling theory in sociological research; *Social Forces*, 16, 67-74. S- 1
- CORNELL, F. G. (1947) : A stratified random sample of a small finite population; *JASA*, 42, 523-532. M- 7
- CORNFIELD, J. (1942) : On certain biases in samples of human populations; *JASA*, 37, 63-68. P-13
- CORNFIELD, J. (1944) : On samples from finite populations; *JASA*, 39, 236-239. M- 3
- CORNFIELD, J. (1951) : Determination of sample size; *AJPH*, 41, 654-. M- 4
- CORSA, L. (JR.) (1964) : The sample survey in a national population programme, *POQ*, 28, 383-388. P- 1
- COSTELLO, D. F., WILM, N. G. and KLIPPLE, C. E. (1944) : Estimating forage yield by the double samples method; *Journal of American Society of Agronomy*, 36, 194-203. C- 1
- COX, D. R. (1952) : Estimation by double sampling; *Biometrika*, 39, 217-227. M-11
- CRAIG, A. T. (1939) : On the mathematics of the representative method of sampling; *AMS*, 10, 26-34. M- 1

- CRAIG, C C (1953a) On the utilization of marketed specimens in estimating populations of flying insects, *Biometrika*, 40, 170-176 W- 1
- CRAIG, C C (1953b) On a method of estimating biological populations in the field, *Biometrika*, 40, 216-218 W- 1
- CROSETTI, A H and SCHMITT, R C (1956) A method of estimating the intercensal population of countries, *JASA*, 51, 587-590 P- 1
- CROSSLEY, A M (1941) Theory and application of representative sampling as applied to marketing, *J. Marketing*, 5, 456-461 O- 1
- CROSSLEY, A M (1950) See Parry, H J
- CRUM, W L (1933) An analytical interpretation of straw root samples, *JASA*, 28, 152-163 O- 1
- CRUMP, S L (1948) See Nordakog, A W
- CYERT, R M (1954) See Trueblood R M
- CYERT, R M and TRUEBLOOD, R M (1957) Statistical sampling techniques in the ageing of accounts receivable in a department store, *Management Science*, 3 185-195 E- 1
- DALENIUS T (1950) The problem of optimum stratification—I, *Skand Akt*, 33, 203-213 M- 7
- DALENIUS, T and GURNEY, M (1951) The problem of optimum stratification—II, *Skand Akt*, 34 133-148 M- 7
- DALENIUS, T (1952) The problem of optimum stratification in a special type of design, *Skand Akt*, 35, 61-70 M- 7
- DALENIUS, T (1953a) Multivariate sampling problem, *Skand Akt*, 36, 92-122 M- 1
- DALENIUS, T (1953b) The economics of one stage stratified sampling, *Sankhya*, 12, 351-356 M- 7
- DALENIUS, T (1956) The Survey Research Centre of the Central Bureau of Statistics Sweden, *Sankhya*, 17, 225-244 M- 1
- DALENIUS, T (1957a) Possibilities and limits of sampling in regional enquiries, *BISI*, 35, (4), 337-353 M- 1
- DALENIUS, T and HODGES, J L (JR) (1957b) The choice of stratification points, *Skand Akt*, 40 198-203 M- 7
- DALENIUS, T and HODGES, J L (JR) (1959) Minimum variance stratification, *JASA*, 54 88-101 M- 7
- DALENIUS, T (1960) Training in sampling for a government statistical system, *BISI*, 37, (2), 201-217 M- 1
- DALENIUS, T, HAJEK, J and ZUBRIZZEKI, S (1961) On plane sampling and related geometrical problems, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 125-150, University of California Press M-14
- DALENIUS, T (1962) Recent advances in sample survey theory and methods, *AMS*, 33, 325-349 M- 2

- DALENTIUS, T. (1963) : Lattice sampling by means of Lahiri's sampling scheme; *Contributions to Statistics*, 49-56, Presented to Professor P. C. Mahalanobis on the occasion of his 70th Birthday, Pergamon Press, London, and Statistical Publishing Society, Calcutta. M-14
- DALY, J. F. (1949) : See Jessen, R. J.
- DAMES, J. (1962) : See Wells, D. W.
- DARROCH, J. N. (1958) : The multiple recapture census, I—Estimation of a closed population; *Biometrika*, 45, 343-359. W-14
- DARROCH, J. N. (1959) : The multiple recapture census II—Estimation when there is immigration or death; *Biometrika*, 46, 336-351. W-14
- DAS, A. C. (1949) : Two dimensional systematic sampling; *Science and Culture*, 15, 157-158. M- 5
- DAS, A. C. (1950) : Two dimensional systematic sampling and the associated stratified and random sampling; *Sankhyā*, 10, 95-108. M- 5
- DAS, A. C. (1951a) : On two phase sampling and sampling with varying probabilities; *BISI*, 33, (2), 105-112. M- 6
- DAS, A. C. (1951b) : Systematic sampling; *BISI*, 33, (2), 119-132. M- 5
- DAS, A. C. (1962) : On MVU estimates of parameters of a finite population based on varying probability samples; *BCSA*, 11, 39-48. M- 6
- DAS, N. C. (1959) : See Som, R. K. (1959b).
- DAS GUPTA, A. K. and DUTTA, N. C. (1951) : Public Preference Survey in Calcutta in 1941; *BISI*, 33, (5), 377-384. O- 1
- DASGUPTA, P. (1964) : On the estimation of the total number of units and of the probabilities of detecting an event from information supplied by different agencies; *BCSA*, 13, 89-99. P-14
- DAVID, M. (1962) : The validity of income reported by a sample of families who received welfare assistance during 1959; *JASA*, 57, 680-685. L-13
- DAVID, S. T. (1952) : See Booker, H. S.
- DEDRICK, C. L. (1948) : Some problems of the 1950 Census of the Americas; *Estadística*, 6, 354-359. P- 1
- DELURY, D. B. (1947) : On the estimation of biological populations; *Biometrics*, 3, 145-167. W- 1
- DEMING, W. E. (1940a) : See Stephan, F. F. (1940a).
- DEMING, W. E. and STEPHAN, F. F. (1940b) : On a least square adjustment of a sampled frequency table when the expected marginal totals are known; *AMS*, 11, 427-444. M- 1
- DEMING, W. E. and STEPHAN, F. F. (1941a) : On the interpretation of censuses of samples; *JASA*, 36, 45-49. M- 1
- DEMING, W. E. and GEOFREY, L. (1941b) : On sample inspection in the processing of census returns; *JASA*, 36, 351-360. M-13
- DEMING, W. E. and GURNEY, M. (1943a) : Government standards of sampling practice in the United States; *Estadística*, 1, 124-126. M- 1
- DEMING, W. E. (1943b) : See Hansen, M. H. (1943a).
- DEMING, W. E. (1943c) : See Tepping, B. J.

- DEMING, W E (1944) On errors in surveys, *ASR*, 9, 359-369 S-13
- DEMING, W E (1945) On training in sampling, *JASA*, 40, 307-316 M- 1
- DEMING, W E and SIMMONS, W R (1946) On the design of a sample for dealers' inventories, *JASA*, 41, 16-33 E- 1
- DEMING, W E (1947a) Some criteria for judging the quality of surveys, *J Marketing* 12 145-157 M-13
- DEMING W E (1947b) See Jessen, R J
- DEMING, W E (1949a) See Sekhar, C C
- DEMING W E (1949b) See Jessen, R J
- DEMING, W E (1950a) On the sampling of physical materials, *BISI*, 18, (1-2), 1-20 E- 1
- DEMING, W E (1950b) See Hansen, M H
- DEMING, W E (1953a) On the distinction between enumerative and analytic surveys, *JASA*, 48, 244-245 M- 1
- DEMING, W E (1953b) On a probability mechanism to attain an economic balance between the resultant error of response and the bias of non-response, *JASA*, 48, 743-772 M-13
- DEMING, W E (1954) On the presentation of results of sample surveys as legal evidence, *JASA*, 49 814-825 M- 1
- DEMING, W E (1956) On simplification of sampling design through replication with equal probabilities and without stages, *JASA*, 51, 24-53 M- 1
- DEMING, W E and GLASSER, G J (1959) On the problem of matching lists by samples, *JASA*, 54, 403-415 M-14
- DEMING, W E (1961) Uncertainties in statistical data and their relation to the design and management of statistical surveys, *BISI*, 38 (4), 365-383 M- 1
- DEMING, W E (1963a) Some stratified sampling plans in replicated designs, *Estadística*, 21, 716-738 M- 7
- DEMING, W E (1963b) On some of the contributions of interpenetrating networks of samples, *Contributions to Statistics*, 57-66, Presented to Professor P C Mahalanobis on the occasion of his 70th Birthday, Pergamon Press, London, and Statistical Publishing Society, Calcutta M-13
- DENT, J K (1950) See Lansing, J B
- DES RAJ (1954) Ratio estimation in sampling with equal and unequal probabilities, *JISAS*, 6, 127-138 M- 6
- DES RAJ (1956a) On the method of overlapping maps in sample surveys, *Sankhya*, 17, 89-98 M- 6
- DES RAJ (1956b) A note on the determination of optimum probabilities in sampling without replacement, *Sankhya*, 17, 197-200 M- 6
- DES RAJ (1956c) Some estimators in sampling with varying probabilities without replacement, *JASA*, 51, 269-284 M- 6
- DES RAJ (1957) On estimating parametric functions in stratified sampling designs, *Sankhya*, 17, 361-366 M- 7
- DES RAJ (1958a) On the relative accuracy of some sampling techniques, *JASA*, 53, 98-101, M- 6

- DES RAJ and KHAMIS, H. S. (1958b) : Some remarks on sampling with replacement; *AMS*, 29, 550-557. M- 3
- DES RAJ (1962) : On matching lists by samples; *JASA*, 56, 251-255. M-14
- DES RAJ (1963) : Some apparently unconnected problems encountered in sampling work; *Contributions to Statistics*, 67-72, Presented to Professor P. C. Mahalanobis on the occasion of his 70th Birthday, Pergamon Press, London, and Statistical Publishing Society, Calcutta. M- 6
- DES RAJ (1964a) : On double sampling for pps estimation; *AMS*, 35, 900-902. M-11
- DES RAJ (1964b) : The use of systematic sampling with probability proportionate to size in a large scale survey; *JASA*, 59, 251-255. M- 6
- DES RAJ (1964c) : On forming strata of equal aggregate size; *JASA*, 59, 481-486. M-7
- DES RAJ (1964d) : A note on the variance of the ratio estimate; *JASA*, 59, 895-898. M-10
- DES RAJ (1964e) : On sampling with probability proportionate to aggregate size; *JISAS*, 16, 317-319. M- 6
- DES RAJ (1965a) : On a method of using multi-auxiliary information in sample surveys; *JASA*, 60, 270-277. M-11
- DES RAJ (1965b) : On sampling over two occasions with probability proportionate to size; *AMS*, 36, 327-330. M-11
- DES RAJ (1965c) : Variance estimation in randomized systematic sampling with probability proportionate to size; *JASA*, 60, 278-284. M- 6
- DHANALAL (1940) : See Kalamkar, R. J. (1940b).
- DIEUTE-FAIT, C. E. (1942) : Note on a method of sampling; *AMS*, 13, 94-97. M- 1
- DONALD, M. N. (1960) : Implications of non-response for the interpretation of mail questionnaire data; *POQ*, 24, 99-114. O-13
- DOWNHAM, J. S. (1954) : Social class in sample surveys; *Inc. Stat.*, 5, 17-38. S- 1
- DURANT, H. (1954) : The Gallup poll and some of its problems; *Inc. Stat.*, 5, 101-112. O- 1
- DURBIN, J. (1950) : See Stone, J. R. N.
- DURBIN, J. and STUART, A. (1951) : Differences in response rates of experienced and inexperienced interviewers; *JRSS*, (A), 114, 163-206. S-13
- DURBIN, J. (1953) : Some results in sampling theory when the units are selected with unequal probabilities; *JRSS*, (B), 15, 262-269. M- 6
- DURBIN, J. (1954a) : Non-response and call-backs in surveys; *BISI*, 34, (2), 72-86. M-13
- DURBIN, J. and STUART, A. (1954b) : Call-backs and clustering in sample surveys; *JRSS*, (A), 117, 387-428. S-13
- DURBIN, J. (1958) : Sampling theory for estimates based on fewer individuals than the number selected; *BISI*, 36, (3), 113-119. M- 1
- DURBIN, J. (1959) : A note on the application of Quenouille's method of bias reduction to the estimation of ratios; *Biometrika*, 46, 477-480. M-10
- DUTTA, N. C. (1951) : See Das Gupta, A. K.
- DUTTA, N. C. (1964) : See Sengupta, J. M.

- EAPEN, A. T (1959) *See* Lansung, J. B
- EBLINE, W H (1953) Some problems of tabulation in agricultural mail sampling, *Estadística* 11, 103-111 C- 1
- ECIMOVIC, J P (1956) Three stage sampling with varying probabilities of selection, *JISAS*, 8, 14-44 M- 9
- ECKLER, A R and STAUDT, E P (1943) Marketing and sampling uses of population and housing data, *JASA*, 38, 87-92 S- 1
- ECKLER, A R (1945) The revised census series of current employment estimates, *JASA*, 40, 187-196 F- 1
- ECKLER, A R and PALTZKER, L (1951) Measuring the accuracy of enumerative surveys, *BISI*, 33 (4), 7-24 P-13
- ECKLER, A R (1953) Extent and character of errors in the 1950 census, *Amer Stat*, 7, (5), 15-19, 21 P-13
- ECKLER, A R (1955) Rotation sampling *AMS*, 26, 664-685 M-11
- ECKLER, A R and HURWITZ, W N (1958) Response variances and biases in censuses and surveys, *BISI*, 36, (2), 12-35 P-13
- EDGAR, J L (1938) Hoblyn, T N
- EDWARDS F (1953) Aspects of random sampling for a commercial survey, *Ind Stat*, 4, 9-26 E- 1
- EHRENBORG, A S C (1960) A study of some potential biases in the operation of a consumer panel *Appl Stat*, 9, 20-27 L-13
- EHRLICH, J S and RIESMAN, D (1961) Age and authority in interview, *POQ*, 25, 39-56 S-13
- EKMAN, G (1959a) A limit theorem in connection with stratified sampling, *Skand Alt*, 42, 208-223 M- 7
- EKMAN, G (1959b) An approximation useful in university stratification, *AMS*, 30, 219-229 M- 7
- EKMAN, G (1960) A limit theorem in connection with stratified sampling, Part II, *Skand Alt*, 43, 1-26 M- 7
- EL BADRY, M A and STEPHAN, F F (1955) On adjusting sample tabulations to census counts, *JASA*, 50, 738-762 S-13
- EL BADRY, M A (1956) A sampling procedure for mailed questionnaires, *JASA*, 51, 209-227 M-13
- ELKIN, J M (1953) Estimating the ratio between the proportions of two classes when one is a sub-class of the other, *JASA*, 48, 128-130 M-10
- EL SAYEH, M A (1961) *See* Husein, H M
- EMMETT, B P (1964) Reflections on the state of population sampling in the United Kingdom, *Appl Stat*, 13, 146-157 F- 2
- ENDRISS J (1961) *See* Thompson, W A (Jr)
- ERDŐS, P and RENYI, A (1959) On the central limit theorem for samples from a finite population, *Publication of the Mathematical Institute of the Hungarian Academy of Science*, 49-61 M- 1
- ERICSSON, W A (1965) Optimum stratified sampling using prior information, *JASA*, 60, 750-771. M- 7

BIBLIOGRAPHY

607

- EVANS, D. H. (1963a) : Multiplex sampling; *AMS*, 34, 1322-1346. M- 1
- EVANS, D. H. (1963b) : Applied multiplex sampling; *Technometrics*, 5, 341-359. M- 1
- EVANS, W. D. (1951a) : On stratification and optimum allocation; *JASA*, 46, 95-104. M- 7
- EVANS, W. D. (1951b) : On the variance of estimates of the standard deviation and variance; *JASA*, 46, 220-224. M- 3
- EVANS, W. D. (1958) : The control of non-sampling errors in social and economic surveys; *BISI*, 38, (2), 36-43. S-13
- FAN, C. T., MULLER, M. E. and REZUCHA, I. (1962) : Development of sampling plan by using sequential (item by item) selection techniques and digital computers; *JASA*, 57, 387-402. M-14
- FASTEAU, H. H., INGRAM, J. J. and MINTON, G. (1964) : Control of quality of coding in the 1960 censuses; *JASA*, 59, 120-132. P-13
- FATTU, N. A. and RAO, V. R. (1964) : On a bias in voting; *Amer. Stat.*, 18, (5), 19-20. O-13
- FEDERER, W. T. (1946) : See Houseman, E. E.
- FEDERER, W. T. (1953) : See Brunk, M. E.
- FELDT, A. (1959) : See Sharp, H.
- FELLEGI, I. P. (1959) : Some sequential techniques applied by the Dominion Bureau of Statistics; *Estadística*, 17, 532-543. M-14
- FELLEGI, I. P. (1963) : Sampling with varying probabilities without replacement—rotating and non-rotating samples; *JASA*, 58, 183-201. M-11
- FELLEGI, I. P. (1964) : Response variance and its estimation, *JASA*, 59, 1016-1041. M-13
- FERBER, R. (1946) : The disproportionate method of market sampling; *J. Business*, 67-75. O- 1
- FERBER, R. (1955a) : On the reliability of responses secured in sample surveys; *JASA*, 50, 788-810. S-13
- FERBER, R. (1955b) : Sales forecasting by sample surveys; *J. Marketing*, 20, 1-13. O- 1
- FERBER, R. (1956) : The effect of respondent ignorance of survey results; *JASA*, 51, 576-586. M-13
- FERBER, R. (1965) : The reliability of consumer surveys of financial holdings : Time deposits; *JASA*, 60, 148-163. E-13
- FERBER, R. (1966) : The reliability of consumer surveys of financial holdings : Demand deposits; *JASA*, 61, 91-103. E-13
- FIELLER, E. C. and HARTLEY, H. O. (1954) : Sampling with control variables; *Biometrika*, 41, 494-501. M- 1
- FINKNER, A. L., MORGAN, J. N. and MONROE, R. J. (1943) : Methods of estimating farm employment from sample data in North Carolina; *North Carolina Agricultural Experimental Station Technical Bulletin*, No 75, A- 1
- FINKNER, A. L. (1950) : Methods of sampling for estimating commercial peach production in North Carolina; *North Carolina Agricultural Experimental Station Technical Bulletin* No. 91. C- 1

- FINKNER, A L (1952) Adjustment for non response biases in a rural mailed survey, *AER*, 4, 77-82 C-13
- FINKNER, A L (1954) See Sen, A R
- FINKNER, A L (1958) See Fleischer, J
- FINNEY, D J (1942) See Yates, F
- FINNEY, D J (1946) Field sampling for the estimation of wire worm populations, *Biometrics*, 2, 1-7 W- 1
- FINNEY, D J (1947) Volume estimation of standing timber by sampling, *Forestry*, 21, 179-203 F- 1
- FINNEY, D J (1948) Random and systematic sampling in timber surveys, *Forestry*, 22, 64-99 F- 5
- FINNEY, D J (1949a) The efficiency of enumeration I Volume estimation of standing timber by sampling II Random and systematic sampling in timber surveys, *Forest Research Institute, Dehradun, Bulletin*, No 146 F- 1
- FINNEY, D J and PALCA, H (1949b) The elimination of bias due to edge effects in forest sampling, *Forestry* 23 31-47 F-13
- FINNEY, D J (1950) An example of periodic variation in forest sampling, *Forestry* 23 96-111 F- 1
- FINNEY, D J (1953) The estimation of error in the systematic sampling of forests, *JISAS*, 5, 6-16 F- 5
- FISHER, G (1962) A discriminant analysis of reporting errors in health interviews *Appl Stat* 11, 148-163 P-13
- FISHER, R A (1950) The Sub Commission on Statistical Sampling of the United Nations *BISI*, 32 (2), 207-209 M- 1
- FISHER, W D (1958) On grouping for maximum homogeneity, *JASA*, 53, 789-798 M- 7
- FISKE, M (1938) See Lazarsfeld, P F
- FITZPATRICK, T B (1961) See Hess, I
- FLEISCHER, J, HORVITZ, D G, AIRTH J M, and FINKNER, A L (1958) Measurement of errors associated with obtaining acreage estimates of cotton fields, *Biometrics*, 14, 401-407 C-13
- FLEMING, W E (1936) See Baker, E F
- FLORES, A M (1958) The Theory of duplicated samples and its use in Mexico, *BISI*, 36, (3), 120-126 M- 1
- FOG, D (1948) The geometrical method in the theory of sampling *Biometrika*, 35, 46-54 M- 1
- FOGH, I F (1943) Sampling methods in log scaling, *Forest Chronicle*, No 19, 127-138 F- 1
- FOLKS, J L and ANTLE, C E (1965) Optimum allocation of sampling units to strata when there are R responses of interest, *JASA*, 60, 225-233 M- 7
- FORD, R N (1947) See Clausen, J A
- FORD, R N and ZEISEL, H (1949) Bias in mail surveys cannot be controlled by one mailing, *POQ*, 13, 495-501 O-13
- FRANKEL, A (1940) See Kullback, S
- FRANKEL, L R (1939) See Stock, J S

BIBLIOGRAPHY

609

FRANKEL, L. R. and STOCK, J. S. (1942) : On the sample survey of unemployment; <i>JASA</i> , 37, 77-80.	S- 1
FRANKEL, L. E. (1950) : Sample to estimate tire inventories; <i>J. Marketing</i> , 14, 584-586.	E- 1
FRANZEN, R. and WILLIAMS, R. (1956) : A method for measuring error due to variance among interviewers; <i>POQ</i> , 20, 587-592.	O-13
FREEDMAN, R. (1964) : Sample surveys for family planning research in Taiwan; <i>POQ</i> , 28, 373-382.	P- 1
FREEMAN, G. H. and BOLAS, B. D. (1956) : A method for the rapid determination of the leaf areas in the field; <i>Report for the East Malling Experimental Research Station for 1955</i> , 104-107.	C- 1
FREEMAN, G. H. (1958) : A comparison of methods of measuring leaf areas in the field; <i>Report of the East Malling Experimental Research Station for 1957</i> , 83-86.	A- 1
FULLER, W. A. (1966) : Estimation employing post-strata; <i>JASA</i> , 61, 1172-1183.	M- 7
FUNG, A. F. (1961) : Interviewer differences among automobile purchasers; <i>Appl. Stat.</i> , 10, 93-97.	E-10
GAGE, R. P. (1943) : Contents of Tippet's "random sampling numbers"; <i>JASA</i> , 38, 223-227.	M- 1
GALES, K. and KENDALL, M. G. (1957) : An inquiry concerning interviewer variability; <i>JRSS</i> , (A), 120, 121-147.	S-13
GALLUP, G. (1938) : Government and the sampling referendum; <i>JASA</i> , 33, 131-142.	O- 1
GANGULI, M. (1941) : A note on nested sampling; <i>Sankhyā</i> , 5, 449-452.	M- 9
GANGULY, A. (1951) : See Lahiri, D. B.	
GARWOOD, F. (1962) : The sampling and use of traffic flow statistics; <i>Appl. Stat.</i> , 11, 1-15.	W- 1
GAUTSCHI, W. (1957) : Some remarks on systematic sampling; <i>AMS</i> , 28, 385-394.	M- 5
GAYLOR, D. W. (1956) : Equivalence of two estimates of product variance; <i>JASA</i> , 51, 451-453.	M- 1
GEARY, R. C. (1950) : Most efficient sample sizes for the two-stage sampling process in the case of limited universe; <i>BISI</i> , 32, (2), 228-239.	M- 1
GEOFFREY, L. (1941) : See Deming, W. E. (1941b).	
GEORGE, R. F. (1936) : A sample investigation of the 1931 Population Census with reference to earners and non-earners; <i>JRSS</i> , (A), 99, 147-161.	P-13
GEVORKIANTZ, S. R. (1934) : See Mudgett, B. D.	
GHOSH, A. (1946) : See Mahalanobis P. C. (1946d).	
GHOSH, A. (1953) : Accuracy of family budget data with reference to period of re-call; <i>BCSA</i> , 4, 16-23.	L-13
GHOSH, B. (1941) : On some methods of random sampling in a region having some known characteristics; <i>Science and Culture</i> , 7, 117.	M- 1
GHOSH, B. (1943) : On sampling in unknown fields; <i>Science and Culture</i> , 9, 129-130.	M- 1

- GHOSH, B (1947a) Crop estimation in India. A brief review of the sampling methods, *BCSA*, 1, 5-12 C- 1
- GHOSH, B (1947b) Bias due to change in stratification, *BCSA*, 1, 43-46 M- 7
- GHOSH, B (1947c) Double sampling with many auxiliary variates *BCSA*, 1, 91-93 M-11
- GHOSH, B (1949a) A multi stage stochastic model for natural fields, *BCSA*, 2 21-31 M- 9
- GHOSH, B (1949b) Interpenetrating (net works of) sample, *BCSA*, 2 108-119 M-13
- GHOSH, B (1954) A variance in areal sampling, *BCSA*, 5, 73-81 M- 6
- GHOSH, B (1956a) Optimum structure of rectangular sample units *BCSA*, 6, 176-180 C- 8
- GHOSH, B (1956b) A model for perimeter errors, *BCSA*, 6, 189-192 C-13
- GHOSH, B (1957a) Enumerational errors in surveys, *BCSA*, 7, 50-59 M-13
- GHOSH, B (1957b) A practical problem of random samples, *BCSA*, 7, 167-171 M- 1
- GHOSH, J K (1963) A game theory approach to the problem of optimum allocation in stratified sampling multiple characters, *BCSA*, 12, 4-12 M- 7
- GHOSH, M N (1947) Survey of public opinion, *BCSA* 1, 13-18 O- 1
- GHOSH, M N (1961) See Shaligram, G C
- GHOSH, S P (1958) A note on stratified random sampling with multiple characters *BCSA* 8, 81-90 M- 7
- GHOSH S P (1963a) Post cluster sampling *AMS*, 34, 587-597 M-14
- GHOSH S P (1963b) Optimum stratification with two characters, *AMS*, 34 866-872 M- 7
- GHOSH, S P (1963c) Estimating the mean by two stage sampling with replacement *BCSA* 12, 97-103 M- 9
- GHOSH S P (1965) Optimum allocation in stratified sampling with replacement, *Metrika*, 9, 212-221 M- 7
- GHOSH S P (1966) Polychotomy sampling, *AMS* 37, 657-665 L- 1
- GHOSH T (1940) Sampling in family budget enquiries, *Sankhya*, 4, 501-504
- GILFORD, D M and MARKS, C L (1956) Use of a sample survey for estimating aggregate quarterly financial statement for a population of corporations *Improving the Quality of Statistical Surveys*, 15-30 A Memorial to Samuel Weiss, American Statistical Association, Washington, D C E- 1
- GLASSER, G J (1959) See Deming, W E
- GLASSER, G J (1961) An unbiased estimator for powers of the arithmetic mean, *JRSS, (B)*, 23 154-159 M- 1
- GLASSER, G J (1962a) On estimators of variances and covariances, *Bio metrika*, 49, 259-262 M- 1
- GLASSER G J (1962b) Estimators for the product of arithmetic means, *JRSS, (B)*, 24, 180-184 M- 1
- GLASSER, G J (1962c) On the complete coverage of large units in a statistical survey, *RISI*, 30, 28-32 M- 3

- GLASSER, G. J. (1963) : Random numbers, sample selection and occupancy problems; *JRSS*, (A), 126, 115-119. M- 1
- GODAMBE, V. P. (1951) : On two-stage sampling; *JRSS*, (B), 13, 216-218. M- 9
- GODAMBE, V. P. (1955) : A unified theory of sampling from finite populations; *JRSS*, (B), 17, 269-278. M- 1
- GODAMBE, V. P. (1960) : An admissible estimate for any sampling design; *Sankhyā*, 22, 285-288. M- 1
- GODAMBE, V. P. (1965a) : A review of the contributions towards a unified theory of sampling from finite populations; *RISI*, 33, (2), 242-258. M- 1
- GODAMBE, V. P. and JOSHI, V. M. (1965b) : Admissibility and Bayes estimation in sampling finite populations I; *AMS*, 36, 1707-1722. M- 1
- GODAMBE, V. P. (1966a) : A new approach to sampling from finite populations I; *JRSS*, (B), 28, 310-319. M- 1
- GODAMBE, V. P. (1966b) : A new approach to sampling from finite populations II; *JRSS*, (B), 28, 320-328. M- 1
- GOLDBERG, S. A. (1958) : Non-sampling error in household surveys : A general review of some Canadian work; *BISI*, 36, (2), 44-59. S-13
- GOLHAR, M. B. (1961) : See Shaligram, G. C.
- GOODMAN, L. A. (1949) : On the estimation of the number of classes in a population; *AMS*, 20, 572-579. M- 1
- GOODMAN, L. A. (1952) : On the analysis of samples from k lists; *AMS*, 23, 632-634. M- 1
- GOODMAN, L. A. (1953) : A simple method for improving some estimators; *AMS*, 24, 114-117. M- 1
- GOODMAN, L. A. and HARTLEY, H. O. (1958) : The precision of unbiased ratio-type estimators; *JASA*, 53, 491-508. M-10
- GOODMAN, L. A. (1960) : On the exact variance of products; *JASA*, 55, 708-713. M- 1
- GOODMAN, L. A. (1961) : Snow-ball sampling; *AMS*, 32, 148-170. M-14
- GOODMAN, L. A. (1962) : The variance of the product of k random variables; *JASA*, 57, 54-60. M- 1
- GOODMAN, R. (1947) : Sampling for the 1947 survey of consumer finances; *JASA*, 42, 439-448. L- 1
- GOODMAN, R. and KRISH, L. (1950) : Controlled selection—A technique in probability sampling; *JASA*, 45, 350-372. M- 7
- GOODMAN, R. and MACCOBY, E. E. (1948) : Sampling methods and sampling errors in surveys of the consumer finances; *IJOAR*, 2, 349-360. L- 1
- GOODSELL, W. D., JESSEN, R. J. and WILCOX, W. W. (1940) : Procedures which increase the usefulness of farm management research; *JFE*, 22, 753-761. A- 1
- GOSWAMI, J. N. and SUKHATME, B. V. (1965) : Ratio method of estimation in multi-phase sampling with several auxiliary variables; *JISAS*, 17, 83-103. M-10
- GRAHAM, J. E. (1964) : See Rao, J. N. K.
- GRAY, P. G. and CORLETT, T. (1950) : Sampling for the social survey; *JRSS*, (A), 113, 150-206. S- 1

- GRAY, P G (1955) The memory factor in social surveys, *JASA*, 50, 344-363 S-13
- GRAY, P G (1956) Examples of interviewer variability taken from two sample surveys, *Appl Stat*, 5, 73-85 S-13
- GRAY, P G (1957) A sample survey with both a postal and an interview stage, *Appl Stat*, 6, 139-153 A- 1
- GRAYBILL, F A (1958) Determining sample size for a specified width confidence interval, *AMS*, 29, 282-287 M- 4
- GREENBERG, A and LISSANCE, D (1955) The accuracy of journalistic poll, *POQ* 19, 45-52 O-13
- GREFWOOD, J A and SANDOMIRE, M M (1950) Sample size required for estimating the standard deviation as a percent of its true value, *JASA*, 45, 257-260 M- 4
- GREVILLE, T N E (1949) Opinion polls and sample surveys, *Estadística*, 7, 92-93 O- 1
- GRIFFITH, A L (1945, 1946) The efficiency of enumerations Forest Research Institute, Dohradun, *Indian Forest leaflets*, 83 to 91, 93 and 96. F- 1
- GRIFFITHS, W (1965) See Yankey, D
- GROSENBAUM, L R (1952) Plotless timber estimates—new, fast, easy, *J Forestry*, 50, 32-37 F- 1
- GRUBBS F E and WEAVER, C L (1947) The best unbiased estimate of population standard deviation based on group ranges *JASA*, 42, 224-241 M- 1
- GRUBBS, F E (1948) On estimating precision of measuring instruments and product variability, *JASA*, 43, 243-261, 564 M-13
- GRUNDY, P M (1951) The expected frequencies in a sample of an animal population in which the abundances of species are log normally distributed I, *Biometrika*, 38, 427-434 W- 1
- GRUNDY, P M (1953) See Yates, F
- GRUNDY, P M (1954) A method of sampling with probability exactly proportional to size, *JRSS. (B)* 16, 236-238 M- 6
- GURNEY, M (1943) See Deming, W E (1943a)
- GURNEY, M (1946) See Hansen, M H (1946a)
- GURNEY, M (1951) See Dalenius, T
- GUTERMANN, H E (1958) See Rosander, A C
- HAGOOD M J and BERTERT, E H (1946) Component indexes as a basis for stratification in sampling, *JASA*, 40, 330-341 M- 7
- HAJFK, J (1958) Some contributions to the theory of probability sampling, *B ISI*, 36, (3), 127-133 M- 1
- HAJFK, J (1960a) On the theory of ratio estimates, *B ISI*, 37, (2), 219-226 M-10
- HAJFK, J (1960b) Limiting distributions in simple random sampling from a finite population, *Publication of the Mathematical Institute of the Hungarian Academy of Science*, 5, 361-374 M- 3
- HAJFK, J (1961) See Dalenius, T
- HAJFK, J (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population, *AMS*, 35, 1491-1523, M- 6

- HALDAR, A. (1961) : A note on the amount of rejection in Lahiri's method of pps sampling; *Sankhyā*, 23, (B), 329-330. M- 6
- HALL, O. (1949) : The use of sampling procedures and role theory in sociological research; *Canadian Journal of Economics and Political Science*, 15, 1-13. S- 1
- HALPERIN, M. (1961) : Almost linearly-optimum combination of unbiased estimates; *JASA*, 56, 36-43. M- 1
- HAMILTON, E. L. (1946) : The problem of sampling rainfall in mountainous areas; *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, 469-476. M- 1
- HAMMERSLEY, J. M. (1953) : Capture-recapture analysis, *Biometrika*, 40, 265-278. W-14
- HANER, C. F. (1951) : See Meier, N. C.
- HANNA, H. S. (1934) : Adequacy of sample in budgetary studies; *JASA*, (Supplement), 29, 131-134. L- 1
- HANNAN, E. J. (1962) : Systematic sampling; *Biometrika*, 49, 281-283. M- 5
- HANSEN, M. H. (1940) : See Stephan, F. F. (1940a).
- HANSEN, M. H. and HURWITZ, W. N. (1942) : Relative efficiencies of various sampling units in population inquiries; *JASA*, 37, 89-94. P- 8
- HANSEN, M. H. and DEMING, W. E. (1943a) : On some census aids to sampling; *JASA*, 38, 353-357. M- 1
- HANSEN, M. H. and HURWITZ, W. N. (1943b) : On the theory of sampling from finite populations; *AMS*, 14, 333-362. M- 6
- HANSEN, M. H. and HURWITZ, W. N. (1944a) : A new sample of the population; *Estadística*, 2, 483-497. P- 1
- HANSEN, M. H. (1944b) : See Hauser, P. M.
- HANSEN, M. H. and HAUSER, P. M. (1945) : Area sampling—some principles of sampling design; *POQ*, 9, 183-193.
- HANSEN, M. H., HURWITZ, W. N. and GURNEY, M. (1946a) : Problems and methods of the sample survey of business; *JASA*, 41, 173-189, 529. F- 1
- HANSEN, M. H. and HURWITZ, W. N. (1946b) : The problem of non-response in sample surveys; *JASA*, 41, 517-529. M-13
- HANSEN, M. H. (1947) : Sampling of human populations; *Proceedings of International Statistical Conference*, 3, 113-128. P- 1
- HANSEN, M. H. and HURWITZ, W. N. (1949a) : On the determination of optimum probabilities in sampling; *AMS*, 20, 426-432. M- 6
- HANSEN, M. H. and HURWITZ, W. N. (1949b) : Dependable samples for market surveys; *J. Marketing*, 14, 363-372. O- 1
- HANSEN, M. H. and DEMING, W. E. (1950) : On an important limitation to the use of data from samples; *BISI*, 32, (3), 214-219. M- 1
- HANSEN, M. H., HURWITZ, W. N., MARKS, E. S. and MAULDIN, W. P. (1951a) : Response errors in surveys; *JASA*, 46, 147-190. M-13
- HANSEN, M. H. and HURWITZ, W. N. (1951b) : Modern methods in the sampling of human populations; *AJPH*, 41, 647-653. M-1

- HANSEN, M H, HURWITZ, W N and PRITZKER, L (1953). The accuracy of census results, *ASR*, 18, 418-423 P-13
- HANSEN, M H, HURWITZ, W N, NISSELSON, H and STEINBERG, J (1955) The redesign of the current population survey, *JASA*, 50, 701-719 S- 1
- HANSEN, M H and STEINBERG, J (1956) Control of errors in surveys, *Biometrika*, 42, 462-474 M-13
- HANSEN, M H (1957) Effect of the new design for the current population survey, *Estadística*, 15, 418-421 S- 1
- HANSEN, M H and MARKS, E S (1958) Influence of the interviewer on the accuracy of survey results, *JASA*, 53, 635-655 S-13
- HANSEN, M H, HURWITZ, W N and BERSHAD, M A (1961) Measurement of error in census and survey, *BISI*, 38, (2), 359-374 M-13
- HANSEN, M H, HURWITZ, W N and JABINE, T B (1963a) The use of imperfect list for probability sampling at the U S Bureau of the Census, *BISI*, 40, (1), 497-517 M- 1
- HANSEN, M H, HURWITZ, W N and PRITZKER, L (1963b) The estimation and interpretation of gross differences and the simple response variance, *Contribution to Statistics*, 111-136, Presented to Professor P C Mahalanobis on the occasion of his 70th Birthday, Pergamon Press, London and Statistical Publishing Society, Calcutta M-13
- HANUMANTHA RAO, T V (1962a) An existence theorem in sampling theory, *Sankhya*, 24, (A), 327-330 M- 1
- HANURAV, T V (1962b) Some sampling schemes in probability sampling, *Sankhya*, 24, (A), 421-428 M- 4
- HANURAV, T V (1962c) On Horvitz and Thompson estimator, *Sankhya*, 24, (A), 429-436 M- 6
- HANURAV, T V (1966) Some aspects of unified sampling theory, *Sankhya*, 28, (A), 175-203 M-1
- HARLAN, C L (1939) See Becker, J A
- HAERIS, F F (1954) The use of sampling methods for ascertaining total morbidity in the Canadian sickness survey, 1950-51, *World Health Organisation Bulletin*, 11, 25-50 S- 1
- HART H (1926) Reliability of a percentage, *JASA*, 21, 40-46 M- 3
- HARTLEY, H O and Ross, A (1954a) Unbiased ratio estimators, *Nature*, 174, 270-271 M-10
- HARTLEY, H O (1954b) See Fieller, E C
- HARTLEY, H O (1958) See Goodman, L A
- HARTLEY, H O (1960) See Bryant, E C
- HARTLEY, H O and RAO, J N K (1962a) Sampling with unequal probabilities and without replacement, *AMS*, 33, 350-374 M- 6
- HARTLEY, H O (1962b). See Rao, J N K (1962b)
- HARTLEY, H O (1966) Systematic sampling with unequal probability and without replacement, *JASA*, 61, 739-748 M- 6
- HASEL, A A (1938) Sampling error in timber surveys, *JAB*, 57, 713-736 F- 1

- HASEL, A. A. (1941) : Estimation of vegetation type areas by linear measurement; *J. Forestry*, 39, 34-40. F- 1
- HASEL, A. A. (1942) : Estimation of volume in timber stands by strip sampling; *AMS*, 13, 179-206. F- 1
- HATHEWAY, W. H. and WILLIAMS, E. J. (1958) : Efficiency estimation of the relationship between plot size and the variability of crop yields; *Biometrics*, 14, 207-222. C- 8
- HAUSER, P. M. (1941) : The use of sampling in the census; *JASA*, 36, 369-375. P- 1
- HAUSER, P. M. (1942) : Proposed annual sample census of population; *JASA*, 37, 81-88. P- 1
- HAUSER, P. M. and HANSEN, M. H. (1944) : On sampling in market surveys; *J. Marketing*, 9, 26-31. P- 1
- HAUSER, P. M. (1945) : See Hansen, M. H. O- 1
- HAUSER, P. M. (1950) : Some aspects of methodological research in the 1950 census; *POQ*, 14, 5-13. P- 1
- HAUSER, P. M. (1954) : The use of sampling for vital registration and statistics; *World Health Organisation Bulletin*, 11, 5-24. P- 1
- HAYASHI, C. (1950a) : Sampling design in literary survey; *AISM*, 2, 49-59. S-1
- HAYASHI, C. (1950b) : Sampling design in the social survey of language at the city of Shirakawa; *AISM*, 2, 69-76. S- 1
- HAYASHI, C. (1950c) : See Maruyama, H. S- 1
- HAYASHI, C. (1957) : Note on sampling from a sociometric pattern; *AISM*, 9, 49-52. S- 1
- HAYASHI, C. (1961) : Sample survey and theory of quantification; *BISI*, 38, (4), 505-514. M- 1
- HAYES, S. P. (JR.) (1948) : Commercial surveys as an aid in the determination of public policy : A case study; *J. Marketing*, 12, 475-482. O- 1.
- HEALY, W. C. (1960) : See Birnbaum, A. M- 1
- HEGE, V. S. (1965) : Sampling designs which admit uniformly minimum variance unbiased estimates; *BCSA*, 14, 160-162. M- 1
- HEMMERLE, W. J. (1963) : See Clyde, R. W. M- 1
- HENDRICKS, W. A. (1944) : The relative efficiencies of groups of farms as sampling units; *JASA*, 39, 366-376. A- 8
- HENDRIKS, W. A. (1949) : Adjustment for bias by non-response in mailed surveys; *AER*, 1, 52-. O-13
- HENDRIKS, W. A. (1951) : Variance components as a tool for the analysis of sample data; *Biometrics*, 7, 97-101. M- 1
- HENDRIKS, W. A. (1956) : Non-sampling errors in agricultural surveys; *Improving the Quality of Statistical Surveys*; 31-39, A Memorial to Samuel Weiss, American Statistical Association, Washington, D.C. C-13
- HENDRIKS, W. A. (1957) : Research on sample survey procedures and crop estimating methods; *Estadistica*, 15, 346-355. C- 1
- HENDRIKS, W. A. (1964) : Estimation of the probability that an observation will fall into a specified class; *JASA*, 59, 225-232. M- 3

- HENEMAN, H G (JR) (1949) *See Patterson, D G*
- HESS I (1958) *See Kish, L*
- HESS, I (1959a) *See Kish, L (1959a)*
- HESS, I (1959b) *See Kish, L (1959b)*
- HESS I, RIEDEL, D C and FITZPATRICK, T B (1961) *Probability Sampling of Hospitals and Patients* University of Michigan, Ann Arbor S- 1
- HESS, I, SETHI, V K and BALAKRISHNA, T R (1966) *Stratification A practical investigation*, *JASA*, 61, 74-90 M- 7
- HILGARD, E R and PAYNE, S L (1944) Those not at home riddle for pollsters *POQ*, 8 254-261 O-13
- HILL, T P, KLEIN, L R and STRAW, K H (1955) The savings survey, response rates and reliability of data *Bulletin of Oxford University Institute of Statistics*, 17, 89-126 L-13
- HILTON, J (1924) Enquiry by sample an experiment and its results, *JRSS*, 87, 544-561 S- 1
- HILTON, J (1928) Some further enquiries by sample, *JRSS*, 91, 519-540 S- 1
- HITT, H L (1940 41) A sampling technique for studying population changes in rural areas, *Social Forces*, 20, 208 P- 1
- HOBLYN, T N and EDGAR, J L (1938) Experiments in sampling techniques II, size and colour of Allington Pippin, 1936 Crop, *Report of East Malling Experimental Research Station for 1937*, 168-172 C- 1
- HOCHSTIM, J R (1947) *See Stock, J S*
- HOCHSTIM, J R and SMITH, D M K (1948) Area sampling or quota control ? three sampling experiments, *POQ*, 12, 73-80 O-14
- HOCHSTIM, J R (1951) *See Stock J S*
- HODGES J L (JR) (1957) *See Dalenius, T (1957b)*
- HODGES, J L (JR) (1959) *See Dalenius, T*
- HOEL P G (1943) The accuracy of sampling methods in ecology, *AMS*, 14 289-300 W- 1
- HOLLAND, D A (1957) *See Pearce, S C*
- HOLLAND, D A (1958) *See Jolly G M*
- HOLLAND, D A (1965) Sampling errors in an orchard survey involving unequal numbers of orchards of different type, *Biometrics*, 21, 55-62 C- 1
- HOLMES, I (1939) Results of four methods of sampling individual farms, *JFE*, 21, 365-374 C- 1
- HOLMES, I (1943) Some sampling uses of data from the Census of Agriculture, *JASA*, 38, 78-87 A- 1
- HORVITZ, D G and THOMPSON, D J (1952) A generalization of sampling without replacement from a finite universe, *JASA*, 47, 663-685 M- 6
- HORVITZ, D G (1958) *See Fleischer, A J*
- HOUSEMAN, E E (1944) *See Jessen, R J*
- HOUSEMAN, E E, WEBER, C R and FEDERER, W T (1946) Preharvest sampling of soyabean for yield and quality, *Iowa Agricultural Experimental Research Station Bulletin No 341*. C- 1

- HOUSEMAN, E. E. (1947) : The sample design for a national farm survey, by the Bureau of Agricultural Economics; *JAE*, 29, 241-245. A- 1
- HOUSEMAN, E. E. (1950) : Sampling methods in marketing research; *AER*, 2, 73-81. O- 1
- HOUSEMAN, E. E. (1957) : Sample design for the survey of farm operators' 1955 expenditure; *Estatistica*, 15, 591-600. A- 1
- HUBBACk, J. A. (1927) : Sampling for rice yield in Bihar and Orissa; *Indian Agricultural Institute, Pusa, Bulletin*, No. 166, reprinted in *Su Dutt*, 7, (1946), 281-294. C- 1
- HUDDLESTON, H. F. (1950) : Methods used in a survey of orchards; *ALR*, 2, 126-130. C- 1
- HUDSON, H. G. (1939) : Population studies with wheat, I, sampling; *JAS*, 29, 76-110. C- 1
- HURLEY, R. (1961) : See Jabine, T. B.
- HURWITZ, W. N. (1942) : See Hansen, M. H.
- HURWITZ, W. N. (1943a) : See Hansen, M. H. (1943b).
- HURWITZ, W. N. (1943b) : See Tepping, B. J.
- HURWITZ, W. N. (1944) : See Hansen, M. H. (1944a).
- HURWITZ, W. N. (1946a) : See Hansen, M. H. (1946a).
- HURWITZ, W. N. (1946b) : See Hansen, M. H. (1946b).
- HURWITZ, W. N. (1949a) : See Hansen, M. H. (1949a).
- HURWITZ, W. N. (1949b) : See Hansen, M. H. (1949b).
- HURWITZ, W. N. (1951a) : See Hansen, M. H. (1951a).
- HURWITZ, W. N. (1951b) : See Hansen, M. H. (1951b).
- HURWITZ, W. N. (1953) : See Hansen, M. H.
- HURWITZ, W. N. (1955) : See Hansen, M. H.
- HURWITZ, W. N. (1961a) : See Hansen, M. H.
- HURWITZ, W. N. (1961b) : See Jabine, T. B.
- HURWITZ, W. N. (1963a) : See Hansen, M. H. (1963a).
- HURWITZ, W. N. (1963b) : See Hansen, M. H. (1963b).
- HUSEIN, H. M. and EL. SAYEH, M. A. (1961) : The 1958-1959 family budget sample survey in Egypt (UAR); *BISI*, 38, (2), 191-205. L- 1
- HYUYETT, M. J. (1958) : See Sobel, M.
- HYMAN, H. (1949a) : See Manheimer, F. D.
- HYMAN, H. (1949b) : See Mosteller, G.
- IMMER, F. R. (1932) : A study in sampling technique with sugar-beets; *JAS*, 44, 633-647. C- 1
- INDIAN COUNCIL OF AGRICULTURAL RESEARCH (1951) : Sample Surveys for the estimation of yield of food crops, 1944-49; *Bulletin* No. 72, Government of India, New Delhi. C- 1
- INDIAN STATISTICAL INSTITUTE (1952) : *National Sample Survey General Report No. 1 on the First Round (October 1950-March 1951)*; Government of India, New Delhi. S- 1

- INGRAM, J J (1964) *See* Fastean, H H
- IRWIN, J O, COCHRAN, W G and WISHART, J (1938) Crop estimation and its relation to agricultural meteorology, *JRSS, (Supplement)* 5, 1-45 C- 1
- IRWIN, J O and KENDALL, M G (1943) Sampling moments of moments for a finite population *Annals of Eugenics* 12, 138-142 M- 1
- IRWIN, J O (1951) Contribution to a discussion on crop prediction, *BISI*, 33, (2), 289-294 C- 1
- ISIDA, M D (1950) *See* Maruyama, E
- ISIDA, M D (1964) 10,000 spots forest survey, *AISM*, 16, 255-276 F- 1
- IYER, P V K and SINGH, D (1951) Problem of distance in sampling, *BISI*, 33, (2), 113-118 M- 1
- JABINE, T B and HURLEY, R and HURWITZ, W N (1961) Sample design and estimation procedure for the 1960 sample survey of agriculture in the U S A, *RISI*, 29, (3) 1-12 A- 1
- JABINE, T B (1963a) *See* Hansen, M H (1963a)
- JAGANNATHAN, R (1963a) The programming approach in multiple character studies *Econometrica*, 33, 236-237 M- 7
- JAGANNATHAN, R (1965b) A method for solving a non linear programming problem in sample surveys, *Econometrica* 33, 841-846 M- 1
- JAMES, S P (1956) Some sampling problems in connection with accounting records, *Appl Stat* 5 86-105 E- 1
- JARRATT P (1964) *See* Chambers M L
- JEBE, E H (1940) *See* King A J (1940a)
- JEBE E H (1952) Estimation for sub sampling designs employing the county as a primary sampling unit, *JASA*, 47, 49-70 A- 9
- JENSEN A (1926) Report on the representative method in statistics, *BISI*, 22 359-437 M- 1
- JENSEN A (1928) Purposive selection *JRSS, (A)*, 91, 541-547 M- 1
- JENSEN A (1952) A short remark on the theory of random sampling and the theory of variance, *Skand Akt*, 35, 195-200 M- 1
- JENSEN E L (1959) Optimum stratification of the logarithmic normal distribution A comment, *Skand Akt*, 42, 144-147 M- 7
- JESSEN, R J (1939) An experiment in the design of agricultural surveys, *JFE*, 21, 856-863 C- 1
- JESSEN, R J (1940) *See* Goodsell W D
- JESSEN, R J (1942) Statistical investigation of a sample survey for obtaining farm facts *Iowa Agricultural Experimental Station Research Bulletin*, No 304 A- 1
- JESSEN, R J (1943) *See* Strand, N V
- JESSEN, R J and HOUSEMAN, E E (1944) Statistical investigation of farm sample surveys taken in Iowa, Florida and California, *Iowa Agricultural Experimental Station Research Bulletin*, No 329 A- 1
- JESSEN, R J (1945) *See* King, A J

- JESSEN, R. J. (1947a): The master sampling project report on agricultural economics; *JFE*, 29, 531-540.
A- 1
- JESSEN, R. J., BLYTHE, R. H., KIRKTHORPE, O. and DAVIES, W. E. (1947 b): On a population sample for Greece; *JAS*, 1, 42, 357-384.
P- 1
- JESSEN, R. J., KIRKTHORPE, O., DALY, J. F. and DAVIES, W. E. (1949): Observations on the 1946 election in Greece; *ASB*, 14, 11-16.
P- 1
- JESSEN, R. J. (1955): Determining the fruit count on a tree by random branch sampling; *Biometrics*, 11, 99-109.
C- 6
- JESSEN, R. J. (1960): See Bryant, E. C.
- JOHANSEN, K. (1958): A preliminary report of a sample survey of housing requirements in the Copenhagen metropolitan housing area; *BISI*, 36, (4), 69-77.
S- 1
- JOHNSON, D. G. (1951): See Rosender, A. C.
- JOHNSON, F. A. (1943): A statistical study of sampling methods for tree nursery inventories; *J. Forestry*, 41, 671-679.
P- 1
- JOHNSON, F. A. (1949): Statistical aspects of timber volume sampling in Pacific North-West; *J. Forestry*, 47, 292-295.
P- 1
- JOHNSON, F. A. (1950): Estimating forest areas and volumes for large tracts; *J. Forestry*, 48, 340-342.
P- 1
- JOHNSON, N. L. (1957): Optimal sampling for quota fulfilment; *Biometrika*, 44, 518-523.
M- 7
- JOHNSON, P. O. and RAO, M. S. (1959): *Modern Sampling Methods*; University of Minnesota Press, Minneapolis.
M- 9
- JOLLY, G. M. and HOLLAND, D. A. (1958): Sampling methods for the measurement of extension growth of apple trees; *Report of East Malling Research Station for 1957*, 87-90.
C- 1
- JOLLY, G. M. (1963): Estimates of population parameters from multiple recapture data with both death and dilution—deterministic model; *Biometrika*, 50, 113-128.
W-14
- JONES, A. E. (1948): Systematic sampling of continuous parameter populations; *Biometrika*, 35, 283-290.
M- 7
- JONES, E. W. (1937): Practical field methods of sampling soil for micro-organisms; *JAR*, 54, 123-134.
W- 1
- JONES, H. L. (1955): The application of sampling procedures to business operations; *JASA*, 50, 763-776.
E- 1
- JONES, H. L. (1956): Investigating the properties of a sample mean by employing random sub-sample means; *JAS*, 1, 51, 54-83.
M- 1
- JONES, H. L. (1959): How many of a group of random numbers will be usable in selecting a particular sample; *JASA*, 54, 102-122.
M- 1
- JOSHI, V. M. (1965a): Admissibility and Bayes estimation in samples finite populations II; *AMS*, 36, 1723-1729.
M- 1
- JOSHI, V. M. (1965b): Admissibility and Bayes' estimation in samples finite populations III; *AMS*, 36, 1730-1742.
M- 1
- JOSHI, V. M. (1965c): See Godambe, V. P.

- JOSHI, V M (1966) Admissibility and Bayes estimation in sampling finite populations IV, *AMS*, 37, 1658-1670 M- 1
- JOWETT, H H (1952) The accuracy of systematic sampling from conveyor belts *Appl Stat*, 1, 50-59 M- 5
- JUNGE, C D and CHAPMAN D G (1956) Estimation of the size of a stratified animal population, *AMS* 27, 375-389 W- 7
- JUSTESSEN S H (1932) Influence of size and shape of plots on the precision of field experiments with potatoes, *JAS*, 22, 366-372 C- 8
- KALAMKAR R J (1932) A study in sampling technique with wheat *JAS*, 22, 783-796 C- 1
- KALAMKAR, R J (1940a) A note on crop cutting experiments on cotton, *Sankhya*, 4, 589 C- 1
- KALAMKAR R J and DHANALAL (1940b) Sampling studies in cotton varietal trial, *Sankhya*, 4, 567-576 C- 1
- KALAMKAR R J (1941) A note on crop cutting experiments on cotton, *Sankhya* 5, 345-348 C- 1
- KALAMKAR R J (1944a) See Panse, V G (1944a)
- KALAMKAR R J (1944b) See Panse, V G (1944b)
- KALAMKAR, R J (1945) See Panse, V G
- KALYANASUNDARAM, G (1963) See Roy, J
- KAMEDA I (1930) Application of the method of sampling to the first Japanese population census *BISI*, 25 (2), 121-132 P- 1
- KATONA G (1950) See Lansing J B
- KATZ, D (1941) The public opinion poll and the 1940 election *POQ*, 5, 52-78 O- 1
- KATZ D (1942) Do interviewers bias poll results *POQ*, 6, 248-268 O- 13
- KATZ D (1944) The polls and the 1944 election *POQ*, 8, 468-482 O- 1
- KATZ, L (1953) Confidence interval for the number showing a certain characteristic in the population when sampling is without replacement, *JASA* 48, 256-261 M- 3
- KELLY B W (1958) Objective methods for forecasting Florida citrus population *Estadistica*, 16, 56-64 C- 1
- KEMPTHORNE, O (1947) See Jessen R J (1947b)
- KEMPTHORNE O (1949) See Jessen, R J
- KEMSLEY, W F F (1950) Designing a budget survey, *Appl Stat*, 8, 114-123 L- 1
- KEMSLEY, W F F (1960a) Interviewer variability and a budget survey, *Appl Stat*, 9, 122-128 L- 13
- KEMSLEY, W F F and NICHOLSEN, J L (1960b) Some experiments in methods of conducting family expenditure surveys *JRSS*, (A) 123, 307-328 L- 13
- KEMSLEY, W F F (1961) The household expenditure enquiry of the Ministry of labour—variability in the 1953-54 enquiry, *Appl Stat*, 10, 117-135 L- 13

- KEMSLY, W. F. F. (1962) : Some technical aspects of a pilot survey into professional earnings; *Appl. Stat.*, 11, 93-105.
L-13
- KEMSLY, W. F. F. (1965) : Interviewer variability in expenditure surveys; *JRSS, (A)*, 128, 118-142.
L-13
- KEMSLY, W. F. F. (1966) : Sampling errors in family expenditure survey; *Appl. Stat.*, 15, 1-14.
L-1
- KENDALL, M. G. and SMITH, B. B. (1938) : Randomness and random sampling numbers; *JRSS, (A)*, 101, 147-166.
M- 1
- KENDALL, M. G. (1939) : Second paper on random sampling numbers; *JRSS, Supplement*, 6, 51-61.
M- 1
- KENDALL, M. G. (1943) : See Irwin, J. O.
M- 1
- KENDALL, M. G. (1952) : Moment-statistics in sampling from a finite population; *Biometrika*, 39, 14-16.
M- 3
- KENDALL, M. G. (1957) : See Gales, K.
M- 3
- KEYFITZ, N. (1945) : The sampling approach to economic data; *Canadian Journal of Economics and Political Science*, 11, 167-177.
S- 1
- KEYFITZ, N. and ROBINSON, H. L. (1949) : The Canadian sample for labour force and other population data; *Population Studies*, 2, 427-443.
P- 1
- KEYFITZ, N. (1951) : Sampling with probability proportional to size: adjustment for changes in probabilities; *JASA*, 46, 105-109.
M- 6
- KEYFITZ, N. (1957a) : Calculation of variances in a monthly population survey; *BISI*, 35, (2), 181-186.
M- 1
- KEYFITZ, N. (1957b) : Estimates of sampling variance when two units are selected from each stratum; *JASA*, 52, 503-510.
M- 1
- KEYFITZ, N. (1960) : The design of surveys to provide experimental contrasts; *BISI*, 37, (2), 227-230.
M- 1
- KHAMIS, H. S. (1958a) : A report on a pilot infant mortality survey of rural Lebanon; *BISI*, 36, (2), 60-69.
S- 1
- KHAMIS, H. S. (1958b) : See Des Raj (1958b).
S- 1
- KHAN, S. and KOKAN, A. R. (1965) : A note on the stabilities of estimates of the sampling variances of the ordinary mean estimate and the regression estimate; *BCSA*, 14, 171-174.
M- 11
- KINDAHL, J. K. (1962) : Estimation of means and totals from finite population of unknown size; *JASA*, 57, 61-91.
M- 1
- KING, A. J. and JEBE, E. H. (1940a) : An experiment in the preliminary sampling of wheat fields; *Iowa Agricultural Experimental Station Research Bulletin No. 273*.
C- 1
- KING, A. J. and SIMPSON, G. D. (1940b) : New developments in agricultural sampling; *JAE*, 22, 311-349.
C-12
- KING, A. J. and McCARTHY, D. E. (1941) : Application of sampling to agricultural statistics with emphasis on stratified sampling methods; *J. Market. Res.*, 5, 462-474.
C- 7
- KING, A. J. (1912a) : See Snedecor, G. W.

SAMPLING THEORY AND METHODS

- ING, A J, McCARTHY, D E and McPEAK M (1942b) An objective method of sampling wheat fields to estimate production and quality of wheat, *United States Department of Agriculture Technical Bulletin No 814* C- 1
- ING, A J and JESSEN, R J (1945) The master sample of agriculture, *JASA*, 40 38-56 A- 1
- ING, A J (1952) See Robson D S
- ISER, C V (1934) Pitfalls in sampling for population study, *JASA*, 29, 250-256 P- 1
- ISH, L (1949) A procedure for objective respondent selection within the household, *JASA*, 44, 380-387 M- 9
- ISH, L (1950a) See Goodman, R
- ISH, L (1950b) See Lansing J B
- ISH, L (1952) A two-stage sample of a city, *ASR*, 17, 761-769 M- 9
- ISH, L (1953) Selection of the sample, *Research Methods in the Behavioural Sciences*, edited by Festinger and Katz 175-240 M- 1
- ISH, L and LANSING, J B (1954) Response errors in estimating value of homes *JASA*, 49 520-538 S-13
- ISH, L (1957) Confidence intervals for clustered samples *ASR* 22, 154-165 M- 8
- ISH, L and HESS I (1958) On non coverage of sample dwellings *JASA*, 53 509-524 M-13
- ISH, L and HESS, I (1959a) On variances of ratios and their differences in multi stage samples *JASA*, 54 416-446 M-10
- ISH, L (1959b) A "replacement" procedure for reducing the bias of non response *Amer Stat*, 13, (4) 17-19 M-13
- ISH, L (1961) Efficient allocation of a multi purpose sample, *Econometrica*, 29, 363-385 M- 7
- ISH, L (1962a) Studies of interviewer variance for attitudinal variables, *JASA*, 57, 92-115 O-13
- ISH, L, NAMBOODIRI, N K and PILLAI, R K (1962b) The ratio bias in survey, *JASA*, 57, 863-876 M-10
- KISHEN, K (1951) See Sukhatme, P V (1951b)
- ITAGAWA, T (1955) Some contributions to the design of sample surveys, parts I, II & III, *Sankhya* 14, 317-362 M- 1
- ITAGAWA, T (1956) Some contributions to the design of sample surveys, parts IV, V & VI, *Sankhya*, 17, 1-36 M- 1
- ITAGAWA, T (1957) Successive process of statistical inferences and applications to the design of sample surveys *BISI*, 35, (2) 151-162 M- 1
- KLEIN, L R (1951a) Estimating patterns of savings behaviour from sample survey data, *Econometrica*, 19, 438-454 L- 1
- KLEIN, L R and MORGAN, J N (1951b) Results of alternative statistical treatments of sample survey data, *JASA*, 46 442-460 M- 1
- KLEIN, L R (1955) See Hill T P
- KLEIN, L R and VANDOME, P (1957) Sampling errors in savings surveys, *Bulletin of Oxford University Institute of Statistics*, 19, 85-95 L- 1

- KLIPPLE, C. E. (1944): *See* Costello, D. P.
- KOKAN, A. R. (1963): A note on the stability of the estimates of sampling errors of the ordinary estimates and ratio estimate; *BCSI*, 12, 139-158. M-16
- KOKAN, A. R. (1963): *See* Khan, S.
- KOLLER, S. (1958): On the problems of replicated sampling in German governmental statistics; *BISI*, 36, (3), 137-143. M- 1
- KOLLER, S. (1960): The use of prior statistical information in problems of estimation; *BISI*, 37, (2), 231-239. M- 1
- KONIJN, H. S. (1956): Some estimates which minimize the least upper bound of a probability together with the cost of observation; *AISM*, 7, 143-158. M- 1
- KOOR, J. C. (1950): On a large-sample method of estimating unemployment in large cities; *Current Science*, 19, 232-233. S- 1
- KOOR, J. C. (1951a): Notes on the estimation of gross and net reproduction rates by methods of statistical sampling; *Biometrics*, 7, 155-166. S- 1
- KOOR, J. C. (1951b): A note on the bias of the ratio estimate; *BISI*, 23, (1), 141-146. M-10
- KOOR, J. C. (1962): On upper limits to the difference in bias between two ratio estimates; *Metrika*, 5, 143-149. M-10
- KOOR, J. C. (1963): On the axioms of sample formation and their bearing on the construction of linear estimators in sampling theory for finite universes, Parts I, II & III; *Metrika*, 7, 81-114; 165-201. M- 1
- KOOR, J. C. (1964): On an identity for the variance of a ratio of two random variables, *JRSS*, (B), 26, 484-486. M-10
- KOSHAL, R. S. (1947): *See* Sukhatme, P. V. (1947f).
- KOSHAL, R. S. (1959): *See* Sukhatme, B. V.
- KOZNIEWSKA, J. (1957): Comparison of the efficiency of drawing samples with and without replacement when the variance of the general population is unknown; *Colloquium Mathematicum*, 4, 232-234. M- 3
- KRAMER, R. C. and STRAFFER, J. D. (1954): The error for the mail survey; *JFE*, 36, 575-589. O- 1
- KRIESBERG, M. (1952): *See* Voight, R. B.
- KRISTOFF, W. (1963): Statistical inferences about the error variance; *M- 13 Psychometrika*, 28, 129-143.
- KUDO, A. and YAO, J. S. (1964): Some considerations on the multiple interval sampling method; *BMS*, 11, 1-2, 61-77. M-14
- KULKARNI, R. K. (1936): *See* Mahalanobis, P. C. (1936b).
- KULLBACK, S. and FRANKEL, A. (1940): A simple sampling experiment on confidence intervals; *AMS*, 11, 209-212. M- 1
- KULLDOOR, G. (1963): Some problems of optimum allocation for sampling on two occasions; *RISI*, 31, (1), 24-57. M-11
- LAHIRI, D. B. (1951a): A method of sample selection providing unbiased ratio estimates; *BISI*, 33, (2), 133-140. M- 6
- LAHIRI, D. B. and GANGULY, A. (1951b): An overall measure of precision of a simple table with applications in the study of relative effectiveness of different sampling units in population census; *BISI*, 33, (4), 55-74. M- 1

- LAHIRI, D B (1954a) Technical paper on some aspects of the development of the sample design, *National Sample Survey Report No 5*, Government of India, reprinted in *Sanhyā*, 14, 264-316 M- 1
- LAHIRI, D B (1954b) On the question of bias of systematic sampling, *Proceedings of World Population Conference*, 6, 349-362 M- 6
- LAHIRI, D B (1958a) Recent developments in the use of technique for assessment of errors in nation wide surveys in India *BISI*, 36, (2), 71-93 M-13
- LAHIRI, D B (1958b) Observations on the use of interpenetrating samples in India, *BISI*, 36, (3) 144-152 M-13
- LAHIRI D B (1961) See Mahalanobis, P C
- LAHIRI, D B (1963) Multi subject sample survey system—Some thoughts based on Indian experience, *Contributions to Statistics*, 175-220 Presented to Professor P C Mahalanobis on the occasion of his 70th Birthday, Pergamon Press, London and Statistical Publishing Society, Calcutta M-1
- LANSING, J B, KATONA, G, KISH, L and DENT J K (1950) Methods of the survey of consumer finances *Federal Reserve Bulletin* 36, 795-809 L- 1
- LANSING J B (1954) See Kish L
- LANSING, J B and EAPEV, (1959) Dealing with missing information in surveys, *J Marketing*, 24, 21-27 M 13
- LARSEN, O N and LUNDBERG, (1949) Characteristic of 'hard to reach' individuals *POQ*, 13 487-494 O-13
- LARSON N G (1941) The sampling method in social and economic research, a partial list of references *United States Bureau of Agricultural Economics Bibliography* No 90. S- 1
- LARSON, R F and CARTOY W R (Jr) (1959) Can the mail back bias contribute to a study's validity, *ASR*, 24, 243-245 W-14
- LAZARSFELD, P F and FISKE M (1938) The 'Panel' as a new tool for measuring opinion *POQ*, 2 596-615 O- 1
- LAZARSFELD, P F (1940) Panel studies *POQ* 4 122-128 O- 1
- LEAVENS, D H and BROWN, T H (1951) A sampling device, *Amer S tat* 5, (4), 16-21 M- 1
- LEGATT, C W (1941) A study of the relative efficiency of several sampling methods *Canadian J Research* 19, 156-162 M- 1
- LESLIE, P H and CHITTY, D (1951) The estimation of population parameters from data obtained by means of capture recapture method *Biometrika*, 38 269-292 W-14
- LESLIE, P H (1952) Estimation by capture recapture method, *Biometrika*, 39, 363-388 W-14
- LESLIE, P H, CHITTY, D and CHITTY, H (1953) Estimation of population parameters from data obtained by means of the capture recapture method, *Biometrika*, 40, 137-169 W-14
- LEXEN, B (1941) The application of sampling to log scaling, *J Forestry*, 36, 624-631 F- 1
- LEXEN B (1947) The determination of net volume by sample tree measuring, *J Forestry*, 45, 21-32 F- 1

BIBLIOGRAPHY

625

- LIEBERMAN, M. D. (1958) : Philippine Statistical Program development and the survey of households; *JASA*, 53, 78-88. S- 1
- LINK, H. C. (1937) : How many interviews are necessary for results of a certain accuracy; *J. Applied Psychology*, 21, (1), 1-17. M- 4
- LIPSTEIN, B. (1951) : See Cohen, S. E.
- LISSANCE, D. E. (1955) : See Greenberg, A.
- LOOMIS, R. D. (1946) : Accuracy in timber estimating; *Forest Chronicle*, 22, 1-1 201-202.
- LOWE, F. E. and McCORMICK, T. C. (1955) : Some survey sampling biases; *POQ*, 19, 303-315. O-13
- LUNDBERG, G. A. (1949) : See Larsen, O. N.
- LYDALL, H. F. (1954) : The methods of the savings survey; *Bulletin of Oxford University Institute of Statistics*, 16, 197-214. L- 1
- MACCOBY, E. E. (1948) : See Goodman, R. L- 1
- MACURA, M. and BALABAN, V. (1961) : Yugoslav experience in evaluation of population censuses and sampling; *B ISI*, 38, (2), 375-399. P-13
- MADHAVA, K. B. (1939) : Technique of random sampling; *Sankhyā*, 4, 532-534. M- 1
- MADOW, L. H. (1944) : See Madow, W. G.
- MADOW, L. H. (1946) : Systematic sampling and its relation to other sampling designs; *JASA*, 41, 204-217. M- 5
- MADOW, L. H. (1950) : On the use of the county as a primary sampling unit for State estimates, *JASA*, 45, 30-47. M- 9
- MADOW, W. G. and MADOW, L. H. (1944) : On the theory of systematic sampling; *AMS*, 15, 1-24. M- 5
- MADOW, W. G. (1948) : On the limiting distributions of estimates based on samples from finite universes; *AMS*, 19, 535-545. M- 1
- MADOW, W. G. (1949) : On the theory of systematic sampling II; *AMS*, 20, 333-354. M- 5
- MADOW, W. G. (1953) : On the theory of systematic sampling. III; *AMS*, 24, 101-106. M- 5
- MAHALANOBIS, P. C., BOSE, S. S., RAY, P. R. and BANERJEE, S. K. (1934) : Tables for random samples from a normal population; *Sankhyā*, 1, 289-328, (for the revised tables see Sengupta, J. M. and Bhattacharya, N. (1958)). M- 1
- MAHALANOBIS, P. C. (1936a) : Editorial note on the margin of error in the calculation of the cost of cultivation and profit; *Sankhyā*, 2, 362-378. A- 1
- MAHALANOBIS, P. C., KULKARNI, R. K. and BOSE, S. S. (1936b) : On the influence of shape and size of plots on the effective precision of field experiments with Juar; *Indian J. Agriculture Science*, 6, 460-. C- 8
- MAHALANOBIS, P. C. (1938a) : A note on grid sampling; *Science and Culture*, 4, 300. M- 1
- MAHALANOBIS, P. C. (1938b) : *Statistical Report on the Experimental Project Census, 1937*; Indian Central Jute Committee. C- 1

- MAHALANOBIS, P. C (1939) *First Report of the Crop Census of 1938*, Indian Central Jute Committee C- 1
- MAHALANOBIS, P. C (1940a) A Sample survey of the acreage under jute in Bengal, *Sankhya*, 4, 511-530 C- 1
- MAHALANOBIS, P. C (1940b) *Report on the Sample Census of Jute in 1939*, Indian Central Jute Committee C- 1
- MAHALANOBIS, P. C (1940c) *Statistical Report on Crop Cutting Experiments on Jute*, 1940, Indian Central Jute Committee C- 1
- MAHALANOBIS, P. C (1941a) Statistical survey of public opinion, *Modern Review*, 393-397 O- 1
- MAHALANOBIS, P. C (1941b) A note on random fields, *Science and Culture*, 7, 54 M- 1
- MAHALANOBIS, P. C (1941c) *Statistical Report on Crop-cutting Experiments on Jute in Bengal*, 1940, Indian Central Jute Committee C- 1
- MAHALANOBIS, P. C (1941d) *General Report on the Sample Census of Area under Jute in Bengal*, 1941, Indian Central Jute Committee C- 1
- MAHALANOBIS, P. C (1942) Sample Surveys 1942 Presidential Address (Mathematics and Statistics Section), *Proceedings of the Indian Science Congress*, 25-46 C- 1
- MAHALANOBIS, P. C (1943) An enquiry into the prevalence of drinking tea among middle class families in Calcutta 1939 *Sankhya*, 6, 283-312 S- 1
- MAHALANOBIS, P. C (1944) On large scale sample surveys, *Philosophical Transactions of Royal Society*, London, 231 (B), 329-451 M- 1
- MAHALANOBIS, P. C (1945) Report on the Bihar crop survey Rabi season, 1943-44, *Sankhya* 7, 29-106 C- 1
- MAHALANOBIS, P. C (1946a) Recent experiments in statistical sampling in the Indian Statistical Institute *JRSS*, (A), 109, 325-378, reprinted in *Sankhya*, 20, (1958), 1-68 M-13
- MAHALANOBIS, P. C (1946b) Sample surveys of crop yields in India *Sankhya*, 7, 269-280 C- 1
- MAHALANOBIS, P. C (1946c) Use of small size plots in sample surveys for crop yields, *Nature*, 158, 798-799 C- 8
- MAHALANOBIS, P. C, MUKHERJEE, R. K and GHOSH, A (1946d) A sample survey of after effects of the Bengal famine of 1943 *Sankhya*, 7, 337-400 S- 1
- MAHALANOBIS, P. C (1950) Cost and accuracy of results in sampling and complete enumeration *BISI*, 32, (2), 210-213 C- 1
- MAHALANOBIS, P. C and SENGUPTA, J. M (1951) On the size of sample cuts in crop cutting experiments in the ISI 1939-1950, *BISI*, 33, (2), 359-403 C- 8
- MAHALANOBIS, P. C (1952) Some aspects of the design of sample surveys, *Sankhya*, 12, 1-7 M- 1
- MAHALANOBIS, P. C and SEN, S. B (1954) On some aspects of the Indian National Sample Survey, *BISI*, 34, (2), 5-14 S-13
- MAHALANOBIS, P. C (1958a) Some observations on the 1960 world Census of Agriculture, *BISI*, 36, (4), 214-220 A- 1

- MAHALANOBIS, P. C. (1958): A method of fractile graphical analysis with some surmises of results; *Transactions of the Box Research Institute*, 22, 223-230. M- 1
- MAHALANOBIS, P. C. (1960): A method of fractile graphical analysis; *Biometrika*, 28, (2), 325-351, reprinted in *Sankhyā*, 23, (1), 1961, 41-74. M- 1
- MAHALANOBIS, P. C. and LAHIRI, D. B. (1961): Analysis of errors in censuses and surveys with special reference to experience in India; *BJSI*, 35, (2), 401-433, reprinted in *Sankhyā*, 23, (4), 325-358. M- 13
- MAHALANOBIS, P. C. (1966): Some concepts of sample surveys in demographic investigations; *Sankhyā*, 28, (A), 199-204. P-13
- MALLIE, A. K., SATAGOPAN, V. and RAO, S. G. (1945): A study of the estimation of the yield of wheat by sampling; *Ind. J. Agr. Sci.*, 15, 219-225. C- 1
- MANDAL, B. J. (1953): Sampling the federal old age and survivors insurance records; *JASA*, 48, 462-475. S- 1
- MANDEVILLE, J. P. (1946): Improvements in methods of census and survey analysis; *JRSS*, 109, 111-129. M- 1
- MANOUS, A. R. (1934): Sampling in the field of rural relief; *JASA*, 29, 410-415. S- 1
- MANHEIMER, D. and HYMAN, H. (1949): Interviewer performance in area sampling; *POQ*, 13, 83-92. O-13
- MANN, H. B. (1945): On a problem of estimation occurring in public opinion polls; *AMS*, 16, 85-90; *AMS*, 17, (1946), 87. O- 1
- MANTEL, N. (1951): Rapid estimation of standard errors of means for small samples; *Amer. Stat.*, 5, (4), 26-27. M- 1
- MARCUSE, S. (1949): Optimum allocation and variance components in nested sampling with an application to chemical analysis; *Biometrika*, 36, 189-206. M- 2
- MARKET RESEARCH TECHNICAL COMMITTEE (1946): Design, size and validation of sample for Market Research; *J. Marketing*, 10, 221- . O- 1
- MARKS, C. L. (1956): See Gilford, D. M.
- MARKS, E. S. (1949): See Mosteller, F.
- MARKS, E. S. and MAULDIN, W. P. (1950a): Problems of response in enumeration surveys; *ASR*, 15, 649- . P-13
- MARKS, E. S. and MAULDIN, W. P. (1950b): Response errors in census research; *JASA*, 45, 424-438. P-11
- MARKS, E. S. (1951): See Hansen, M. H.
- MARKS, E. S., MAULDIN, W. P. and NISSELSON, H. (1953): The post-enumeration survey of the 1950 census: A case history in the survey designs; *JASA*, 48, 220-243. P-13
- MARKS, E. S. (1958): See Hansen, M. H.
- MARKS, E. S. (1962): The fetish of sample size; *POQ*, 26, 92-97. M- 4
- MARUYAMA, E., HAYASHI, C. and ISIDU, M. D. (1950): On some criteria for stratification; *AISM*, 2, 77-86. M- 7
- MASUYAMA, M. (1951): Recent advances in Sample Surveys in Japan. *BJSI*, 33, (2), 147-151. M- 2

- MASUYAMA, M (1953) A rapid method of estimating basal area in timber survey, *Sankhya*, 12, 291-302 F- 1
- MASUYAMA, M (1954a) Analysis of the 1939 model sample survey results from the view point of integral geometry, *Sankhya*, 13, 229-234 C- 1
- MASUYAMA, M (1954b) Mathematical note on area sampling *Sankhya* 13, 241-242 M- 1
- MASUYAMA, M (1954c) On the error in crop cutting experiment due to the bias on the border of the grid, *Sankhya*, 14 181-186 C-13
- MASUYAMA, M and SENGUPTA J M (1955) On a bias in a crop cutting experiment, *Sankhya*, 15, 373-376 C-13
- MASUYAMA, M (1957) Ratio estimate in line grid sampling, *BMS*, 7, 73-76 M-10
- MATHEN, K K (1948) Studies on the sampling procedure for a general health survey, *BCSA*, 1, 106-113 S- 1
- MATTHAI A (1951) Estimation of parameters from incomplete data with application to design of sample surveys, *Sankhya*, 11, 145-152 M- 1
- MATTHAI, A (1954) On selecting random numbers for large scale sampling, *Sankhya*, 13 257-260 M- 1
- MATUSITA N and OTHERS (1955) Some problems of sampling in the forest survey, *AISM*, 7, 1-24 F- 1
- MAULDIN, W P (1950a) See Marks, E S (1950a)
- MAULDIN, W P (1950b) See Marks E S (1950b)
- MAULDIN, W P (1951) See Hansen, M H
- MAULDIN, W P (1953) See Marks, E S
- MAYNES, E S (1965) See Neter, J
- MCCANDLISS, D A (1941) Objective sampling in estimating southern crops *JFE*, 23, 246-255 C- 1
- MCCAETHY, D E (1941) See King, A J
- MCCAETHY, D E (1942) See King A J (1942b)
- MCCAETHY, P J (1947) See Stephan, F F
- MCCAETHY, P J (1949) See Mosteller, F
- MCCAETHY, P J (1951) Sampling—Elementary principles, *New York State School of Industrial and Labour Relations Bulletin*, No 15 M- 1
- MCCARTHY, P J (1965) Stratified sampling and distribution free confidence intervals for a median, *JASA*, 60, 772-783 M- 7
- MCCORMICK, T C (1937) Sampling theory in sociological research, *Social Forces*, 16, 67-74 S- 1
- MCCORMICK T C (1955) See Lowe, F E
- MCHUGH, R B (1961) Confidence interval inference and sample size determination, *Amer Stat*, 15, (2) 14-17 M- 4
- McKEON, A J (1953) See Price, O O
- McKEON, A J (1958) See Rosander, A C
- MCMENAMER, Q (1940) Sampling in psychological research, *Psychological Bulletin*, 37, 331-365 S- 1
- MCPEAK, M (1942) See King, A J (1942b),

- MCVAY, F. E. (1947): Sampling methods applied to estimating numbers of commercial orchards in a commercial peach area; *JASA*, 42, 533-540 C- 1
- MEDIN, K. (1965): Crop yield estimation and crop insurance in Sweden; *RISI*, 83, (3), 414-412. C- 1
- MEIER, N. C. and BURKE, C. J. (1947): Laboratory tests of sampling techniques; *POQ*, 11, 586-593. C- 1
- MEIER, N. C., BURKE, C. J. and BANKS, S. (1948): Laboratory tests of sampling techniques: comments and rejoinders; *POQ*, 12, 316-324. N- 1
- MEIER, N. C. and HANER, C. F. (1951): The adoptability of area probability sampling to public opinion measurements; *POQ*, 15, 337-352. O- 14
- MELLOR, J. W. (1952): Sample bias from the elimination of poultry men who don't keep financial records; *JFE*, 34, 119-123. O- 6
- MEYER, M. T. and PATCH, L. H. (1937): A statistical study of sampling in field surveys of the fall population of the European Corn Borer; *JAR*, 55, 840-872. A-13
- MICKEY, M. R. (1959): Some finite population unbiased ratio and regression estimators; *JASA*, 54, 594-612. M-10
- MIDZUNO, H. (1950): An outline of the theory of sampling systems; *AISM*, 1, 149-156. M- 1
- MIDZUNO, H. (1952a): On the sampling system with probability proportional to sum of sizes; *AISM*, 3, 99-107. M- 6
- MIDZUNO, H. (1952b): Report of the survey design for agricultural production estimates in Ryuku Islands; *AISM*, 3, 109-121. C- 1
- MIDZUNO, H. (1961): On the post-enumeration survey; *BISI*, 39, (2), 436-441. P-13
- MILNE, A. (1959): The centric systematic area sample treated as a random sample; *Biometrics*, 15, 270-297. M- 5
- MENTON, G. (1964): See Fastau, H. H.
- MITCHELL, W. (JR.) (1939): Factors affecting the rate of return on mailed questionnaires; *JASA*, 34, 683-692. E- 1
- MOKASHI, V. K. (1950): A note on interpenetrating samples; *JISAS*, 2, 189-195. M-13
- MOKASHI, V. K. (1953): Investigations on sampling for estimation of crop acreages I; *JISAS*, 5, 128-143. A-10
- MOKASHI, V. K. (1954a): Efficiency of stratification in sub-sampling designs for the ratio method of estimation; *JISAS*, 6, 77-82. M- 7
- MOKASHI, V. K. (1954b): Efficiency of sampling methods in forest surveys; *JISAS*, 6, 101-114. P- 5
- MOKASHI, V. K. (1954c): Investigations of sampling for estimation of crop acreages—II; *JISAS*, 6, 115-126. C- 9
- MOLINA, E. C. (1946): Some fundamental curves for the solution of sampling problems; *AMS*, 17, 325-335. M- 1
- MONROE, R. J. (1943): See Finkner, A. L.
- MOORE, P. G. (1957): Sampling techniques and some applications; *J. Institute of Actuaries Students' Society*, 14, 111-128. M- 1

- MORAN, P A P (1951) A mathematical theory of animal trapping, *Biometrika*, 38, 307-311 W- 1
- MORAN, P A P (1952) The estimation of death rates from capture mark recapture sampling, *Biometrika*, 39 181-188 W-14
- MORGAN J N (1943) See Finkner, A L
- MORGAN, J N (1951) See Klein, L R
- MORGAN, J N and SONQUIST, J A (1963) Problems in the analysis of survey data and a proposal *JASA*, 58, 415-434 M- 1
- MORITA Y (1951) Sampling tabulation of the 1950 Population Census in Japan, *BISI* 33 (4) 47-54 P- 1
- MORITA, Y (1958) The accuracy of age reporting in the population Census, *BISI*, 36 (2), 183-189 P-13
- MORRELL A J H (1950) The estimation of age standardized ratios by sampling methods *Inc Stat*, 1, 17-20 P- 1
- MORRIS, K W (1963) A note on direct and inverse binomial sampling, *Biometrika*, 50 544-545 M- 3
- MOSER, C A (1949) The use of sampling in Great Britain, *JASA*, 44, 231-259 M- 2
- MOSER C A (1951a) Remarks on sampling aspects of family expenditure surveys *BISI*, 33 (2), 189-196 L- 1
- MOSER C A (1951b) Interviewer Bias *RISI*, 19, (1), 28-40 S-13
- MOSER, C A (1952) Quota Sampling, *JRSS* (A), 115, 411-423 M-14
- MOSER C A and STUART, A (1953) An experimental study of quota sampling *JRSS* (A) 116, 349-405 M-14
- MOSER C A (1955) Recent Developments in the sampling of human populations in Great Britain *JASA*, 50 1195-1214 P- 2
- MOSHMAN, J (1958) A method of selecting the size of the initial sample in Stein's two sample procedure, *AMS*, 29 1271-1275 M- 4
- Moss L (1950) The Government Social Survey, *Operations Research Quarterly*, 1 55-65 L- 1
- MOSTELLER F, HYMAN, H, McCARTHY, P J and MARKS, E S (1949) *The Pre Election Polls of 1948*, Social Science Research Council, New York O- 1
- MOSTELLER, F (1953) See Cochran, W G
- MOSTELLER, F (1954) See Cochran, W G
- MUDGETT, B D (1929) The application of the theory of sampling to successive observations not independent of each other, *JASA*, (Supplement), 24, 108-113 M- 1
- MUDGETT, B D and GEVORKIANTZ, S R (1934) Reliability of forest surveys, *JASA*, 29, 257-281 F- 1
- MUKHERJEE, R K (1946) See Mahalanobis, P C (1946d)
- MUKHERJEE, V (1965) A note on the optimum sample size when there are non sampling errors, *JISAS*, 17, 30-33 M-13
- MULLER, M E (1962) See Fan C T
- MULLOY, G A (1946) See Robertson, W M
- MURTHY, M N (1957) Ordered and unordered estimators in sampling without replacement, *Sankhya*, 18, 379-390 M- 6

BIBLIOGRAPHY

631

- MURTHY, M. N. (1959a) : See Nanjamma, N. S. (1959a).
- MURTHY, M. N. and NANJAMMA, N. S. (1959b) : Almost unbiased ratio estimators based on interpenetrating sub-sample estimates; *Sampling*, 21, 381-392. M-10
- MURTHY, M. N. and SETHI, V. K. (1961) : Randomized rounded-off multipliers in sampling theory; *JASA*, 56, 329-334. M-12
- MURTHY, M. N. (1962a) : Variance and confidence interval estimation; *Sankhyā*, 24, (B), 1-12. M- 1
- MURTHY, M. N. (1962b) : Almost unbiased estimators based on interpenetrating sub-samples; *Sankhyā*, 24, (A), 303-314. M- 1
- MURTHY, M. N. (1962c) : Technical Paper on Sample Design—National Sample Survey, Fourteenth Round (July 1958-June 1959); Report No. 76, Government of India, New Delhi. M- 1
- MURTHY, M. N. (1963a) : Some recent advances in sampling theory; *JASA*, 58, 737-755. M- 1
- MURTHY, M. N. (1963b) : A note on determination of sample size; *Sankhyā*, 25, (A), 351-382. M- 2
- MURTHY, M. N. (1963c) : Generalized unbiased estimation in sampling from finite populations; *Sankhyā*, 25, (B), 245-262. M- 3
- MURTHY, M. N. (1963d) : Assessment and control of non-sampling errors in censuses and surveys; *Sankhyā*, 25, (B), 263-282. M- 2
- MURTHY, M. N. (1963e) : On Mahalanobis' contributions to the development of sample survey theory and methods; *Contributions to Statistics*, Presented to Professor P. C. Mahalanobis on the occasion of his 70th Birthday, Pergamon Press, London, and Statistical Publishing Society, Calcutta. M- 2
- MURTHY, M. N. (1964a) : Product method of estimation; *Sankhyā*, 26, (A) 69-74. M-14
- MURTHY, M. N. (1964b) : The work of the United States Bureau of the Census with emphasis on sample designs and control of errors in censuses and surveys; *Sankhyā*, 26, (B), 257-300. M- 1
- MURTHY, M. N. and SETHI, V. K. (1965) : Self-weighting design at tabulation stage, *Sankhyā*, 27, (B), 201-210. M-12
- MUSHMAN, H. V. (1959) : Population estimates based on census enumeration and coverage check; *Population Studies*, 13, 278-281. P-13
- MYBURG, C. A. L. (1948) : See Shaul, J. R. H.
- MYBURG, C. A. L. (1949) : See Shaul, J. R. H.
- MYERS, R. J. (1954) : Accuracy of age reporting in the 1950 US Census; *JASA*, 49, 826-831. P-13
- NAGASWA, R. (1930) : The method of statistical investigation concerning agricultural production in Japan; *BISI*, 25, (2), 149-178. C- 1
- NAIR, K. R. (1950) : Sampling techniques; *Indian Forester*, 76, 31-35. F- 1
- NAIR, K. R. and BHARGAVA, R. P. (1951) : Statistical sampling in timber surveys in India; Forest Research Institute, Dehradun, *Indian Forest Leaflet*, No. 153. F- 1

- NAMBOODIRI, N K (1962) *See* Kish, L (1962b)
- NANDA, D N (1951) Application of cluster sampling for estimating loss during storage, *BISI*, 33, (5), 1-5
- NANDI H K (1948) Choosing a random sample, *BCSA*, 1, 143-153
- NANDI, H K (1950) Indian National Sample Survey, *BCSA*, 3, 11-20
- NANJAMMA, N S, MURTHY, M N and SETHI, V K (1959a) Some sampling systems providing unbiased ratio estimators, *Sankhya*, 21, 299-314
- NANJAMMA, N S (1959b) *See* Murthy, M N (1959b)
- NANJAMMA CHINNAPPA (1963) Technical paper on sample designs of working class and middle class family living surveys (1958-59), *Sankhya*, 25, (B), 359-418
- NARAIN, R D (1951) On sampling without replacement with varying probabilities, *JISAS*, 3 169-174
- NARAIN, R D (1952) *See* Sukhatme, P. V (1952b)
- NARAIN, R D (1953) On the recurrence formula in sampling on successive occasions, *JISAS*, 5 96-99
- NARAIN, R D (1954a) The general theory of sampling on successive occasions, *BISI*, 34, (2), 87-89
- NARAIN, R D (1954b) On the recurrence formula in sampling on successive occasions, *BISI*, 34, (2) 201-202
- NARAIN, R D (1955) Problems of sampling in Latin America, *MBAES*, 4 8-12
- NETER, J and WAKSBERG, J (1964) Conditioning effects from repeated household interviews, *J Marketing*, 28, (2), 51-56
- NETER, J, MAYNES, E S and RAMANATHAN, R (1965) The effect of mismatching on the measurement of response errors *JASA*, 60, 1005-1027
- NEYMAN, J (1934) On the two different aspects of the representative method. The method of stratified sampling and the method of purposive selection, *JRSS*, 97, 558-625
- NEYMAN, J (1938) Contributions to the theory of sampling human populations, *JASA*, 33, 101-116
- NICHOLSON, J L (1954) A survey of the living conditions of Arab families in Tripolitania in November December 1950 *RISI*, 22 (1-3), 68-84
- NICHOLSON, J L (1960) *See* Kemsley, W F F (1960b)
- NISSELSON, H (1953) *See* Marks, E S
- NISSELSON, H (1955) *See* Hansen, M H
- NISSELSON, H and WOOLSEY, T D (1959) Some problems of the household interview design for the National Health Survey, *JASA*, 54, 69-87
- NORDBOTTEL, S (1954) On the determination of an optimal sample size, *Skand Akt*, 37, 60-64
- NORDBOTTEL, S (1956) Allocation in stratified sampling by means of linear programming, *Skand Akt*, 39, 1-6
- NORDBOTTEL, S (1957) On errors and optimal allocation in a census, *Skand Akt*, 40, 1-10
- NORDIN, J A (1944) Determining sample size, *JASA*, 39, 497-506

C- 1

M- 1

S- 1

M-10

L- 1

M- 6

M-11

M-11

M-11

M- 1

M- 1

S-13

M-10

M- 7

M- 1

L- 1

S-13

M- 4

M- 7

P- 1

M- 4

BIBLIOGRAPHY

633

- NORDSKOG, A. W. and CRUMPT, S. L. (1948) : Systematic and random sampling for estimating egg production in poultry; *Biometrics*, 4, 223-233. M- 5
- NORTHROP, M. S. (1943) : See Webb, J. N.
- OAKLAND, G. B. (1950) : An application of sequential analysis to white-fish sampling; *Biometrics*, 6, 59-67. W- 1
- OBROCK, R. F. (1958) : A case study of statistical sampling; *J. Accountancy*, 53-59. E- 1
- OLDS, E. G. (1940) : On a method of sampling; *AMS*, 11, 355-358. M- 1
- OLKIN, I. (1958) : Multivariate ratio-estimation for finite populations; *Biometrika*, 45, 154-165. M-10
- OSBORNE, J. G. (1942) : Sampling errors of systematic and random surveys of covertype areas; *JASA*, 37, 256-264. F- 5
- OSGOOD, O. T. (1949) : Results of two sampling methods used in farm management research; *JFE*, 31, 157-167. A- 1
- OVERTON, R. S. (1949) : Use of semi-controlled mail surveys for initiating new statistical series; *AER*, 1, 87-. A- 1
- OWEN, P. C. (1957) : Rapid estimation of the areas of the leaves of crop plants; *Nature*, 180, 611. C- 1
- PALCA, H. (1949) : See Finney, D. J. (1949b).
- PALCA, H. (1953) : An experiment in the sampling of agricultural returns; *Appl. Stat.*, 2, 152-159. A- 1
- PALCA, H. (1955) : Some sampling problems in agriculture; *Inc. Stat.*, 6, 18-33. C- 1
- PALMER, G. L. (1942) : The reliability of response in labour market enquiries; *Technical paper*, No. 22, United States Bureau of Budget, Washington, D. C. S-13
- PALMER, G. L. (1943) : Factors in the variability of response in enumerative surveys; *JASA*, 38, 143-152. P-13
- PANSE, V. G. (1938) : Preliminary studies on sampling in field experiments; *Sankhyā*, 4, 139-148. C- 1
- PANSE, V. G. and SAHASRABUDHE, V. B. (1943) : A rapid method of sampling for fibre weight determination in cotton; *Indian J. Genetics*, 3, 28-44. C- 1
- PANSE, V. G. and KALAMKAR, R. J. (1944a) : Forecasting and estimation of crop yields; *Current Science*, 13, 120-124. C- 1
- PANSE, V. G. and KALAMKAR, R. J. (1944b) : A further note on the estimation of crop yields; *Current Science*, 13, 223-225. C- 1
- PANSE, V. G. and SHALIGRAM, G. C. (1945) : A large scale yield survey on cotton; *Current Science*, 14, 287-291. C- 1
- PANSE, V. G. (1946a) : Plot size in yield surveys on cotton; *Current Science*, 15, 218-219. C- 8
- PANSE, V. G. (1946b) : *Report on the Scheme for the Improvement of Agricultural Statistics*; Imperial Council of Agricultural Research, New Delhi. C- 1
- PANSE, V. G. (1947) : Plot size in yield surveys; *Nature*, 159, 820. C- 8

- PANSE, V. G and SUKHATME, P. V (1948) Crop surveys in India I, *JISAS*, 1, 34-58 C- 1
- PANSE, V. G (1951) See Sukhatme, P. V (1951a)
- PANSE, V. G (1954) *Estimation of Crop Yields*, United Nations Food and Agricultural Organization, Rome C- 1
- PANSE, V. G (1958a) Some comments on the objective and method of the 1960 world census of Agriculture, *BISI*, 36, (4), 222-227 A- 1
- PANSE, V. G (1958b) See Sukhatme, P. V (1958b)
- PANSE, V. G (1963) Plot size again, *JISAS*, 15 151-159 C- 8
- PARKER, R. A (1963) On the estimation of population size, mortality and recruitment, *Biometrika*, 19, 318-323 W-14
- PARRY, H. J and CROSSLEY, A. M (1950) Validity of responses to survey questions, *POQ*, 14, 61-80 O-13
- PARTEV, M (1937) See Schoenberg E
- PASCUAL, J. N (1961) Unbiased ratio estimators in stratified sampling, *JASA*, 56 70-87 M-10
- PATCH, L. H (1937) See Meyer, M. T
- PATHAK, P. K (1961a) Use of 'order statistic' in sampling without replacement, *Sankhya*, 23, (A) 409-414 M- 6
- PATHAK, P. K (1961b) On the evaluation of moments of distinct units in sample, *Sankhya*, 23, (A) 415-420 M- 3
- PATHAK, P. K (1962a) On simple random sampling with replacement, *Sankhya*, 24 (A), 287-302 M- 3
- PATHAK, P. K (1962b) On sampling units with unequal probabilities, *Sankhya*, 24, (A), 315-326 M- 6
- PATHAK, P. K (1964a) On sampling schemes providing unbiased ratio estimates *AMS*, 35 222-231 M-10
- PATHAK, P. K (1964b) Sufficiency in sampling theory, *AMS*, 35, 795-808 M- 1
- PATHAK, P. K (1964c) On inverse sampling with unequal probabilities, *Biometrika*, 51, 185-193 M-14
- PATHAK, P. K (1964d) On estimating the size of a population and its inverse by capture work method, *Sankhya*, 26 (A), 75-80 M-14
- PATHAK, P. K (1966a) An estimator in pps sampling for multiple characteristics *Sankhya*, 28 (A), 35-40 M- 6
- PATHAK, P. K and SHUKLA, N. D (1966b) Non negativity of a variance estimator, *Sankhya*, 28 (A) 41-46 M- 6
- PATTERSON, D. G and HENEMAN, H. G (1949) Refusal rates and interviewer quality, *IJOAR*, 3 392-398 O-13
- PATTERSON, H. D (1950) Sampling on successive occasions with partial replacement of units *JRSS*, (B), 12, 241-255 M-11
- PATTERSON, H. D (1954) The errors of lattice sampling, *JRSS*, (B), 16, 140-149 M-14
- PAYNE, S. L (1943) See Webb, J. N
- PAYNE, S. L (1944). See Hilgard, E. R

- PEAKER, G. F. (1953): A sampling design used by the Ministry of Education; *JRSS*, (A), 116, 140-165. S- 9
- PEARCE, S. C. (1944): Sampling methods for the measurements of fruit crops; *JRSS*, 107, 117-126. C- 1
- PEARCE, S. C. and HOLLAND, D. A. (1957): Randomized branch sampling for estimating fruit number; *Biometrics*, 13, 127-130. C- 6
- PECHANCE, J. (1941): Sampling error in range surveys of sago brush grass vegetation; *J. Forestry*, 39, 52-54. F- 1
- PERRY, P. (1962): Gallup poll election survey experience: 1950-60; *POQ*, 26, 272-279. O- 1
- PHILIPPINES UNIVERSITY STATISTICAL CENTRE (1962): *Seminar on Sampling and Sample Surveys*; Proceedings of the Seminar, Manila. M- 1
- PILLAI, R. K. (1962): See Kish, L. (1962b).
- POLITZ, A. and SIMMONS, W. (1949): An attempt to get the 'not at home' into the sample without callbacks; *JASA*, 44, 9-31; 45, (1950), 136-137. M-13
- POTI, S. J. (1955): Measures of overall efficiency of sample multinomial tables; *BCSA*, 6, 102-112. M- 1
- POTTER, R. C. (JR.) (1961): See Westoff, C. F.
- PRABHU AJGAONKAR, S. G. (1965): On a class of linear estimators in sampling with varying probabilities without replacement; *JASA*, 60, 637-642. M- 6
- PRICE, O. O. and McKEON, A. J. (1953): A comparison of serial number digit sampling with systematic and random sampling; *Appl. Stat.*, 2, 39-43. M- 5
- PRITZKER, L. (1951a): See Ackoff, R. L.
- PRITZKER, L. (1951b): See Eckler, A. R.
- PRITZKER, L. (1953): See Hansen, M. H.
- PRITZKER, L. (1963): See Hansen, M. H. (1963b).
- PROUDFOOT, M. J. (1942): Sampling with transverse traverse lines; *JASA*, 37, 265-270. F- 1
- QUENOUILLE, M. H. (1949): Problems in plane sampling; *AMS*, 20, 355-375. M-14
- QUENOUILLE, M. H. (1956): Notes on bias in estimation; *Biometrika*, 43, 353-360. M-10
- QUENSEL, C. E. (1958): Some sampling problems when a stratification variable follows a logarithmic normal distribution; *Skand. Akt.*, 41, 177-184. M- 7
- QURESHI, O. M. (1955): Crop estimation surveys in the Punjab; Pakistan CSO Bull.; *Proceedings of the Pakistan Statistical Association*, 3 and 4, 126- . C- 1
- RAMACHANDRAN, K. V. (1963): See BASAVARAJAPPA, K. C.
- RAMAMURTI, B. (1954): Agricultural Labour enquiry in India; *BISI*, 34, (2), 580-587. S- 1
- RAMANATHAN, R. (1965): See Neter, J.
- RANGARAJAN, R. (1957): A note on two-stage sampling; *Sankhyā*, 17, 373-376. M- 9

- RAO, J N K and CHAWLA, H K (1956) Efficiency of stratification in sub sampling designs for the ratio method of estimation with varying probabilities of selection, *JISAS*, 8, 91-101 M-10
- RAO, J N K (1957) Double ratio estimate in forest surveys, *JISAS*, 9, 191-204 M-10
- RAO, J N K (1961a) On sampling with varying probabilities with replacement in sub sampling designs *JISAS*, 13, 211-217 M- 6
- RAO, J N K (1961b) On the estimate of variance in unequal probability sampling, *AISM*, 13, 57-60 M- 6
- RAO, J N K (1962a) See Hartley, H O
- RAO, J N K, HARTLEY, H O and COCHRAN, W G (1962b) On a sample procedure of unequal probability sampling without replacement, *JRSS*, (B), 24 482-491 M- 6
- RAO J N K (1962c) On the estimation of relative efficiency of sampling procedures, *AISM*, 14, 143-150 M- 1
- RAO, J N K (1963a) On the three procedures of unequal probability sampling without replacement, *JASA*, 58, 202-215 M- 6
- RAO J N K (1963b) On two systems of unequal probability sampling without replacement, *AISM*, 15, 67-72 M- 6
- RAO, J N K and GRAHAM, J E (1964a) Rotation designs for sampling on repeated occasions, *JASA*, 59, 492-509 M-11
- RAO, J N K (1964b) Unbiased ratio and regression estimators in multi-stage sampling, *JISAS*, 16, 175-183 M- 9
- RAO, J N K (1964c) See Seth, G R
- RAO, J N K (1965) A note on estimation of ratios by Quenouille's method, *Biometrika* 52, 647-649 M-10
- RAO, J N K (1966a) Alternative estimators in pps sampling for multiples characteristics *Sankhya*, 28, (A), 28, 47-60 M-6
- RAO, J N K (1966b) On the relative efficiency of some estimators in pps-sampling for multiple characteristics, *Sankhya*, 28, (A), 61-70 M-6
- RAO, M S (1959) See Johnson P O
- RAO, S G (1945) See Mallik, A K
- RAO, M J (1966a) On certain unbiased ratio estimators, *AISM*, 18, 117-121 M-10
- RAO, T J (1966b) On the variance of the ratio estimator for Midzuno Sen sampling scheme, *Metrika*, 10, 89-91 M-10
- RAO, V R (1964) See Fattu, N A
- RAY, P R (1934) See Mahalanobis, P C
- RENYI, A (1959) See Erdos, P
- REZUCHA, I (1962) See Fan, C T
- RICE, S A (1929) Contagious bias in interview, *American J. Sociology*, 35, 420-423 M-13
- RIDDERSTRÖM, S (1955) On ratio estimates in simple random sampling with some practical applications, *Skand Akt*, 38, 135-162, M-10
- RIEDEL, D C (1961) See Hess, I.

- RIESMAN, D. (1961): *See* Ehrlich, J. S.
- RIGNEY, J. A. and BLASER, R. E. (1948): Sampling Alyce Clover for chemical analysis; *Biometrics*, 4, 234-239. C- 1
- ROBERTS, B. J. (1965): *See* Yankey, D.
- ROBERTSON, W. M. and MULLOY, G. A. (1946): *Sample Plot Methods*; Dominion Forest Service, Ottawa. F- 1
- ROBINSON, H. L. (1949): *See* Keyfitz, N.
- ROBSON, D. S. and LING, A. J. (1952): Multiple sampling of attributes; *JASA*, 47, 203-215; 48, (1953), 911. M- 1
- ROBSON, D. S. (1957): Applications of multivariate polykays to the theory of unbiased ratio-type estimation; *JASA*, 52, 511-522. M-10
- ROBSON, D. S. (1960): An unbiased sampling and estimation procedure for Creel censuses of fisherman; *Biometrics*, 16, 261-277. W- 1
- ROBSON, D. S. and VITHAYASAI, C. (1961): Unbiased component-wise ratio estimation; *JASA*, 56, 350-358. M-10
- ROPER, E. (1940) Sampling public opinion; *JASA*, 35, 325-334. O- 1
- ROPER, E. (1941a): Checks to increase polling accuracy; *POQ*, 5, 87-90. O-13
- ROPER, E. (1941b): Problems and possibilities of the sampling technique; *Journalism Quarterly*, 1-9. O- 1
- ROSANDER A. C., BLYTHE, R. H. and JOHNSON, D. G. (1951): Sampling 1949 corporation income tax returns; *JASA*, 46, 233-241. E- 1
- ROSANDER, A. C., GUTERMAN, H. E. and McKEON, A. J. (1958): The use of random work sampling for cost analysis and control; *JASA*, 53, 382-397. E- 1
- ROSHWAHL, I. (1953): Effect of weighting by card duplication on efficiency of survey results; *JASA*, 48, 773-777; 49 (1954), 906. M- 1
- ROSS, A. (1961): Variance estimates in optimum sample designs; *JASA*, 56, 135-142. M- 7
- ROSS, A. (1954): *See* Hartley, H. O. (1954a).
- ROY, J. (1957): A note on estimation of variance components in multistage sampling with varying probabilities; *Sankhyā*, 17, 367-372. M- 9
- ROY, J. and CHAKRAVARTI, I. M. (1960): Estimating the mean of finite population; *AMS*, 31, 392-398. M- 3
- ROY, J. and KALYANASUNDARAM, G. (1963): Use of randomized rounded-off weights in sample surveys; *Sankhyā*, 25, (B), 333-340. M-12
- ROY, S. (1964): *See* Sengupta, J. M.
- ROY, S. N. (1940): *See* Banerjee, K. S.
- ROY CHOUDHURY, D. K. (1956): Integration of several pps surveys; *Science and Culture*, 22, 119-120. M- 6
- ROYER, J. (1959): Note on rural surveys covering food consumption and household expenditure in tropical West Africa; *MBAES*, 8, (1), 1-6. L- 9
- SAGI, P. C. (1961): *See* Westoff, C. F.
- SAHAI, R. (1947): Partial enumeration in the Dehradun forest division; *Indian Forester*, 73, 437-444. F- 1
- SAHASRABUDHE, V. B. (1943): *See* Panse, V. G.

- SAITO, K (1956) Maximum likelihood estimate of proportion using supplementary information, *BMS*, 7, 11-17 M-11
- SAITO, K (1957) Some results in the theory of sampling on successive occasions with partial replacement of units, *JUSE*, 4, 15-22 M-11
- SAMPFORD, M R (1962) Methods of cluster sampling with and without replacement for clusters of unequal sizes, *Biometrika*, 49, 27-40 M- 6
- SANDIFORD, P J (1960) A new Binomial approximation for use in sampling from finite populations, *JASA*, 55, 718-722 M- 3
- SANDOMIRE M M (1950) See Greenwood, J A
- SARC O C (1957) A household budget enquiry along sampling lines in Istanbul, *BISI*, 35 (2), 133-139 L- 1
- SARKAR D (1951) See Sengupta, J M (1951a)
- SARLE, C F (1929) The theory of sampling as applied to crop-estimation, Bureau of Agricultural Economics, United States Department of Agriculture, Washington D C C- 1
- SARLE, C F (1932) Adequacy and reliability of crop yields estimates, *United States Department of Agriculture Technical Bulletin*, No 311 C- 1
- SARLE, C F (1938) Methods in sample census research, *JFE*, 20, 669-672 M- 1
- SARLE, C F (1939a) Development of partial and sample census methods, *JFE*, 21, 356-364 M- 1
- SARLE, C F (1939b) Future improvement in agricultural statistics, *JFE*, 21 838-845 C- 1
- SARLE, C F (1940) The possibilities and limitations of objective sampling in strengthening agricultural statistics, *Econometrica*, 8, 45-61 C- 1
- SARLE, C F (1947) See Callander, W F
- SARLE, C F (1949) Need for special purpose sampling in estimating agricultural production *AER* 1, 134- C- 1
- SASTRY, K V R (1959) See Sukhatme, P V (1958b)
- SASTRY, K V R (1961) A note on sample design of the combined survey of coconut and arecanut in India, *JISAS*, 13, 80-86 C- 6
- SASTRY, K V R (1965) Unbiased ratio estimators, *JISAS*, 17, 19-29 M-10
- SATAGOPAN, V (1945) See Mallick, A K
- SAXENA, P N (1957) See Singh, D
- SCHAFFER, K A (1963) See Szamenteit, K
- SCHMITT, R C (1952) Short cut methods of estimating county population, *JASA*, 47, 232-238 P- 1
- SCHMITT, R C (1956) See Crosetti, A H
- SCHOENBERG, E and PARTEV, M (1937) Methods of problems of sampling presented by the urban study of consumer purchases, *JASA*, 32 311-322 L- 1
- SCHULTZ, T W (1933) Testing the significance of mean values drawn from stratified samples, *JFE*, 15, 452-473 M- 7
- SCHULTZ, T W (1952) Ten years of family surveys, *Bulletin of Oxford University Institute of Statistics*, 14, 83-95, L- 2

- SCHUMACHER, F. X. and BULL, H. (1932) : Determination of the errors of estimates in a forest survey with special reference to the Bottom Land Hardwood Forest Region; *JAR*, 44, 741-756. F- 1
- SCHUMACHER, F. X. (1948) : See Chapman, R. A.
- SCHUMAN, R. S. (1963) : See Carter, R. E. (Jr.).
- SCHUTZ, H. H. (1937a) : See Shepard, J. B.
- SCHUTZ, H. H. (1937b) : Selection of area sample for agricultural enumeration II : Tests of various sampling methods; *JFE*, 19, 464-469. A- 1
- SCOTT, C. (1961) : Research on mail surveys; *JRSS*, (A), 124, 143-205. S-13
- SEAL, K. C. (1951) : On errors of estimates in various types of double sampling procedure; *Sankhyā*, 11, 125-144. M-11
- SEAL, K. C. (1953) : On certain extended cases of double sampling; *Sankhyā*, 12, 357-362. M-11
- SEAL, K. C. (1962) : Use of out-dated frames in large-scale sample surveys; *BCSA*, 11, 68-84. E- 1
- SEARLE, S. R. (1958) : Sampling variances of estimates of components of variance; *AMS*, 29, 167-178. M- 1
- SEARLS, D. T. (1964) : The utilization of a known coefficient of variation in the estimation procedure; *JASA*, 59, 1225-1226. M- 3
- SEARLS, D. T. (1966) : An estimator for population mean which reduces the effect of large true observations ; *JASA*, 61, 1200-1204. M- 3
- SEBER, G. A. F. (1962) : The multi-sample single recapture census; *Biometrika*, 49, 339-350. W-14
- SEBER, G. A. F. (1965) : A note on multiple recapture census; *Biometrics*, 21, 249-259. W-14
- SEDRANSK, J. (1965a) : A double sampling scheme for analytical surveys; *JASA*, 60, 985-1004. M-11
- SEDRANSK, J. (1965b) : Analytical surveys with cluster sampling; *JRSS*, (B), 27, 264-278. M- 9
- SEELBINDER, B. M. (1953) : On Stein's two-stage sampling scheme; *AMS*, 24, 640-649. M- 4
- SEKHAR, C. C. and DEMING, W. E. (1949) : On a method of estimating birth and death rates and the extent of registration; *JASA*, 44, 101-115. P-13
- SEN, A. R. (1953a) : Recent advances in sampling with varying probabilities; *BCSA*, 5, 1-15. M- 6
- SEN, A. R. (1953b) : On the estimate of the variance in sampling with varying probabilities; *JISAS*, 5, 119-127. M- 6
- SEN, A. R., ANDERSON, R. L. and FINKNER, A. L. (1954) : A comparison of stratified two-stage sampling systems; *JASA*, 49, 539-558. A- 9
- SEN, A. R. (1955a) : On the selection of ' n ' primary sampling units from a stratum structure ($n \geq 2$); *AMS*, 26, 744-751. M- 6
- SEN, A. R. (1955b) : A simple design in sampling with varying probabilities; *JISAS*, 7, 57-69. M- 6

- SEN, A R and CHAKRABARTHY, R P (1964) Estimation of loss of crop from pests and diseases of tea from sample surveys *Biometrics*, 20, 492-504 C- 9
- SEN, P K. (1960) On the estimation of the population size by capture recapture methods, *BCSA*, 9, 93-110 W-14
- SEN, S B (1954) See Mahalanobis P C
- SENG, Y P (1949) Practical problems in sampling for social and demographic inquiries in undeveloped countries *Population Studies*, 3 170-191 P- 1
- SENG, Y P (1951) Historical survey of the development of sampling theories and practice *JRSS*, (A) 114 214-231 M- 2
- SENGUPTA J M, CHAKRAVARTI, I M, and SARKAR, D (1951a) Experimental survey for the estimation of cinchona yield, *BISI*, 33 (2), 313-330 C- 1
- SENGUPTA, J M (1951b) See Mahalanobis, P C (1951)
- SENGUPTA J M (1954) Some experiments with different types of area sampling for winter paddy in Gurdih, Bihar, 1945, *Sanhyā*, 13, 235-240 C- 8
- SENGUPTA J M (1955) See Masuyama, M
- SENGUPTA, J M and BHATTACHARYA, N (1958) Tables of random normal deviates *Sanhyā* 20, 249-286 M- 1
- SENGUPTA, J M (1963) A study of the field cost for the collection of household consumption data by an interview method *Contributions to Statistics*, 429-448 Presented to Professor P C Mahalanobis on the occasion of his 70th Birthday, Pergamon Press, London, and Statistical Publishing Society Calcutta. M- 1
- SENGUPTA J M (1964a) On perimeter bias in sample-cuts of small size, *Sanhyā* 26 (B) 53-68 C- 8
- SENGUPTA, J M, DUTTA N C and Roy, S (1964b) Experimental land utilization surveys in cadastrally unsurveyed area through direct plot to plot observation *Sanhyā* 26 (B) 69-88 C- 1
- SENGUPTA, J M (1965) On the use of motor vehicles in crop cutting surveys and sampling of road side plots only, *Sanhyā* 27, (B), 159-174 C-14
- SENGUPTA, J M (1966a) On the validity of fertility data collected through interviewers, *Sanhyā*, 28 (B), 256-268 P-13
- SENGUPTA, J M (1966b) Enumeration of fruit trees in land utilization surveys, *Sanhyā*, 28, (B), 269-292 C- 1
- SENGUPTA, J M. (1966c) Exploitation of fish in tanks and ponds, West Bengal, 1957-61, *Sanhyā*, 28, (B), 293-306 W- 1
- SETH, G R (1952) See Sukhatme, P V (1952a)
- SETH, G R (1961) On some aspects of sampling theory and practice, Presidential address to the Section of Statistics, *Proceedings of the Indian Science Congress*, 48th Session Roorkee M- 2
- SETH, G R and Rao, J N K (1964) On the comparison between simple random sampling with and without replacement, *Sanhyā*, 26, (A), 81-90 M- 3
- SETH, G R (1966a) On collapsing strata, *JISAS*, 18, 1-3 M- 7

- SETHI, G. R. (1966b): On estimates of variance of estimate of population total in varying probabilities; *JISAS*, 18, 52-56. M- 6
- SETHI, V. K. (1959): See Nanjamma, N. S. (1959a).
- SETHI, V. K. (1961): See Murthy, M. N.
- SETHI, V. K. (1962): Some consequences of an interpretation of varying probability sampling; *Sankhyā*, 24, (B), 215-222. M- 6
- SETHI, V. K. (1963): A note on optimum stratification of populations for estimating the population means; *AJS*, 5, 20-33. M- 7
- SETHI, V. K. (1965a): On optimum pairing of units; *Sankhyā*, 27, (B), 315-320. M- 5
- SETHI, V. K. (1965b): See Murthy, M. N.
- SETHI, V. K. (1966): See Hess, I.
- SHAFFER, J. D. (1954a): A plan for sampling a changing population over time; *JFE*, 36, 153-163. M-11
- SHAFFER, J. D. (1954b): See Kramer, R. C. (1954).
- SHALIGRAM, G. C. (1945): See Panse, V. G.
- SHALIGRAM, G. C., GOLUR, M. B. and GHOSH, M. N. (1961): On relative sampling of various regions of the field; *JISAS*, 18, 137-146. C- 1
- SHANKLEMAN, E. (1955): Measuring the readership of newspapers and magazines; *Appl. Stat.*, 4, 183-191. O- 1
- SHARP, H. and FELDT, A. (1959): Some factors in a probability sample survey of a metropolitan community; *ASR*, 24, 650-661. P-13
- SHARTLE, R. B. (1952): How scientific sampling controls accuracy in interviewing; *J. Accountancy*, 167- . E- 1
- SHAUL, J. R. H. and MYBURGH, C. A. L. (1948): A sample survey of the African population of Southern Rhodesia; *Population Studies*, 2, 339-353. P- 1
- SHAUL, J. R. H. and MYBURGH, C. A. L. (1949): Provisional results of the sample survey of the African population of Southern Rhodesia, 1948; *Population Studies*, 3, 274-285. P- 1
- SHAUL, J. R. H. (1952): Sampling surveys in Central Africa; *JASA*, 47, 239-253. M- 1
- SHAUL, J. R. H. (1957): An African agricultural sample survey in Central Africa in 1955; *BISI*, 35, (2), 141-150. C- 1
- SHEATSLEY, P. B. (1950): An analysis of interviewer characteristics and their relationship to performance; *IJOAR*, 4, 473-498. M- 1
- SHEATSLEY, P. B. (1951): An analysis of interviewer characteristics and their relationship to performance, Part II; *IJOAR*, 5, 79-94. M- 1
- SHERPARD, J. B. and SCHUTZ, H. H. (1937): Selection of areas for sample agricultural enumeration I; *JFE*, 19, 454-464. A- 1
- SHERWHART, W. A. (1931): Random sampling; *American Mathematical Monthly*, 38, 245-270. M- 1
- SHOWELL, M. (1951): How much stratification (on the basis of community size); *IJOAR*, 5, 229-240. O- 7
- SHUKLA, N. D. (1966): See Pathak, P. K. (1966b).

- SIEBEL, H S (1947) An experimental and theoretical investigation of bias error in mine sampling with special reference to narrow gold reefs, *Bulletin of Institute of Mining and Metallurgy*, 26-41, 55-59 E- 1
- SILBER, J (1948) Multiple sampling for variables, *AMS*, 19, 246-257 M-11
- SILBERMAN, L (1954) The social survey of Port Louis (Mauritius), *RISI*, 22 (1 3), 85-94 L- 1
- SILCOCK, H (1952) Estimating by sample the size and age sex structure of population, *Population Studies*, 6, 55-68 P- 1
- SILCOCK, H (1954) Precision in population estimates, *Population Studies*, 8, 142-147 P-13
- SILVEY, R J E (1944) Methods of listener research employed by the B B C; *JRSS*, (A), 107, 190-230 O- 1
- SIMMONS, W R (1946) See Deming, W. E
- SIMMONS, W R (1949) See Politz, A N
- SIMMONS, W R (1953) Prelisting in marketing and media surveys, *J Marketing*, 18, 6-17 O- 1
- SIMMONS, W R (1954) A plan to account for 'not at home' by combining weighting and call backs, *J Marketing*, 19, 42-53 S-13
- SIMPSON, G D (1940) See King, A J (1940b)
- SIMPSON, H (1943) On a theorem concerning sampling, *JRSS*, (A), 106, 266-267 M- 1
- SINGH, B D and SINGH, D (1965a) Some remarks on double sampling for stratification, *Biometrika*, 52, 587-590 M- 7
- SINGH, B D (1965b) See Singh, D (1965a)
- SINGH, D (1951) See Iyer, P V K.
- SINGH, D (1954) On efficiency of sampling with varying probabilities without replacement, *JISAS*, 6, 48-57 M- 6
- SINGH D (1956) On efficiency of cluster sampling *JISAS*, 8, 45-55
- SINGH, D and SAXENA, P N (1957) On recent developments and prospects in sampling and surveys with special reference to work done in India, *BISI*, 35, (2), 163-178 M- 2
- SINGH, D (1958) Estimates of variance components in finite population, *JISAS*, 10, 1-15 M- 9
- SINGH, D (1965a) See Singh, B D (1965a)
- SINGH, D and SINGH, B D (1965b) Double sampling for stratification on successive occasions, *JASA*, 60, 784-792 M- 7
- SINGH, M P (1965) On the estimation of ratio and product of population parameters, *Sankhya*, 27, (B), 321-328 M-10
- SIRKEN, M G (1950a) See Birnbaum, Z W (1950a)
- SIRKEN, M G. (1950b) See Birnbaum Z W (1950b)
- SIRKEN, M G (1965) See Birnbaum, Z W
- SKELLAM, J G (1949) The distribution of the moment statistics of samples drawn without replacement from a finite population, *JRSS*, (B), 11, 291-296 M- 3

- SLONIM, M. J. (1957) : Sampling in a nutshell; *JASA*, 52, 143-161. M- 1
- SMITH, B. B. (1938) : See Kendall, M. G.
- SMITH, B. B. (1939) : See Kendall, M. G.
- SMITH, D. M. K. (1948) : See Hochstim, J. R.
- SMITH, E. D. (1939) : Market sampling; *J. Marketing*, 4, 45-50. O- 1
- SMITH, H. F. (1938) : An empirical law describing heterogeneity in the yields of agricultural crops; *JAS*, 28, 1-23. C- 8
- SMITH, H. F. (1947) : Standard errors of means in sampling surveys with two-stage sampling; *JRSS*, (A), 110, 257-259. M- 9
- SMITH, H. F. (1948) : A sampling survey of tappings on small holdings (1939-40); *J. Rubber Research Institute, Malaya*, 12, 70-125. C- 1
- SNEDECOR, G. W. (1939) : Design of sampling experiments in the social Sciences; *JFE*, 21, 346-355. S- 1
- SNEDECOR, G. W. and KING, A. J. (1942) : Recent developments in sampling for agricultural statistics; *JASA*, 37, 95-102. C- 2
- SNEDECOR, G. W. (1948) : On the design of sampling investigations; *Amer. Stat.*, 2, (6), 6-9. M- 1
- SOBEL, M. and HUYETT, M. J. (1958) : Non-parametric definition of the representativeness of a sample with tables; *The Bell System Technical Journal*, 37, 135-162. M- 1
- SOBEL, M. G. (1959) : Panel mortality and panel bias; *JASA*, 54, 52-68. M-13
- SOM, R. K. (1958) : On sampling design in opinion and marketing research; *POQ*, 32, 564-566. O-12
- SOM, R. K. (1959a) : Self-weighting sample design with an equal number of ultimate stage units in each of the selected penultimate stage units; *BCSA*, 8, 59-66. M-12
- SOM, R. K. and DAS, N. C. (1959b) : On recall-lapse in infant death recording; *Sankhyā*, 21, 205-208. P-13
- SOM, R. K. (1960) : On response errors in economic classification of population; *Economic Affairs*, 5, 543-555. P-13
- SOMMERVILLE, P. N. (1954) : Some problems of optimum sampling; *Biometrika*, 41, 420-429. M- 1
- SOMMERVILLE, P. N. (1957) : Optimum sampling in binomial populations; *JASA*, 52, 491-502. M- 1
- SONQUIST, J. A. (1963) : See Morgan, J. N.
- SRIKANTAN, K. S. (1963) : A note on interpenetrating sub-samples of unequal sizes; *Sankhyā*, 25, (B), 345-350. M- 1
- SRINIVASAN, P. S. (1943) : Studies on the estimation of growth and yield of Jowar by sampling; *Indian J. Agricultural Science*, 13, 399-412. C- 1
- STANTON, F. N. (1941) : Problems of sampling in market research; *J. Consulting Psychology*, 154-163. O- 1
- STARCK, D. (1932) : Factors in the reliability of samples; *JASA*, (Supplement), 27, (177A), 190-201. M- 1
- STAUDT, E. P. (1943) : See Eckler, A. R.

- STEIV, C (1945) A two sample test for a linear hypothesis whose power is independent of the variance, *AMS*, 16, 243-258 M- 4
- STEINBERG, J (1955) See Hansen, M H
- STEINBERG, J (1956) See Hansen, M H
- STEIVER, P O (1951) A source of bias in one of the samples of the 1950 census, *JASA*, 46, 110-113 P-13
- STEPHAN F F (1934) Sampling errors and interpretation of social data ordered in time and space, *JASA (Supplement)* 29, 163-166 S- 1
- STEPHAN F F (1936) Practical problems of sampling procedure, *ASR*, 41, 567-580 S- 1
- STEPHAN F F (1939) Representative sampling in large scale surveys, *JASA*, 34, 343-352 M- 1
- STEPHAN F F, DEMING, W E and HANSEN, M H (1940a) The sampling procedure of the 1940 population census *JASA*, 35, 615-630 P- 1
- STEPHAN, F F (1940b) See Deming, W E (1940b)
- STEPHAN F F (1941a) See Deming, W E (1941a)
- STEPHAN F F (1941b) Stratification in representative sampling *J. Marketing* 6, 38-46 O- 7
- STEPHAN F F and McCARTHY, P J (1947) Sampling opinions, attitudes and consumer wants, *Amer Stat*, 1, (3), 6-7 O- 1
- STEPHAN, F F (1948a) Sampling for old age Research *Social Science Research Council Bulletin*, 5 S- 1
- STEPHAN, F F (1948b) History of the uses of modern sampling procedures, *JASA*, 43, 12-39 M- 2
- STEPHAN, F F (1950) Sampling, *Amer J Soc*, 55, 371-375 M- 1
- STEPHAN, F F (1955) See El Badry M A
- STEPHAN, F F (1957) Advances in survey methods and measurement techniques, *POQ*, 21, 79-90 O- 2
- STEPHAN, F F (1963) Purposive and stratified random sampling today, *Estadística* 21, 691-695 M- 7
- STEVENS W L (1952) Samples with the same number in each stratum, *Biometrika*, 39, 414-417 M- 7
- STEVENS, W L (1955) Estimation of the Brazilian coffee harvest by sampling survey, *JASA*, 50, 775-787 C- 1
- STEVENS, W L (1958) Sampling without replacement with probability proportional to size, *JRSS, (B)* 20, 393-397 M- 6
- STOCK, J S and FRANKEL, L R (1939) The allocation of sampling among several strata, *AMS*, 10, 288-293 M- 7
- STOCK, J S and FRANKEL, L R (1942) On the sample survey of unemployment, *JASA*, 37, 77-80 S- 1
- STOCK, J S and HOCHSTIM, J R (1947) Commercial uses of sampling, *Proceedings of International Statistical Conference*, 3 (A), 129-143, *JASA*, 43, (1948), 509-522 O- 1
- STOCK, J S and HOCHSTIM, J R (1951) A method of measuring interviewer variability, *POQ*, 15, 322-334 M-13

- STONE, J. R. N., UTTING, J. E. G. and DURBIN, J. (1950) : The use of sampling methods in national income statistics and social accounting; *RISI*, 18, (1-2), 21-44. S- 1
- STRAND, N. V. and JESSEN, R. J. (1943) : Some investigations on the stability of the township as the unit for sampling Iowa agriculture; *Iowa Agricultural Experimental Research Station Bulletin*, No. 315. A- 1
- STRAW, K. H. (1955) : See Hill, T. P.
- STUART, A. (1951) : See Durbin, J.
- STUART, A. (1953) : See Moser, C. A.
- STUART, A. (1951a) : A simple presentation of optimum sampling results; *JRSS*, (B), 16, 239-241. M- 1
- STUART, A. (1954b) : See Durbin, J. (1954b).
- STUART, A. (1963) : Some remarks on sampling with unequal probabilities; *RISI*, 40, (2), 773-779. M- 6
- STUART, A. (1964) : Multi-stage sampling with preliminary random stratification of first stage units; *RISI*, 32, 193-201. M- 6
- STYCOS, J. M. (1956) : Sample survey: Its uses and problems; *United Nations Population Bulletin*, No. 5. M- 1
- STYCOS, J. M. (1961) : Survey research and population control in Latin America; *POQ*, 28, 367-372. P- 1
- SUBRAHMANYA, M. T. (1965) : On generalized estimators in sampling from finite populations; *Metrika*, 9, 234-239. M- 1
- SUBRAHMANYA, M. T. (1966) : A note on a biased estimator in sampling with probability proportional to size with replacement; *AMS*, 37, 1045-1047. M- 6
- SUCHMAN, E. A. (1962) : An analysis of 'bias' in survey research; *POQ*, 26, 102-110. M-13
- SUDMAN, S. (1966) : Probability sampling with quotas; *JASA*, 61, 749-771. M-14
- SUKHATME, B. V. and KOHARAL, R. S. (1950) : A contribution to double sampling; *JISAS*, 9, 128-144. M-11
- SUKHATME, B. V. (1962a) : Generalized Hartley-Ross unbiased ratio-type estimator; *Nature*, 196, (4860), 1238. M-10
- SUKHATME, B. V. (1962b) : Some ratio-type estimators in two-phase sampling; *JASA*, 57, 628-632. M-10
- SUKHATME, B. V. (1965a) : See Avadhani, M. S.
- SUKHATME, B. V. (1965b) : See Goswami, J. N.
- SUKHATME, B. V. (1966) : See Avadhani, M. S.
- SUKHATME, P. V. (1935) : Contributions to the theory of the representative method; *JRSS*, (Supplement), 2, 253-268. M- 7
- SUKHATME, P. V. (1945) : Random sampling for estimating rice yields in Madras Province; *IJAS*, 15, 308-318. C- 1
- SUKHATME, P. V. (1946a) : Bias in the use of small-size plots in sample surveys for yield; *Current Science*, 15, 119-120; *Nature*, 157, 630. C- 6
- SUKHATME, P. V. (1946b) : Size of sampling unit in yield surveys; *Nature*, 158, 345. C- 8

- SUKHATME, P V (1947a) Use of small size plots in yield surveys, *Nature*, 160, 542 C- 8
- SUKHATME, P V (1947b) The problem of plot size in large scale yield surveys, *JASA*, 42, 297-310, 460 C- 8
- SUKHATME, P V (1947c) Report on Scheme for Crop Cutting Experiments for Comparing Large and Small size Plots Moradabad Dist, United Provinces, 1944-45, Superintendent, Printing and Stationery, United Provinces, India, 1-15 C- 8
- SUKHATME, P V (1947d) Report on Random Sample Survey for Estimating the Outturn of Paddy in Central Province and Berar, 1945-46, Imperial Council of Agricultural Research, New Delhi C- 1
- SUKHATME, P V (1947e) Report on a Random Sample Survey for Estimating the Outturn of Paddy in Madras, 1945-46, Imperial Council of Agricultural Research, New Delhi C- 1
- SUKHATME P V and KOSHAL, R S (1947f) Report on Scheme for Crop Cutting Experimental Survey on Paddy in the Bombay Province, 1945-46, Imperial Council of Agricultural Research New Delhi C- 1
- SUKHATME, P V (1948a) Experimental Survey for Estimating the Outturn of Wheat, Punjab, 1943-44, Government of India Press, Calcutta, 1-68 C- 1
- SUKHATME, P V (1948b) See Panse V G
- SUKHATME, P V (1950a) Sample surveys in agriculture, Presidential Address to the Statistics Section, *Proceedings of the Indian Science Congress*, 37th Session, Poona C- 1
- SUKHATME P V (1950b) Efficiency of sub sampling designs in yield surveys, *JISAS*, 2 212-228 C- 9
- SUKHATME, P V and PANSE, V G (1951a) Crop surveys in India II, *JISAS*, 3, 97-168 C- 1
- SUKHATME, P V and KISHEN, K (1951b) Assessment of the accuracy of patwaries area records, *Agricultural and Animal Husbandry* 1, (9), 36-47 C-13
- SUKHATME, P V and SETH, G R (1952a) Non sampling errors in surveys, *JISAS* 4 5-41 M-13
- SUKHATME, P V and NARAIN, R D (1952b) Sampling with replacement, *JISAS*, 4, 42-49 M- 9
- SUKHATME, P V (1952c) Random sampling for improvement of agricultural statistics, *MBAES*, 1, 2-6 C- 1
- SUKHATME, P V (1953) Measurement of observational errors in surveys, *RISI*, 20, (2-3), 121-134 M-13
- SUKHATME, P V (1958a) The 1960 World Census of Agriculture, *BISI*, 36, (4), 239-250 A- 1
- SUKHATME, P V, PANSE, V G and SASTRY, K V R (1958b) Sampling techniques for estimating the catch of seafish in India, *Biometrics*, 14, 78-96 W- 5
- SUKHATME, P V (1959) Major developments in the theory and applications of sampling during the last twenty five years, *Estadistica*, 17, 652-679 M- 2

BIBLIOGRAPHY

647

- SUNDRUM, R. M. (1953) : A method of systematic sampling based on order properties; *Biometrika*, 40, 452-456. M- 5
- SWAIN, A. K. P. C. (1964) : The use of systematic sampling in ratio estimate; *JISA*, 2, 160-164. M-10
- SZAMEITAT, K. and SCHAFER, K. A. (1963) : Imperfect frames in statistics and the consequences for their use in sampling; *BISI*, 40, (2), 517-538. M- 1
- TAEUBER, C. (1944) *See* Tolley, H. R.
- TAGA, Y. (1953) : On optimum balancing between sample size and number of strata in sub-sampling; *AISM*, 4, 95-102. M- 7
- TAKEUCHI, K. (1961) : Some remarks about unbiased estimation in sampling from finite populations; *JUSE*, 8, 35-38. M- 1
- TALACKO, J. (1959) : Special methods of inventory by sampling if the population sets have approximately negative exponential distribution; *Trab. Est.*, 10, 19-29. E- 1
- TANNER, J. C. (1957) : The sampling of road traffic; *Appl. Stat.*, 6, 161-170. E- 1
- TANNER, J. C. (1958) : *See* Chandler, K. N.
- TAYLOR, J. (1951) : The estimation of fruit size of cherries by sampling methods; *Report of East Malling Research Station for 1950*, 93-99. C- 1
- TAYLOR, W. B. and CLEMENT, D. V. P. (1956) : The New Zealand agricultural sample survey; *JRSS*, (A), 119, 409-424. A- 1
- TEICHROEW, D. (1965) : A history of distribution sampling prior to the era of the computer and its relevance to simulation; *JASA*, 60, 27-49. M- 2
- TEPPING, B. J., HURWITZ, W. N. and DEMING, W. E. (1943) : On the efficiency of deep stratification in block sampling; *JASA*, 38, 93-100. M- 7
- TEPPING, B. J. and WITTRICH, W. J. (1960) : Sample surveys for legal evidence; *J. Marketing*, 25, (2), 57-59. M- 1
- THOMAS, G. (1944) *See* Box, K.
- THOMPSON, A. P. (1945) : A sampling approach to New-Zealand timber cruising problems; *New-Zealand J. Forestry*, 5, 103-117. F- 1
- THOMPSON, D. J. (1952) : *See* Horvitz, D. G.
- THOMPSON, W. A. (JR.) and ENDRISS, J. (1961) : The required sample size when estimating variances; *Amer. Stat.*, 15, (3), 22-23. M- 4
- TIKKIWAL, B. D. (1953) : Optimum allocation in successive sampling; *JISAS*, 5, 100-102. M-11
- TIKKIWAL, B. D. (1956) : A further contribution to univariate sampling on successive occasions; *JISAS*, 8, 84-90. M-11
- TIKKIWAL, B. D. (1960) : On the theory of classical regression and double sampling estimation; *JRSS*, (B), 22, 131-138. M-11
- TIKKIWAL, B. D. (1964) : A note on two-stage sampling on successive occasions; *Sankhya*, 26, (4), 97-100. M-11
- TIN, M. (1965) : Comparison of some ratio estimators; *JASA*, 60, 294-307. M-10
- TOLLEY, H. R. (1929) : Economic data from the sampling point of view; *JASA*, 24, (*Supplement*), 69-72. S- 1

- TOLLEY, H R and TAEUBER, C (1944) War time developments in agricultural statistics, *JASA*, 39, 411-427 C- 2
- TOMASSOV, R F (1961) Bias in estimates of the US non white population as indicated by trends in death rates, *JASA*, 56, 44-51 P-13
- TOSTLEBE, A S (1945) Estimate of series E bond purchases by farmers, *JASA*, 40, 317-329 S- 1
- TROLDAHL, V C (1963) *See Carter, R E (Jr)*
- TRUEBLOOD, R M and CYERT, R M (1954) Statistical sampling applied to ageing of accounts receivable, *J Accountancy*, 293- E- 1
- TRUEBLOOD, R M (1957) *See Cyert, R M*
- TRUESDELL, L E (1941) New features of the 1940 population census, *JASA*, 36, 361-368 P- 1
- TSCHUPROW, A A (1923) On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, *Metron*, 2, 461-493 and 646-680 M- 1
- TUKEY, J W (1950) Some sampling simplified, *JASA*, 45, 501-519 M- 1
- TUKEY, J W (1953) *See Cochran, W G (1953a)*
- TUKEY, J W (1954a) Unsolved problems of experimental statistics, *JASA*, 49, 706-731 M- 1
- TUKEY, J W (1954b) *See Cochran, W G*
- TULSE, R (1957) Sampling for variables with a very skew distribution, *Appl Stat*, 6, 40-44, *Estadistica*, 15, 399-404 M- 1
- TURNER, R (1961) Inter week variation in expenditure recorded during a two week survey of family expenditure, *Appl Stat*, 10, 136-146 L-13
- TSYKIBAYASHI, S (1964) Stratified sampling and systematic sampling with supplementary data, *JUSE*, 11, (4), 137-146 M- 7
- UNDY G C (1962) *See Brewer, K R W*
- UNITED NATIONS (1947-51) *Reports of the Sub Commission on Statistical Sampling to the Statistical Commission*, E/CN. 3/37, 52, 83, 114, 140, New York, reprinted in *Sankhya*, 8, (1948), 393-402, 9, (1949), 377-391, 10, (1950), 128-158, 11, (1951), 63-95, 12, (1952) 165-204 M- 1
- UNITED NATIONS (1949a) Population census methods, *Population Studies*, No 4, New York P- 1
- UNITED NATIONS (1949b) Recommendations concerning the preparation of reports on sampling surveys, *Statistical Papers*, Series, C, No 1, New York, reprinted in *Sankhya*, 9, 392-398 M- 1
- UNITED NATIONS (1952) Sample surveys of current interest, *Statistical Papers*, Series C, 2, New York M- 1
- UNITED NATIONS (1964a) Recommendations for the preparation of sample survey reports, *Statistical Papers*, Series C, No 1, Rev 2, New York M- 1
- UNITED NATIONS (1964b) Handbook of household surveys, *Studies in Methods*, Series F, No 10, New York L- 1
- UNITED NATIONS (1966) Report of the seminar on sampling methods, Tokyo, 1965, *Statistical Papers*, Series M, No 42 M- 1

- UNITED STATES OF AMERICA (1951): Standards for statistical surveys; *RISI*, 19, (3), 241-245.
- US BUREAU OF AGRICULTURAL ECONOMICS (1936): *Proceedings of Conference on Statistical Methods of Sampling Agricultural Data*; Iowa State College, Ames, Iowa.
- US BUREAU OF AGRICULTURAL ECONOMICS (1953a): Sampling methods for agricultural estimating and forecasting and elements to be considered in their adoption; *Estadistica*, 11, 72-90.
- US BUREAU OF AGRICULTURAL ECONOMICS (1953b): Use of check data in crop estimation; *Estadistica*, 11, 91-102.
- US BUREAU OF THE CENSUS (1945): *Notes on Precision of Sample Estimates*; US Government Printing Office, Washington, D.C.
- US BUREAU OF THE CENSUS (1947): *A Chapter in Population Sampling*; U.S. Government Printing Office, Washington, D.C.
- US BUREAU OF THE CENSUS (1953): The sample survey of retail stores—a report on methodology; *Technical Paper*, No. 1, U.S. Department of Commerce, Washington, D.C.
- US BUREAU OF THE CENSUS (1954): Concepts and methods used in the current labour force statistics prepared by the U.S. Bureau of the Census; *Estadistica*, 13, 280-290.
- US BUREAU OF THE CENSUS (1956): Expansion of the current population survey sample 1956; *Current Population Reports*, P. 23, (3), Washington, D.C.
- US BUREAU OF THE CENSUS (1957): Description of the sample design of the U.S. National Health Survey; *Estadistica*, 15, 428-431.
- US BUREAU OF THE CENSUS (1959): Accuracy of census statistics with and without sampling; *Technical Paper*, No. 2, U.S. Department of Commerce, Washington, D.C.
- US BUREAU OF THE CENSUS (1960): The Post-enumeration survey 1950, *Technical Paper*, No. 4, U.S. Department of Commerce, Washington, D.C.
- US BUREAU OF THE CENSUS (1963a): The Current population re-interview survey, some notes and discussions; *Technical Paper*, No. 6, U.S. Department of Commerce, Washington, D.C.
- US BUREAU OF THE CENSUS (1963b): The Current population survey—a report on methodology, *Technical Paper*, No. 7, U.S. Department of Commerce, Washington, D.C.
- US BUREAU OF THE CENSUS (1965a): Response errors in collection of expenditures data by household interviews: An experimental study; *Technical Paper* No. 11, Department of Commerce, Washington, D.C.
- US BUREAU OF THE CENSUS (1965b): Sampling applications in censuses of population and housing; *Technical Paper*, No. 13, U.S. Department of Commerce, Washington, D.C.
- US PUBLIC HEALTH SERVICE (1958): National Health Survey—the statistical design of the household-interview survey; *Health Statistics Series A-2*.

M- 1

C- 1

C- 1

C- 1

M- 1

P- 1

E- 1

S- 1

S- 1

S- 1

P-13

P-13

S-13

S- 1

L-13

P- 1

S- 1

- UTTING, J E G (1950) *See* Stone, J R N.
- UTTING, J E G and COLE, D (1953) Sample surveys for the social accounts of the household sector, *Bulletin of Oxford University Institute of Statistics*, 15, 1-23 S- 1
- UTTING, J E G and COLE, D (1954) Sampling for social accounts—some aspects of the Cambridgeshire survey, *BISI*, 34, (2), 301-328 S- 1
- UTTING, J E G and COLE, D (1956) Estimating expenditure, saving and income from household budgets *JRSS*, (A), 119, 371-392 S- 1
- VANCE, L L (1960) A review of development in statistical sampling for accountants, *Accounting Review*, 19-32 E- 2
- VAN DEN BERG, N and VERBURGH, C (1956) The accuracy of forecasts in the South African business opinion surveys, *South African J. Economics*, 24 37-62 O-13
- VANDOME, P (1957) *See* Klein L R
- VERBURGH, C (1956) *See* Van Den Berg
- VICKERY, R E (1958) Recent experiences with area sampling for agricultural statistics *Biometrics*, 14 250-256 C- 2
- VIJAYAN, K (1966) On Horvitz Thompson and Des Raj estimators, *Sankhya*, 28, (A) 87-92 M- 6
- VINCENT, W H (1959) A farm panel as a source of income and expenditure data *AER*, 11 97-101 A- 1
- VITHYASAI C (1961) *See* Robson D S
- VOIGHT, R B and KEESBERG, M (1952) Some principles of processing census and survey data *JASA*, 47, 222-231 M-1
- VOS, J W E (1964) Sampling in space and plane *RISI*, 32 223-241 M-14
- WADLEY, F M (1945) An application of the Poisson series to some problem of enumerations *JASA*, 40 85-92 W- 1
- WAKSBERG, J (1964) *See* Neter, J
- WALD, A (1942) Setting of tolerance limits when the sample is large, *AMS*, 13 389-399 M-1
- WALLACE, D (1947) Mail questionnaires can produce good sample of homogeneous groups *J. Marketing*, 12 53-62 O- 1
- WALLACE, D (1954) A case for-and against-mail questionnaires *POQ*, 18 40-52 O- 1
- WALLIS, W A (1949) Statistics of the Kinsey Report, *JASA*, 44 463-484 S-13
- WARNER, L (1939) The reliability of public opinion surveys *POQ*, 3, 376-390 O-13
- WARNER, L (1947) Estimating the character of unsampled segments of a universe, *J. Marketing*, 12, 186-192 M- 1
- WARNER, S L (1965) Randomized response—A survey technique for eliminating evasive answer bias, *JASA*, 60, 63-69 M-13
- WATSON, A N (1946) Respondent pre selection 1946, *Technical Series*, No 1, Research Department, Curtis Publishing Co, Philadelphia M- 1

BIBLIOGRAPHY

651

- WATSON, A. N. (1947) : Respondent pre-selection within sample areas, No. 2; *Technical Series*, No. 2, Research Department, Curtis Publishing Co., Philadelphia.
- WATSON, D. J. (1934) : See Yates, F. M- 1
- WATSON, D. J. (1936) : See Cochran, W. G. (1936).
- WATSON, D. J. (1937) : The estimation of leaf areas; *JAS*, 27, 474-483. C- 1
- WEAVER, C. L. (1947) : See Grubbs, F. E.
- WEBB, J. N., NORTHRUP, M. S. and PAYNE, S. L. (1943) : Practical applications of theoretical sampling methods; *JASA*, 38, 69-77. S- 1
- WEBER, C. R. (1946) : See Houseman, E. E.
- WEIBULL, M. (1950) : The distribution of the t and z variables in the case of stratified sample with individuals taken from normal parent populations with varying means; *Skand. Akt.*, 33, 137-167. M- 7
- WEIBULL, M. (1951) : The regression problem involving non-random variates in the case of a stratified sample from normal parent populations with varying regression coefficients; *Skand. Akt.*, 34, 53-71. M- 11
- WEIBULL, M. (1953) : The distribution of t and F statistics and of correlation and regression coefficients in stratified samples from normal population with different means; *Skand. Akt.*, 36, (Supplement), 1-106. M- 7
- WEIBULL, M. (1959) : Moments of the difference between means in two samples from a finite population; *Skand. Akt.*, 42, 36-60. M- 3
- WELLS, D. W. and DAMES, J. (1962) : Hidden errors in survey data; *J. Marketing*, 26, (4), 50-54. M-13
- WEST, Q. M. (1951) : The results of applying a simple random sample process to farm management data; *Report of Agricultural Experiment Station*, Cornell University, Ithaca, New York. A- 1
- WEST, Q. M. (1952) : Some alternative sampling techniques in the measurement of farm-business characteristics; *JFE*, 34, 982-987. A-1
- WESTFALL, R. (1955) : See Boyd, H. W. (Jr.).
- WESTOFF, C. F., POTTER, R. G. (Jr.) and SACI, P. C. (1961) : Some estimates of the reliability of survey data on family planning; *Population Studies*, 15, 52-69. P-13
- WILCOX, W. W. (1940) : See Goodsell, W. D.
- WILKS, S. S. (1940) : Representative sampling and poll reliability; *POQ*, 4, 261-269. O- 1
- WILKS, S. S. (1960) : A two-stage scheme for sampling without replacement; *BISI*, 37, (2), 241-248. M- 9
- WILLIAMS, E. J. (1958) : See Hatheway, W. H. E- 1
- WILLIAMS, N. (1953) : The use of samples in auditing; *Appl. Stat.*, 2, 180-183.
- WILLIAMS, R. (1950) : Probability sampling in the field—A case study; *POQ*, 14, 316-330. O- 1
- WILLIAMS, R. (1956) : See Franzen, R.
- WILLIAMS, R. M. (1956) : Variance of the mean of systematic samples; *Biometrika*, 43, 137-148. N- 5

- WILLIAMS W H (1961) Generating unbiased ratio and regression estimators
Biometrika 48, 267-274 M-11
- WILLIAMS, W H (1962a) On two methods of unbiased estimation with auxiliary variates *JASA*, 57 184-186 M- 1
- WILLIAMS, W H (1962b) On the variance of an estimator with post stratification *JASA*, 57, 622-627 M- 7
- WILLIAMS W H (1963) The precision of some unbiased regression estimators *Biometrika*, 50 352-361 M-11
- WILLIAMS W H (1964) Sample selection and the choice of estimators in two way stratified population *JASA*, 59 1054-1062 M- 7
- WILM H G (1944) See Costello, D F
- WILM H G (1946) The design and analysis of methods for sampling micro climatic factors *JASA*, 41, 221-232 M- 1
- WILSON, C P (1951) The sample survey as a research technique in marketing livestock and livestock products *JFE*, 33 1013-1018 A- 1
- WILSON E C (1950) Adapting probability sampling to Western Europe, *POQ* 14 215-223 O- 1
- WINDLE C (1959) The accuracy of census literacy statistics in Iran, 1959, *JASA* 54 578-581 P-13
- WISHART J and CLAPHAM A R (1929) A study in sampling technique the effect of artificial fertilisers on the yield of potatoes *JAS*, 19 600-618 C- 1
- WISHART, J (1938) See Irwin, J O
- WISHART J (1952) Moment coefficients of the *L* statistics in samples from a finite population *Biometrika* 39 1-13
- WITTREICH W J (1960) See Tepping, B J
- WOODBURY R M (1940) Methods of family living studies *International Labour Office Series N, Statistics* No 23 L- 1
- WOODRUFF E S (1963) The use of rotating samples in the Census Bureau's monthly surveys, *JASA*, 58 454-468 M-11
- WOOTTON T J (1933) Common errors in sampling *Social Forests*, 12, 521-525 M-13
- WOOLSEY, T D (1959) See Nisselson H
- WORKING H (1946) Note on sampling probabilities, *JASA*, 41, 238-239 M- 1
- WORMBLEIGHTON, R (1960) A useful generalization of Stein's two sample procedure, *AMS*, 31, 217-221 M- 4
- YAMAMOTO, S (1955) On the theory of sampling with probability proportional to given values, *AISM*, 7, 23-38 M- 6
- YANKEY, D, ROBERTS, B J and GRIFFITHS, W (1965) Husbands vs wives' responses to a fertility survey *Population Studies* 19 29-43 P-13
- YAO J S (1964) See Kudo, A
- YATES, F and WATSON, D J (1934) Observer's bias in sampling observations on wheat, *Empire Journal of Experimental Agriculture*, 2, 174-177 C-13
- YATES F (1935a) Some examples of biased sampling *Annals of Eugenics*, 6 202-213 M-13

- YATES, F. and ZACOPANAY, I. (1935b) : The estimation of efficiency of sampling with special reference to sampling for yield in cereal experiments; *JAS*, 25, 545-577. C- 1
- YATES, F. (1936) : Crop estimation and forecasting; Indications of the sampling observations on wheat; *J. Ministry of Agriculture*, 43, 156-162. C- 1
- YATES, F. and FINNEY, D. J. (1942) : Statistical problems in field sampling for wireworms; *AAB*, 29, 156-167. W- 1
- YATES, F. (1943) : Methods and purposes of agricultural surveys; *Journal of Royal Society for Arts*, 91, 367-379. C- 1
- YATES, F. (1946) : A review of recent statistical developments in sampling and sampling surveys; *JRSS*, 109, 12-43. M- 2
- YATES, F. (1947) : The influence of agricultural research statistics on the development of sampling theory; *Proceedings of International Statistical Conference*, 3, (A), 27-39. C- 1
- YATES, F. (1948) : Systematic sampling; *Philosophical Transactions of Royal Society*, (A), 241, 345-377. M- 5
- YATES, F. (1950) : Agriculture, sampling and operational research; *BISI*, 32, (2), 220-227. A- 1
- YATES, F. (1951) : Crop prediction in England; *BISI*, 33, (2), 295-312. C- 1
- YATES, F. and GRUNDY, P. M. (1953a) : Selections without replacement from within strata with probability proportional to size; *JRSS*, (B), 15, 253-261. M- 6
- YATES, F. (1953b) : The work of the United Nations Sub-commission on Statistical sampling; *Sankhyā*, 12, 305-306. M- 1
- YOUNG, D. H. (1961) : Quota fulfilment using unrestricted random sampling; *Biometrika*, 48, 333-342. M- 7
- YULE, G. U. (1938) : A test of Tippett's random sampling numbers; *JRSS*, (A), 101, 167-172. M- 1
- ZACOPANAY, I. (1935) : See Yates, F. (1935b).
- ZARKOVICH, S. S. (1954) : Sampling control of literacy data; *JASA*, 49, 510-519. S-13
- ZARKOVICH, S. S. (1955) : Sampling methods in the Yugoslav 1953 census of population; *JASA*, 50, 720-737. P-13
- ZARKOVICH, S. S. (1956a) : An illustration of some characteristic situations in the application of the difference estimate; *RISI*, 24, 52-63. C-11
- ZARKOVICH, S. S. (1956b) : Note on the history of sampling methods in Russia; *JRSS*, (A), 119, 336-338. M- 2
- ZARKOVICH, S. S. (1956c) : Some remarks on coverage checks in population censuses; *Population Studies*, 9, 271-275. P-13
- ZARKOVICH, S. S. (1957) : Sampling methods and censuses; *MBAES*, 6, 1-9. M- 1
- ZARKOVICH, S. S. (1958a) : On some consequences of the heterogeneity of units in sampling with varying probabilities and replacement; *Statistical Review*, Belgrade, 8, 45-47. M- 6

- ZARKOVICH, S S (1958b): On some problems of sampling work in under-developed countries, *Statistical Review*, Belgrade, 9, 1-11; *BISI*, 37, (2), 249-261 M- 1
- ZARKOVICH, S S (1960) · On the efficiency of sampling with varying probabilities and the selection of units with replacement, *Metrika*, 3, 53-59 M- 8
- ZARKOVICH, S S (1962) A supplement to "Note on the history of sampling methods in Russia", *JRSS*, (A), 125, 580-582 M- 2
- ZEISEL, H (1949) See Ford, R N
- ZINGER, A (1964) Systematic sampling in forestry, *Biometrika*, 50, 553-564 F- 5
- ZUBRZYCKI, S (1958) Remarks on random, stratified and systematic sampling in a plane, *Colloquium Mathematicum*, 6, 251-264 M-14
- ZUBRZYCKI, S (1961) See Dalenius, T

LIST OF JOURNALS WITH THEIR ABBREVIATIONS

AER	Agricultural Economics Research
AJPH	American Journal of Public Health
ASR	American Sociological Review
Amer Stat	American Statistician
AAB	Annals of Applied Biology
AISM	Annals of Institute of Statistical Mathematics
AMS	Annals of Mathematical Statistics
Appl Stat	Applied Statistics
AJS	Australian Journal of Statistics
BCSA	Bulletin of Calcutta Statistical Association
BISI	Bulletin of International Statistical Institute
BMS	Bulletin of Mathematical Statistics
Inc Stat	Incorporated Statistician
IJOAR	International Journal of Opinion and Attitude Research
JAR	Journal of Agricultural Research
JAS	Journal of Agricultural Science
JASA	Journal of American Statistical Association
JFE	Journal of Farm Economics
JISAS	Journal of Indian Society of Agricultural Statistics
JISA	Journal of Indian Statistical Association
JRSS	Journal of Royal Statistical Society
MBAES	Monthly Bulletin of Agriculture and Economic Statistics
POQ	Public Opinion Quarterly
JUSE	Report on Statistics of the Applied Research Union of Japanese Engineers and Scientists
RISI	Review of International Statistical Institute
Skand Akt	Skandinavisk Aktuaritidskrift
Trab Est	Trabajos de Estadistica

BOOKS

- ACKOFF, R. L. (1953) : *Design of Social Research*; University of Chicago Press, Illinois.
- BACKSTRÖM, C. H. and HURSH, G. D. (1963) : *Survey Research*; Northwestern University Press, Evanston.
- CANTRIL, H. (1945) : *Gauging Public Opinion*; University Press, Princeton.
- COCHRAN, W. G. (1953) : *Sampling Techniques*; (Second Edition, 1963), John Wiley & Sons, New York.
- DALENIUS, T. (1957) : *Sampling in Sweden*; Almqvist & Wiksell, Stockholm.
- DEMING, W. E. (1950) : *Some Theory of Sampling*; John Wiley & Sons, New York.
- DEMING, W. E. (1960) : *Sample Design in Business Research*; John Wiley & Sons, New York.
- GALLUP, G. (1948) : *A Guide to Public Opinion Polls*; University Press, Princeton.
- HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953) : *Sample Survey Methods and Theory*; Volumes I and II; John Wiley & Sons, New York.
- HENDRICKS, W. A. (1956) : *The Mathematical Theory of Sampling*; Scarecrow Press, New Jersey and Bailey Bros. & Swinfen, London.
- HYMAN, H. H. (1955) : *Survey Design and Analysis*; Free Press, Illinois.
- JAMBUNATHAN, M. V. (1953) : *Some Aspects of Sampling*; India Book Co., Bangalore.
- JONES, C. D. (1949) : *Social Surveys*; Hutchinson, London.
- KENDALL, M. G. and STUART, A. (1966) : *The Advanced Theory of Statistics, Vol. 3—Design, Analysis and Time Series*; Charles Griffin & Co., London.
- KISH, L. (1965) : *Survey Sampling*; John Wiley & Sons, New York.
- LANSING, J. B., GINSBERG, G. P. and BRAATEN, K. (1961) : *An Investigation of Response Error*; University of Illinois, Urbana.
- MACE, A. E. (1964) : *Sample-size Determination*; Chapman & Hall, London.
- MAHALANOBIS, P. C. (1961) : *Experiments in Statistical Sampling in the Indian Statistical Institute*; Statistical Publishing Society, Calcutta.
- MAHALANOBIS, P. C. (1967) : *Sample Census of Area under Jute in Bengal in 1940*; Statistical Publishing Society, Calcutta.
- MURTHY, M. N. (1967) : *Sampling Theory and Methods*; Statistical Publishing Society, Calcutta.
- PANSE, V. G. (1958) : *Estimation of Crop Yields*; Food and Agricultural Organization, Rome.
- PARTEN, M. B. (1950) : *Surveys, Polls & Samples*; Harper & Bros., New York.
- PEATMAN, J. G. (1947) : *Descriptive and Sampling Statistics*; Harper and Bros., New York.

- SAMPFORD, M R (1962) *Introduction to Sampling Theory with Applications to Agriculture* Oliver & Boyd, London
- SANDEESSON, F H (1954) *Methods of Crop Forecasting*, Harvard University Press, Cambridge, Massachusetts
- SLONIM, M J (1960) *Sampling in a Nutshell* Simon & Schuster, New York
- SOM, R K (1967) *Recall Lapse in Demographic Enquiries*, Asia Publishing House Bombay
- STEPHAN, F F and McCARTHY, P J, (1958) *Sampling Opinions*, Charman and Hall London
- STUART, A (1962) *Basic Ideas of Scientific Sampling*, Charles Griffin & Co , London
- SUKHATME, P V (1953) *Sampling Theory of Surveys with Applications* Indian Society of Agricultural Statistics, New Delhi and Iowa State College Press, Ames Iowa
- TRUEBLOOD, R M and CYERT, R M (1957) *Sampling Techniques in Accountancy*, Engelwood New Jersey
- UNITED NATIONS STATISTICAL OFFICE (1960) *A Short Manual on Sampling* Volume I, United Nations New York
- VANCE, L L and NETER, J (1950) *Statistical Sampling for Auditors and Accountants* John Wiley & Sons New York
- YATES, F (1949) *Sampling Methods for Censuses and Surveys* (Third Edition 1960) Charles Griffin & Co , London
- ZARKOVICH, S S (1961) *Sampling Methods and Census*, Volume I, Food and Agricultural Organization, Rome
- ZARKOVICH, S S (edited by) (1965) *Estimation of Areas in Agricultural Statistics* Food and Agricultural Organization, Rome

APPENDIX 1 : TABLE OF RANDOM NUMBERS

row no.	column number									
	1	2	3	4	5	6	7	8	9	10
1	3436	6833	5809	9169	5081	5655	6567	8793	6830	1332
2	6133	4454	2675	3558	7624	5736	2184	4557	0496	8547
3	9853	3890	5535	3045	9830	5455	8218	9090	7266	4784
4	5807	5692	6971	6162	6751	5001	5533	2386	0004	2855
5	6291	0924	1298	7386	5856	2167	8299	9314	0333	8803
6	4725	9516	8555	0379	7746	9647	2010	0979	7115	6653
7	7697	6486	3720	6191	3552	1081	6141	7613	5455	3731
8	3497	2271	9641	0301	4425	6776	1205	2953	5669	1056
9	8910	4765	1641	0606	4970	7582	7991	6480	2946	5190
10	1122	6364	5264	1267	4027	4749	0338	8406	1213	5355
11	4333	0625	3947	1373	6372	9036	7046	4325	3491	8989
12	7685	1550	0853	4276	1572	9348	6393	2113	8285	9195
13	0592	8341	4430	0196	9613	2643	6442	0870	5449	8560
14	3506	0774	0447	7461	4459	0866	1698	0184	4975	5447
15	8368	2507	3565	4243	6667	8324	3063	8809	4248	1190
16	2630	1112	6680	4863	6813	4149	8325	2271	1963	9569
17	3883	3897	1848	8150	8184	1133	6088	3641	6785	0658
18	1123	3013	5218	0635	9265	4052	1509	1280	0953	9107
19	1167	9827	4101	4496	1254	6814	2479	5924	5071	1244
20	7831	0877	3806	9734	3801	1651	7169	3974	1725	9709
21	2487	9756	9886	6776	9426	0820	3741	5427	5293	3223
22	1245	3875	9816	8400	2938	2530	0158	5267	4639	5428
23	5309	4806	3176	8397	5758	2503	1567	5740	2577	8899
24	7109	0702	4179	0438	5234	9480	9777	2858	4391	0979
25	8716	7177	3386	7643	6555	8665	0768	4409	3647	9286
26	9199	5280	5150	2724	6482	6362	1566	2469	9704	8165
27	3125	4552	6014	0222	7520	1521	8205	0599	5167	1654
28	3788	6257	0632	0693	2263	5290	0511	0229	5951	6808
29	2242	2143	8724	1212	9485	3985	7280	0130	7791	6272
30	0900	4361	6120	8573	9904	2269	6405	9459	3088	6903
31	7909	4528	8772	1876	2113	4781	8678	4873	2061	1835
32	0379	2073	2680	8258	6275	7149	6858	4578	5932	9582
33	0780	6661	0277	0998	0432	8941	8946	9784	0693	2491
34	8478	8093	6990	2417	0290	5771	1304	3306	8825	5937
35	2519	7869	9035	4282	0307	7516	2340	1190	8440	6551
36	2472	0823	6188	3303	0490	9486	2896	0821	5999	3697
37	8418	5111	9245	0857	3059	6689	6523	8386	6674	7081
38	8293	5709	4120	5530	8864	0511	5593	1633	4788	1001
39	9260	1416	2171	0525	6016	9430	2828	6877	2570	4049
40	6568	1568	4160	0429	3488	3741	3311	3733	7882	6985
41	6694	5994	7517	1339	6812	4139	6938	8098	6140	2013
42	2273	6382	2673	6903	4044	3064	6738	7554	7734	7899
43	6361	5762	0322	2592	3452	9002	0264	6009	1311	5873
44	6696	1759	0563	8104	5055	4078	2516	1631	5859	1331
45	3431	2522	2206	3938	7860	1886	1229	7734	3283	8487
46	4842	3765	3484	2337	0587	9885	8568	3162	3028	7091
47	8295	9315	5892	6981	4141	1606	1411	3196	9428	3300
48	4926	4677	8547	5258	7274	2471	4559	6581	8232	7405
49	5439	0994	3794	8444	1043	4629	5975	3340	3793	6060
50	2031	0283	3320	1595	7953	2695	0399	9793	6114	2091

TABLE OF RANDOM NUMBERS (*contd.*)

row no	11	12	13	14	15	16	17	18	19	20
51	0983	2330	1303	4219	0189	4453	0806	1970	4130	7998
52	4634	6385	8760	3555	0567	8815	4700	5092	0231	5757
53	5432	9770	2781	6469	7152	0256	6137	0158	0968	9610
54	2317	5906	3861	0210	8610	5155	9252	4425	7449	0449
55	6836	2472	0385	4924	0569	6486	0819	9121	8586	9478
56	9358	5197	4910	0263	2372	6446	0252	0383	6518	0707
57	5936	9276	7805	3690	7473	5954	3164	3482	1845	7636
58	4306	9165	6438	6777	4671	2360	3382	2686	8767	6827
59	5951	7275	3713	5951	1452	1986	5034	0518	9314	7164
60	2108	6157	6254	7483	2407	8609	2114	4095	2456	8169
61	9566	6198	4546	8964	4473	5657	9152	3956	6235	9991
62	3981	3873	6448	0871	2825	7693	9304	9016	5871	9251
63	8696	2811	5419	9481	4498	1718	7871	1245	7915	2534
64	1433	1167	7332	0970	0159	1218	4679	9568	5533	8206
65	2141	6763	3510	7475	5991	8210	6588	5652	2636	7328
66	5145	6443	2930	1322	7296	4063	9397	4389	1295	3782
67	1339	4168	2508	0980	4184	7238	1406	9956	8366	9846
68	0948	6094	9141	8128	5545	9938	2129	7718	3561	2918
69	4252	3165	2934	4966	8313	0339	3724	9779	3113	9747
70	1898	4922	5411	9237	4511	6360	1905	9126	8473	8258
71	4014	3915	9924	2185	0045	5419	3618	0388	8833	7820
72	2177	3510	0681	6548	5318	7449	5776	5519	2420	5532
73	6625	0747	4812	5649	1408	3724	3681	1637	8352	4305
74	8271	1876	2939	1452	3071	0649	4840	9228	5237	5551
75	5745	1306	9341	2202	9409	3255	7968	6629	6267	4004
76	6164	6330	1234	4065	0816	7058	6369	1947	7346	4723
77	9956	5248	7969	9843	3265	5024	0971	4740	3295	2557
78	9811	9364	8786	4365	7833	0898	5798	9136	3829	5329
79	7346	9293	7714	6558	1103	9861	4270	3645	0912	3498
80	8061	5526	9875	6795	9549	2156	0845	0166	5267	1713
81	8425	0589	3180	4949	9893	8201	4108	6655	5819	1862
82	6464	9513	4697	4312	8602	7950	6790	1419	0407	6701
83	5382	7915	3116	5410	2990	9157	6348	3856	6925	0790
84	1933	3542	9212	3714	7075	1858	9857	1252	0681	5627
85	6426	5146	8050	5391	0055	6730	6866	0829	7953	3239
86	6984	3252	3254	1512	5402	0137	3837	1293	9329	1218
87	9080	7780	2659	8744	2374	6620	2019	2652	1163	7777
88	5583	3674	4040	8915	2860	9783	2497	6507	5084	8877
89	8578	8170	3723	8433	3395	2329	7783	7511	7075	1126
90	3899	0413	0663	3896	2100	3516	7169	0934	8257	9755
91	9372	7493	9462	3932	7468	3383	4358	7037	2542	5480
92	4747	1794	4498	1693	0955	5373	5400	5226	4811	0379
93	3545	6861	4232	3952	9316	1867	0537	2144	1034	9889
94	0836	9910	8303	7618	9262	7540	1802	7089	7172	0442
95	9742	4735	1085	9715	2103	5485	3740	4117	2786	5815
96	9890	5980	2778	5956	6128	2384	8501	3302	7232	6363
97	5960	4185	7079	8917	2378	6868	6472	9093	8609	4008
98	9017	3136	4463	4174	8453	5045	4925	7889	7188	6990
99	8520	7719	6078	0293	0525	7426	8334	2367	5490	4960
100	1436	3124	0072	5146	8555	7584	8382	1378	3848	7323

APPENDIX 2

FORMULAE FOR ESTIMATORS OF Y AND VARIANCE ESTIMATORS.

APPENDIX 2

659

sampling method	estimator (\hat{Y})	variance estimator $v(\hat{Y})$
1.1 with replacement	$N\bar{y}$ or Np	1. Simple Random Sampling $\left\{ \begin{array}{l} N^2 s^2/n \\ N^2(1-f)s^2/n \end{array} \right\}$ $\left\{ \begin{array}{l} s^2 - \sum_{i=1}^n (y_i - \bar{y})^2 \\ N(n-1) \end{array} \right\}$ $\text{or } npq/(n-1)$
1.2 without replacement	$N\bar{y}$ or Np	$N^2(1-f)s^2/n$ $N^2(1-f)\frac{s'^2}{d'} \quad \text{(3.21)}$ $N^2(1-f)\frac{s'^2}{d'} = \sum_{i=1}^d (y'_{i+1} - \bar{y}')^2 \quad \text{(3.25)}$ $\text{or } \frac{dp'q'}{d-1}, f' = \frac{d}{N}$
1.3 distinct units in 1.1	$N\bar{y}'$ or Np'	
2.1 linear	$k\bar{y}\bar{y}$	2. Systematic Sampling $\left\{ \begin{array}{l} \frac{\lambda^2(1-f)}{2n(n-1)} \\ k^2 \end{array} \right\}$ $(Soc. 5.2b)$
2.2 circular	$N\bar{y}$	3. Sampling with probability proportional to size $3.1 \text{ with replacement}$ $\frac{X}{n} \sum_{i=1}^n (y_i/x_i)$ $\frac{1}{n(n-1)} [X^2 - \sum_{i=1}^n (y_i^2/x_i^2) - n\hat{P}^2]$ (6.3) $3.1a \text{ general case}$ $\frac{G}{n} \sum_{i=1}^n p_i$ $\frac{G^2}{n(n-1)} (p\bar{q} - \frac{1}{n} \sum_{i=1}^n p_i q_i), \bar{p} = \frac{1}{n} \sum_{i=1}^n p_i$ (6.23) $3.1b \text{ area sampling (estimation of drop proportion)}$

Note : Notations used are as given in Table 2.5 of Chapter 2 (pp. 50-51) and in the text. The references in brackets denote equation numbers or section numbers where the same or similar expressions are given. The expressions not included in the text are denoted by † within brackets.

APPENDIX 2 (contd.) FORMULAE FOR ESTIMATORS OF Y AND VARIANCE ESTIMATORS

sampling method	estimator (\hat{Y})	variance estimator $v(\hat{Y})$
3.2 without replacement		
3.2a general case	$\sum_{i=1}^n (y_i/\pi_i)$	(6.36) $\left\{ \begin{array}{l} v_1 = \sum_{i=1}^n (1-\pi_i) \left(\frac{y_i}{\pi_i} \right)^2 \\ \quad + \sum_{i=1}^n \sum_{i' \neq i} \frac{\pi_{ii'} - \pi_i \pi_{i'i}}{\pi_{ii'}} \left(\frac{y_i}{\pi_i} - \frac{y_{i'}}{\pi_{i'}} \right)^2 \end{array} \right.$
		(6.39) $v_2 = \sum_{i=1}^n \sum_{i' > i} \frac{\pi_{ii'} - \pi_{i'i}}{\pi_{ii'}} \left(\frac{y_i}{\pi_i} - \frac{y_{i'}}{\pi_{i'}} \right)^2$
3.2b ordered estimator $\frac{1}{n} \sum_{i=1}^n t_i$		(6.41) $(6(n-1))^{-1} (\sum_{i=1}^n t_i^2 - n\bar{t}^2), t_i = \sum_{i=1}^i y_{i'}, (1 - \sum_{i=1}^i p_i) \frac{y_i}{p_i}$
3.2c ordered estimator $\frac{1}{2} \left\{ (1+p_1) \frac{y_1}{p_1} + (1-p_1) \frac{y_2}{p_2} \right\}$ ($n = 2$)		(6.43) $\frac{3(1-p_1)^2}{4} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2$
3.2d unordered estimator ($n = 2$)	$\frac{1}{2-p_1-p_2} \left\{ (1-p_2) \frac{y_1}{p_1} + (1-p_1) \frac{y_2}{p_2} \right\}$	(6.44) $\frac{(1-p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2$
		(6.46) $+ (1-p_1) \frac{y_2}{p_2}$

APPENDIX 2. (contd.) FORMULAE FOR ESTIMATORS OF χ AND VARIANCE ESTIMATORS.

sampling method	estimator (\hat{Y})	variance estimator $v(\hat{Y})$
3.3 random group method	$\sum_{i=1}^n (y_i/p_i) P'_i, \quad P'_i = \frac{N_i}{S} (X_i/N) \quad (6.50)$	$\left(\frac{y_i}{p_i} - \hat{Y} \right)^2 P'_i \mid (N^2 - \sum_{i=1}^n N_i^2) \quad (6.53)$
3.4 pps systematic (random arrangement)	$\frac{1}{n} \sum_{i=1}^n S (y_i/p_i) \quad (6.55)$	$(0.55) \quad \frac{S}{n} \sum_{i=1}^n \left\{ \frac{1-n(p_i+p_i') + n \sum_{i'=1}^{i-1} P_{i'}^2}{n^2(n-1)} \right\} \left(\frac{y_i}{p_i} - \frac{y_{i'}}{p_{i'}} \right)^2 \quad (6.57)$
3.5 probability proportional to total size of sample	$(\bar{y}/v) X \quad (6.58)$	$(0.58) \quad \hat{Y} = \frac{N \bar{X}}{n \bar{v}} \left\{ \frac{\sum_{i=1}^n v_i^2}{n-1} + \frac{N-1}{n-1} \sum_{i=1}^n v_i y_{pi} \right\} \quad (6.60)$
4. Stratified Sampling		
4.1 general case	$\sum_{s=1}^K \hat{Y}_s \quad (7.2)$	$\sum_{s=1}^K v(\hat{Y}_s) \quad (7.3)$
4.2 sis without replacement	$\sum_{s=1}^K \frac{N_s}{n_s} \sum_{i=1}^{n_s} y_{si} \quad (7.23)$	$\sum_{s=1}^K \frac{N_s^2(1-f_s)}{n_s(n_s-1)} \left(\sum_{i=1}^{n_s} y_{si}^2 - n_s \bar{y}_s^2 \right) \quad (7.10)$
4.3 pps with replacement	$\sum_{s=1}^K \frac{1}{n_s} \sum_{i=1}^{n_s} (y_{si}/p_{si}) \quad (7.61)$	$\sum_{s=1}^K \frac{1}{n_s(n_s-1)} \left\{ \sum_{i=1}^{n_s} (y_{si}/p_{si})^2 - n_s \hat{Y}_s^2 \right\} \quad (7.66)$

APPENDIX 2 (contd.) FORMULAE FOR ESTIMATORS OF Y AND VARIANCE ESTIMATORS

Sampling method	estimator (\hat{Y})	variance estimator $v(\hat{Y})$
5 Cluster Sampling		
5.1 srs w/o r (equal cluster size)	$\frac{N}{n} \sum_{i=1}^n \sum_{j=1}^{M_i} y_{ij}$	(See 8.2b)
5.2 srs w/o r (unequal cluster size)	$\frac{N}{n} \sum_{i=1}^n \frac{M_i}{\sum_j M_j} y_{ij}$	(8.29)
5.3 ppsswr (unequal cluster size)	$\frac{NM}{n} \sum_{i=1}^n \bar{y}_{i0}$	(8.30)
6 Two stage Sampling		
6.1. srs w/o r at both stages	$\frac{N}{n} \sum_{i=1}^n \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$	(9.8)
6.2 ppsswr at first stage and srs w/o r at second stage	$\frac{1}{n} \sum_{i=1}^n \frac{1}{p_i} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$ (9.25)	$\left\{ \frac{1}{n(n-1)} \left(\sum_{i=1}^n t_i^2 - n\hat{T}^2 \right), t_i = \begin{cases} \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}/p_i & \text{for } 0.2 \\ \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}/p_i & \text{for } 0.3 \\ \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}/p_i & \text{for } 0.7 \end{cases} \right\}$ (9.27)
6.3 ppsswr at both stages	$\frac{1}{n} \sum_{i=1}^n \frac{1}{p_i} \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{y_{ij}}{p_{ij}}$	(9.11)

APPENDIX 2. (contd.) FORMULAE FOR ESTIMATORS OF Y AND VARIANCE ESTIMATORS.

sampling method	estimator (\hat{Y})	variance estimator $v(\hat{Y})$
7. Three-stage Sampling		
7.1 srs w/o at all the stages	$\frac{N}{n} \sum_{i=1}^n M_i \hat{Y}_i,$	(†) $\frac{(1-f)}{n(n-1)} (N^2 \sum_{i=1}^n M_i^2 \hat{Y}_i^2 - n\hat{Y}^2)$
	$\hat{Y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} L_{ij} \bar{y}_{ij},$	$+ \frac{N}{n} \sum_{i=1}^n \frac{M_i^2(1-f_i)}{m_i(m_i-1)} (\sum_{j=1}^{m_i} L_{ij}^2 \bar{y}_{ij}^2 - m_i \hat{Y}_i^2)$
	$\bar{y}_{ij} = \frac{1}{l_{ij}} \sum_{k=1}^{l_{ij}} y_{ijk},$	$+ \frac{N}{n} \sum_{i=1}^n \frac{M_i}{m_i} \frac{m_i}{l_{ij}(l_{ij}-1)} (\sum_{k=1}^{l_{ij}} y_{ijk}^2 - l_{ij} \bar{y}_{ij}^2), \quad (†)$
		$f_{ij} = l_{ij}/L_{ij}, f_i = m_i/l_i, f = n/N.$
7.2 ppssyr at all the stages	$\frac{1}{n} \sum_{i=1}^n \frac{1}{m_i p_i} \sum_{j=1}^{m_i} \frac{l_{ij}}{l_{ij} p_{ij}} \frac{y_{ijk}}{p_{ijk}}, \quad (9.61)$	$\frac{1}{n(n-1)} (\sum_{i=1}^n t_i^2 - n\hat{Y}^2),$ $t_i = \frac{1}{p_i} \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{1}{p_{ij}} \frac{y_{ijk}}{p_{ijk}} \quad (9.63)$

APPENDIX 2 (contd.) FORMULAE FOR ESTIMATORS OF Y AND VARIANCE ESTIMATORS

sampling method	estimator (\hat{Y}^*)	variance estimator $v(\hat{Y}^*)$
8 Ratio Method of Estimation		
8.1 general case	$\hat{R} X, \hat{R} = \hat{Y}/\hat{X}$	(10.2) $v(\hat{Y}) - 2\hat{R} \operatorname{cov}(\hat{X}, \hat{Y}) + \hat{R}^2 v(\hat{X})$ (10.12)
8.2 srs wor	$\hat{R} X, \hat{R} = \bar{y}/\bar{x}$	(Sec 10.5a) $\frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n (y_i - \hat{R}\bar{x}_i)^2$ (Sec 10.5a)
8.3 systematic sampling	$\hat{R} X, \hat{R} = \bar{y}/\bar{x}$	(Sec 10.5b) $\frac{N^2(1-f)}{2n(n-1)} \sum_{i=1}^{n-1} \{(y_{i+1} - y_i) - \hat{R}(x_{i+1} - x_i)\}^2$ (Sec 10.5b, 5.8a)
8.4 pps with replacement	$\left\{ \sum_{i=1}^n \frac{y_i}{p_i} / \sum_{i=1}^n \frac{x_i}{p_i} \right\} X$ (Sec 10.5c)	$\frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{p_i} - \hat{R} \frac{x_i}{p_i} \right)^2, \hat{R} = \hat{Y}^*/X$ (Sec 10.5c)
8.5 stratified sampling (srs wor)(combined estimator)	$\left\{ \sum_{s=1}^K N_s \bar{y}_s / \sum_{s=1}^K N_s \bar{x}_s \right\} X$ (10.38)	$\sum_{s=1}^K \frac{N_s^2(1-f_s)}{n_s(n_s-1)} \left\{ \sum_{i=1}^{n_s} (y_{si} - \hat{R}_{si} x_{si})^2 - n_s(y_s - \hat{R}_s \bar{x}_s)^2 \right\}, \hat{R} = \hat{Y}^*/X$ (Sec 10.6)
8.6 stratified sampling (srs wor)(separate estimator)	$\sum_{s=1}^K (y_{si}/\bar{x}_s) \bar{X}_s$	(10.41) $\sum_{s=1}^K \frac{N_s^2(1-f_s)}{n_s(n_s-1)} \sum_{i=1}^{n_s} (y_{si} - \hat{R}_{si} \bar{x}_{si})^2, \hat{R}_s = \frac{\bar{y}_s}{\bar{x}_s}$ (Sec 10.6)

APPENDIX 2. (contd.) FORMULAE FOR ESTIMATORS OF Y AND VARIANCE ESTIMATORS.

sampling method	estimator (\hat{Y}^*)	variance estimator $v(\hat{Y}^*)$
9. Product and Regression Estimators		
9.1 general case (product estimator)	$\hat{Y} \cdot \hat{X}/\lambda$	(10.4.6) $v(\hat{Y}) + 2\hat{R} \operatorname{cov}(\hat{X}, \hat{Y}) + \hat{R}^2 v(\hat{X}), \hat{R} = \hat{Y}/\lambda.$ (Sec. 10.8)
9.2 general case (regression estimator)	$\hat{Y} + \hat{\beta}(\lambda - \hat{X})$	(11.7) $v(\hat{Y}) - 2\hat{\beta} \operatorname{cov}(\hat{X}, \hat{Y}) + \hat{\beta}^2 v(\hat{X}).$ (Sec. 11.3)
10. Two-phase Sampling		
10.1 ratio method with srswo and srswo	$N r \bar{x}_1, \quad r = \frac{\bar{y}_2}{\bar{x}_2},$	(10.89) $\frac{N^2(n_1 - n_2)}{n_1 n_2(n_2 - 1)} \sum_{j=1}^{n_2} (y_j - rx_j)^2 + \frac{N^2}{n_1} \frac{n_2 (y_j - \bar{y}_2)^2}{n_2 - 1},$ (Sec. 10.13)
10.2 regression method with srswo and srswo	$N \{\bar{y}_2 + \hat{\beta}(\bar{x}_1 - \bar{x}_2)\}$	(11.21) $\frac{N^2(n_1 - n_2)}{n_1 n_2(n_2 - 1)} \sum_{j=1}^{n_2} \{(y_j - \bar{y}_2) - \hat{\beta}^2(x_j - \bar{x}_2)\}^2 + \frac{N^2}{n_1} \frac{n_2}{n_2 - 1} \frac{(y_j - \bar{y}_2)^2}{n_2 - 1} \quad (\dagger)$
10.3 srswo and ppswo	$N \frac{\bar{x}_1}{n_1} S \sum_{j=1}^{n_2} \frac{y_j}{x_j}$	(9.72) $\frac{N^2 \lambda_1 \bar{x}_1^2}{n_2(n_2 - 1)} \sum_{j=1}^{n_2} \frac{y_j^2}{x_j^2} + \frac{N^2 \bar{y}_1}{n_2(n_2 - 1)} \frac{n_2 y_j^2}{x_j} - \hat{Y}^2(\lambda_1 \lambda_2 - 1), \quad \lambda_1 = \frac{n_1}{n_1 - 1}, \quad \lambda_2 = \frac{n_2}{n_2 - 1} \quad (\ddagger)$
10.4 srswo and stratified srswo	$\frac{N}{n_1} \sum_{s=1}^K \frac{n_{1s}}{n_{2s}} S \sum_{j=1}^{n_s} y_{sj}$	(†) $\frac{N^2 n_1}{n_1 - 1} \sum_{s=1}^K \left\{ \frac{w_{1s}(n_1 w_{1s} - 1)}{n_1 n_{2s}(n_{2s} - 1)} \frac{n_{2s}}{S} (y_{sj} - \bar{y}_{2s})^2 + \frac{w_{1s}}{n_1} (\bar{y}_{2s} - \bar{y}_{sj})^2 \right\}, \quad w_{1s} = n_{1s}/n_1, \quad \bar{y}_{sj} = \hat{Y}^*/ N . \quad (\dagger)$

Note : In (10.4), n_{1s} is the number of units falling in the s -th stratum and n_{2s} is the number of units sampled from that stratum in the second phase.

APPENDIX 3

SOLUTIONS TO PROBLEMS

Chapter 2

- 2.3 cf Section 2.5, p 36
 2.4 cf Sections 2.6, 2.7 and 2.11, pp 38, 40, 45
 2.5 $E(x+y) = p+p$, $E(x-y) = p-p'$, $E(xy) = pp'$,
 $V(x+y) = V(x-y) = pq + p^2$, $V(xy) = pp(1-pp)$
 2.6 cf Section 2.10, p 44 $c < 1/v$
 2.7 (a) t_2 , (b) t_1

$$2.8 \quad \left\{ 1 - \sum_{i=1}^k w_i^2 \right\}^{-1} \left\{ \sum_{i=1}^k w_i^2 t_i^2 - \left(\sum_{i=1}^k w_i^2 \right) t^2 \right\}, \quad w_t = \left(\frac{1}{V_t} \right) / \sum_{i=1}^k \frac{1}{V_i}$$

$$2.10 \quad E \left(\sum_{i=1}^k a_i t_i \right) = \sum_{i=1}^k a_i \theta_i,$$

$$V \left(\sum_{i=1}^k a_i t_i \right) = \sum_{i=1}^k a_i^2 V(t_i) + \sum_{i=1}^k \sum_{i \neq j} a_i a_j \operatorname{Cov}(t_i, t_j)$$

Chapter 3

- 3.2 $\hat{H}_b = 40$, $c(\hat{H}_b) = 29.28\%$, $\hat{P}_b = 175$, $c(\hat{P}_b) = 30.60\%$
 3.3 (i) $c(\hat{Y}) = 32.94\%$, (ii) $c(\hat{X}) = 14.07\%$
 3.4 $\hat{P} = 10.87\%$, $c(\hat{P}) = 3.12\%$
 3.5 $c(\hat{Y}) = 11.67\%$
 3.6

size class	\hat{P}	\hat{A} (acres)	\hat{F} (lbs.)	$c(\hat{P})$	$c(\hat{A})$	$c(\hat{F})$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0.2778	186.96	2400	27.25	33.49	49.29
2	0.2222	240.24	11388	31.62	34.83	50.35
3	0.3333	280.08	11688	23.88	25.98	41.23
4	0.1667	305.40	16476	37.79	38.73	46.73

- 3.7 $\operatorname{Eff}(\hat{Y}_1 | \hat{Y}_2) = 13.11\%$
 3.8 $\rho(\bar{x}, \bar{y}) = \rho$, $\rho(\bar{x} + \bar{y}, \bar{x} - \bar{y}) = (\sigma_x^2 - \sigma_y^2) / \sqrt{(\sigma_x^2 + \sigma_y^2)^2 - 4\rho^2\sigma_x^2\sigma_y^2}$

3.9 $\hat{P} = N_1\bar{x} + N_2\bar{y} + N_3\bar{z}; \quad \hat{S} = (N_1\bar{x}C_1 + N_2\bar{y}C_2 + N_3\bar{z}C_3)/1000;$

$$V(\hat{P}) = N_1^2(1-f_1) \frac{\sigma_x'^2}{n_1} + N_2^2(1-f_2) \frac{\sigma_y'^2}{n_2} + N_3^2(1-f_3) \frac{\sigma_z'^2}{n_3}; \quad f_i = \frac{n_i}{N_i}, \quad i = 1, 2, 3;$$

$$v(\hat{S}) = \left\{ N_1^2 C_1^2 (1-f_1) \frac{s_x^2}{n_1} + N_2^2 C_2^2 (1-f_2) \frac{s_y^2}{n_2} + N_3^2 C_3^2 (1-f_3) \frac{s_z^2}{n_3} \right\} \div (1000)^2.$$

3.10 $\text{Eff} = Q - (P/C^2), \quad C = \sigma/\bar{Y}.$

3.11 yes; $n\bar{y}/(n+C^2); \quad \text{Eff} = 1 + (C^2/n).$

3.12 $E(\bar{y}') = \bar{Y}; \quad V(\bar{y}') = \frac{\sigma^2}{n} \frac{1+3(n'/n)}{\{1+(n'/n)\}^2}; \quad \frac{n'}{n} = \frac{1}{3}.$

3.14 $V(y_t) = \sigma^2; \quad \text{Cov } (y_t, y_{t'}) = -\sigma^2/(N-1); \quad \text{cf. Problem 2.10, p. 666.}$

3.15 $V(v_1) = \frac{\sigma^4}{2n^2(2n-1)}; \quad V(v_2) = \frac{\sigma^4}{2n^2}; \quad V(v_3) = \frac{\sigma^4}{4n^2(n-1)};$

$$V(v_1) < V(v_3) < V(v_2); \quad \text{cf. Sub-section 3.3d, p. 64.}$$

3.16 $P(d=1) = 1/N^2; \quad P(d=2) = 3(N-1)/N^2; \quad P(d=3) = (N-1)(N-2)/N^2;$

$$V(\bar{y}_d) = \frac{2N-1}{6N} \sigma^2; \quad v(\bar{y}_d) = \frac{(2N-1)(N-1)}{6N^2} s_d^2.$$

Note : In Problem 3.16 (p. 93) 'srs wor' should read 'sts wr'.

3.18 $\hat{N} = 2\bar{y}-1; \quad V(\hat{N}) = (N-n)(N+1)/3n; \quad v(\hat{N}) = 4(1-f)s^2/n;$

\bar{y} and $(N-1)s^2$ are the mean and the corrected sum of squares of the original serial numbers of the sample units.

3.19 $\hat{P} = \frac{m-1}{n-1}; \quad V(\hat{P}) = \frac{P^2Q}{n}; \quad v(\hat{P}) = \frac{m(n-m)}{n^2(n-1)};$

n is the number of draws required to select m units possessing the rare item.

3.20 $\hat{N} = \frac{n}{m} (r+1)-1; \quad V(\hat{N}) = \frac{(N+1)(N-r)}{m(r+2)} (r-m+1);$

$$v(\hat{N}) = \frac{n(n-m)(r+1)}{m^2(m+1)} (r-m+1).$$

3.21 $\hat{D} = \frac{MN}{mn} d; \quad V(\hat{D}) = \left(\frac{MN}{mn} - 1 \right) D; \quad v(\hat{D}) = \frac{MN}{mn} \left(\frac{MN}{mn} - 1 \right) d.$

Chapter 4

4.1 (i) 109, 123, 132, 136, 137; (ii) 139 for all the five cases in (i).

4.2 (i) 1279 and 1521; (ii) yes; (iii) 614.

4.3 $n' = (100)^4/n\alpha^4.$

4.4 (i) 2475; (ii) 4950; assumptions : sex ratio is about unity; incidence rate of disease is approximately the same for males and females.

4 5 262, assumption population cv is the same in the two villages

$$4 6 n = C_y \sqrt{\lambda/C_1}, C_y = \sigma'_y/\bar{Y}$$

$$4 8 (i) E(x) = E(y) = 49.5, V(x) = V(y) = 833.25, (ii) E(x-y)^2 = 1666.5$$

4 9 Hint In two tosses $P(HT) = P(TH)$

4 12 (i) No (ii) Continue selection till three paddy growing fields get selected,
(iii) total number of paddy growing fields in the village

4 13 cf Sub section 6.10b of Chapter 6, p 202

$$4 14 (i) m = \sqrt{MN/\lambda(Dx^2+1)} (ii) m = 3326, n = 1663$$

Chapter 5

5 1 (i) $q \geq n-r$, ($N = nq+r$), (ii) integer nearest to (N/n) if $q > n-r$, q otherwise (iii) and (iv) y unbiased for \bar{Y}

5 2 (i) 331.77% (ii) 14389%

5 3 (ii) 113.98% (iii) $\hat{Y} = 18397$ acres, $c(\hat{Y}) = 12.57\%$

5 4 (i) 9270 hectares (ii) $c(\hat{Y}) = 0.59\%$ (iii) $c'(\hat{Y}) = 12.58\%$

5 5 (i) rse = 17.92%, Eff = 70.86%

5 6

n	efficiency %		value of ρ	
	x	y	x	y
2	67.69	91.28	+0.3428	-0.0041
3	163.95	145.18	-0.2505	-0.2183
4	328.51	749.79	-0.2595	-0.3010
6	158.44	142.11	-0.1312	-0.1232

$$5 11 V(y) = \frac{\sigma^2}{n} \left\{ 1 + \frac{2}{n} \sum_{a=1}^{n-1} (n-a)\rho_a \right\}$$

$$5 12 (i) E(V_{xy}) = E(V_r) = \frac{k-1}{k} \frac{\sigma^2}{n}, \sigma^2 = \frac{1}{N} \sum_{i=1}^N \sigma_i^2,$$

$$(ii) E(V_{xy}) = \frac{k-1}{k} \frac{\sigma^2}{n} + \beta^2 \frac{k^2-1}{12} = E(V_r) - \frac{\beta^2}{12} k(k-1)(n-1)$$

$$5 15 C_z^2 < N-1, z = 1, 2, \dots, g$$

$$5 16 (i) V_{xy} = (k^2 + L^2 - 2)/12, V_r = (MN - mn)(N^2 + M^2 - 2)/12(MN - 1)mn$$

5 17 Let y_A , y_B , y_C and y_D be the sample totals

$$(i) A = I_1 y_A, B = N_1 y_B/m$$

$$(ii) A = (I_1 I_2 - I_1 - I_2)y_A/I_1 I_2, C = I_1 y_C, D = I_2 y_D,$$

$$(iii) B = N_1 y_B/m.$$

Chapter 6

6.1 (i) 378.4%; (ii) 75.96%.

6.2 (i) 351644 acres; (ii) 3.84%; (iii) $n = 37$.

6.3 (i) 41.74%; (ii) 99.03%.

6.4

(a)	unit	Π_t	(b)	unit	$\Pi_{tt'}$	units	$\Pi_{tt'}$	units	$\Pi_{tt'}$
	1	0.0781		12	0.00539	24	0.02289	37	0.06572
	2	0.1534		13	0.00825	25	0.02926	45	0.06082
	3	0.2254		14	0.01124	26	0.03597	46	0.07467
	4	0.2939		15	0.01438	27	0.04304	47	0.08929
	5	0.3583		16	0.01768	34	0.03500	56	0.09528
	6	0.4182		17	0.02116	35	0.04472	57	0.11388
	7	0.4727		23	0.01681	36	0.05493	67	0.13961

6.5 $\hat{Y} = Ay'$, $V(\hat{Y}) = \frac{A}{n} \sum_{i=1}^N (Y'_i - Y')^2 A_i$ and $v(\hat{Y}) = \frac{A^2}{n(n-1)} \sum_{i=1}^n (y'_i - \bar{y}')^2$,

where \bar{y}' is the sample mean of yield rates, A is the total area under wheat, Y'_i is the yield rate for i -th unit and Y' is the overall yield rate.

6.6 cf. Sub-section 6.4b p. 188.

6.8 $M_0 = \frac{1}{Q(r)} \left\{ n + \frac{t_\alpha^2}{2} P(r) + \frac{t_n}{2} \sqrt{\{4n + t_\alpha^2 P(r)\} P(r)} \right\},$

where $P(r)$ is the probability of rejection of a draw, $Q(r) = 1 - P(r)$ and t_α is given by $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t_\alpha} e^{-\frac{1}{2}x^2} dx = \alpha$.

6.9 (a)

6.10 $E\{V(\hat{Y}_{HT})\} = a^2 V(\hat{X}_{HT}) + \sum_{i=1}^N \left(\frac{1}{\Pi_i} - 1 \right) \sigma_i^2.$

6.11 (ii) $\sum_{t=1}^N \Pi_t = E\{v(s)\} = v$, $\sum_{i=1}^N \sum_{i' \neq i}^N \Pi_{ti'} = v(v-1) + V\{v(s)\}$,

where $v(s)$ is the number of distinct units in the s -th sample;

(iii) (6.37), p. 210; (iv) (6.39) with n replaced by $v(s)$, p. 211.

6.12 $\Pi_t \Pi_{ti'} \geq \Pi_{ti'}, \quad i' \neq i$.

6.13 $\lambda = \binom{N-1}{n-1}.$

$$6.19 \quad \hat{Y} = \frac{1}{n} \sum_{i=1}^g \sum_{j=1}^{r_i} \frac{y_{ij}}{p_i}, \quad V(\hat{Y}) = \frac{1}{n} \left(\sum_{i=1}^g \sum_{j=1}^{r_i} \frac{N_i Y_{ij}^2}{P_i} - \bar{Y}^2 \right) \\ - \frac{n-1}{n} \sum_{i=1}^g N_i \sigma_i^2,$$

$$v(\hat{Y}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^g \sum_{j=1}^{r_i} \left(\frac{y_{ij}}{p_i} - \hat{Y} \right)^2 - \sum_{i=1}^g \frac{r_i}{N_i p_i} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_i)^2 \right],$$

where r_i is the number of times the i th group gets selected

$$6.20 \quad v(\hat{Y}) = \frac{1}{r(r-1)} \left(\sum_{i=1}^g \frac{y_i^2}{p_i^2} r_i - r \bar{Y}^2 \right)$$

$$6.23 \quad \Pi_{ii'} = 2P_i P_{i'} \left(\frac{1}{1-2P_i} + \frac{1}{1-2P_{i'}} \right) \left(1 + \sum_{i=1}^N \frac{P_i}{1-2P_i} \right)^{-1}$$

6.24

arrangement	proportion of villages		
	common	adjacent	different
existing	0.440	0.474	0.086
modified*	0.569	0.379	0.052

*modified arrangement—14, 13, 12, 15, 16, 8, 10, 11, 9, 6, 7, 3, 4, 5, 1, 2

Chapter 7

$$7.1 \quad (i) \hat{Y} = 1353572, c(\hat{Y}) = 9.12\%, (ii) 407\%, (iii) 86.38\%$$

$$7.2 \quad \text{Eff}(a|c) = 12.56\% \quad \text{Eff}(b|c) = 94.88\%$$

$$7.3 \quad (i) \hat{Y}_1 = 3247, \hat{Y}_2 = 11489, \hat{Y}_3 = 9831, \hat{Y} = 24567, \\ (ii) c(\hat{Y}_1) = 20.38\%, c(\hat{Y}_2) = 6.34\%, c(\hat{Y}_3) = 2.37, c(\hat{Y}) = 6.79\%$$

$$7.4 \quad (i) n_s = 45, 48, 43, 37, 21, 30, 65, 51, 97,$$

$$(ii) n_s = 40, 57, 50, 50, 38, 32, 57, 40, 73, \text{Eff} = 94.24\%$$

$$7.5 \quad (i) n_s = 1222, 167, 611, (ii) \text{Eff} = 108.1\%$$

$$7.6 \quad \text{Eff} = 300.89\%$$

$$7.7 \quad n = 1107$$

$$7.9 \quad (i) \text{same}, (ii) \text{less efficient}$$

$$7.10 \quad n_1 = C' \frac{\sqrt{P_1 Q_1 / C_1}}{\sqrt{P_1 Q_1 C_1 + \sqrt{P_2 Q_2 C_2}}}, \quad n_2 = \frac{1}{C_2} (C' - n_1 C_1).$$

$$7.11 \quad \frac{V(v_2)}{V(v_1)} = 1 + \left\{ 4 \sum_{s=1}^k \sum_{s' \neq s} V_s V_{s'} \mid \sum_{s=1}^k V_s (\beta_s + 1) \right\},$$

where V_s and β_s are the variance and kurtosis of t_{s1} and t_{s2} .

7.13 cf. Section 7.6, pp. 251-252.

$$7.15 \quad \text{(i)} \quad E(V_{st}) = E(V_{sy}); \quad \text{(ii)} \quad E(V_{st}) = \frac{k-1}{k} \frac{\sigma^2}{n} + \beta^2 \frac{k^2-1}{12n} < E(V_{sy}),$$

cf. solution to Problem 5.12, p. 668.

$$7.18 \quad y_0 = 1.6, \quad V(\bar{y}) = \frac{1}{n} \left\{ 1 - \frac{y_0^2 e - y_0}{1 - e - y_0} \right\}.$$

$$7.23 \quad B(\hat{\bar{Y}}) = P_2^* [(P_1 - W_1) \bar{Y}_1 + (P_2 - W_2) \bar{Y}_2]; \quad V(\hat{\bar{Y}}) = P_1^* V(\hat{\bar{Y}'}) + P_2^* V(\hat{\bar{Y}''}),$$

$$\text{where } V(\hat{\bar{Y}'}) = \sum_{t=1}^2 W_t^2 (1-f_t) \frac{\sigma_t'^2}{n_t}, \quad f_t = n_t/N_t, \quad \text{and}$$

$$V(\hat{\bar{Y}''}) = \frac{1}{n} \sum_{t=1}^2 P_t D_t^2 (N_t - n) \frac{\sigma_t'^2}{N_t} + \sum_{t=1}^2 \frac{W_t^2}{P_t} \bar{Y}_t^2 - \bar{Y}^2,$$

P_1^* and P_2^* being the probabilities of the events (i) and (ii) respectively.

Chapter 8

8.1	(i)	cluster size	observed variance	expected variance	cluster size	observed variance	expected variance
		1.00	0.1120	0.1085	12.25	0.0454	0.0466
		2.25	0.0873	0.0825	16.00	0.0419	0.0426
		4.00	0.0659	0.0680	25.00	0.0398	0.0366
		6.25	0.0577	0.0585	36.00	0.0342	0.0324
		9.00	0.0505	0.0517			

(ii) $x = 2.25$ acres; $n = 2449$.

8.2 str : 10 plots; systematic : 10 plots; ppswr and pps systematic : 1 plot.

8.3 One-foot bed.

$$8.4 \quad \text{eff} = \frac{1}{M} \left(1 + \frac{w}{b} \right).$$

$$8.8 \quad V = \frac{\sigma^2}{nM} \{1 + (M-1)\rho_c\} \{1 + (n-1)\rho_c'\}.$$

8.10 $(M_0 - 1)mn V(\hat{Y}) = (M_0 - m)(N\sigma^2 + (N-n)\bar{Y}^2) + N^2(N-n)(m-1)\{V_c/M_0(N-1)\}$,
 where $\sigma^2 = \frac{1}{M_0} \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij}^2 - \bar{Y}^2$ and $V_c = \frac{1}{N} \sum_{i=1}^N \left(M_i \bar{Y}_i - \frac{\bar{Y}}{N} \right)^2$.

Chapter 9

9.1 (i) $\hat{Y} = 35035$, $c(\hat{Y}) = 8.93\%$, (ii) 68.02%

9.2 $\hat{Y} = 3024868$, $c(\hat{Y}) = 28.74\%$

9.3 $m = 2$, $n = 365, 822, 4474$, rse $1.06\%, 0.71\%, 0.23\%$

9.4 (i) $\hat{x} = 57.71\%$, $c(\hat{x}) = 0.72\%$, (ii) 117.52%

9.5 $\hat{Y} = 906.2$ kg, $c(\hat{Y}) = 8.75\%$, (ii) $n = 156$

9.6 (i) $m = \sigma_w \sqrt{C_1} \sqrt{\sigma_b \sqrt{C_2}}$, $n = (C - C_0)/(C_1 + mC_2)$,
 (ii) $m = 7$, $n = 44$

9.7 $\hat{Y} = 3976$ tonnes, $c(\hat{Y}) = 8.65\%$

9.8 $\hat{Y} = M_0 y$, $y = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^{M_i} y_{ij}$, $v(\hat{Y}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2$.

(M_0 is the total number of households in the population)

9.11 (i) $m_t = m \left(\frac{M_t}{P_t} \sigma_t \right) / \sum_{t=1}^N M_t \sigma_t$, (ii) $m_t = nm \left(\frac{M_t}{P_t} \sigma'_t \right) / \sum_{t=1}^n \frac{M_t}{P_t} \sigma'_t$,

$$V_2 - V_1 = \frac{1}{nm} V \left(\frac{1}{n} \sum_{i=1}^n \frac{M_i \sigma_i}{P_i} \right) > 0$$

9.17 $\hat{Y} = \sum_{i=1}^n \frac{t_i}{n}$, $t_i = \sum_{t=1}^{i-1} M_{t'} y_{t'} + \frac{M_i}{P_i} y_i \left(1 - \sum_{t'=1}^{i-1} \frac{P_{t'}}{P_i} \right)$,

$$v(\hat{Y}) = \sum_{i=1}^n \frac{(t_i - \hat{Y})^2}{n(n-1)}$$

9.18 (a) $E(r) = nP$, $V(r) = n(N+n-1) \frac{PQ}{N}$, $v(r) = \frac{(N+n-1)r(n-r)}{(n-1)(N-1)}$,

(b) $E(r) = nP$, $V(r) = nPQ$, $v(r) = r(n-r)/(n-1)$

9.19 (a) $V_1 = \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nm}$, (b) $V_2 = \frac{\sigma^2}{n m'} \{1 + (m-1)\rho_{cm'}\}$,

$$V_1 = V_2 \text{ if } m' = 1 + (m-1)\{\sigma_b^2 / (\sigma^2 \rho_{cm'})\}$$

Chapter 10

- 10.1 $V(\hat{Y}) = 3523$; $M(\hat{Y}_R) = 105.2$; $\text{Eff}(\hat{Y}_R | \hat{Y}) = 3349\%$.
- 10.2 $\hat{R} = 12.11\%$; $c(\hat{R}) = 7.23\%$.
- 10.3 (i) $\hat{Y} = 20276$ acres, $c(\hat{Y}) = 2.74\%$; (ii) $\text{Eff} = 473\%$.
- 10.4 151%.
- 10.5 $\hat{Y}_R = 11010398$ acres; $c(\hat{Y}_R) = 0.32\%$; $\text{eff} = 5930\%$.
- 10.6 $\hat{Y}_R = \text{Rs. } 366433$; $c(\hat{Y}_R) = 2.73\%$; $\text{Eff} = 130\%$.
- 10.7 (i) $V(y_t) = \frac{1}{n P_i^2} \left(\sigma_i''^2 - P_i Q_i \bar{Y}_i^2 \right)$, where $\sigma_i''^2$ is the variance of y with y being 0 for units not belonging to the i -th class, P_i is the proportion of units in that class and $Q_i = 1 - P_i$.
(ii) $V(\bar{y}_t)$ in (i) is to be multiplied by $(N-n)/(N-1)$.
- 10.11 (i) $\frac{\alpha^2}{\beta^2} < \frac{C_x^2}{n} \cdot \frac{(1-f)}{V(1/\bar{x})}$; (ii) $\rho^2 > 1 - \frac{(1-f)}{\bar{X}^2 E(1/\bar{x}^2)} + n \frac{\alpha^2 V(1/\bar{x})}{\sigma_y^2 E(1/\bar{x}^2)}$, $f = \frac{n}{N}$.
- 10.12 $E(V_{rat}) < V(V_{per})$ if $\frac{\bar{X}}{N} \sum_{i=1}^N \frac{1}{X_i} > 1 + C_x^2 \left\{ 1 - \frac{N-1}{N} \left(1 + \frac{\alpha^2}{a} \right)^{-1} \right\}$.
- 10.14 (i) $V\left(\frac{p_2}{p_1}\right) = \frac{N-n}{N-1} \cdot \frac{P_2(P_1-P_2)}{n P_1^3}$; (iii) $P_1 < \frac{2P_2}{1+P_2}$.
- 10.20 $v(\hat{Y}_1) = (1-f) \frac{N^2 \bar{y}_1^2}{n} \sum_{i=1}^n \left\{ \frac{y_{1i}}{\bar{y}_1} - \frac{x_{1i}}{\bar{x}_1} - \frac{y_{2i}}{\bar{y}_2} + \frac{x_{2i}}{\bar{x}_2} \right\}^2$, $f = \frac{n}{N}$.

Chapter 11

- 11.1 $M(\hat{Y}_r) = 1112$; $E(\hat{Y}_r | \hat{Y}) = 317\%$; $\text{Eff}(\hat{Y}_r | \hat{Y}_R) = 9.46\%$.
- 11.2 cf. (11.23), p. 441; $n_1 = 463$, $n_2 = 167$, $C(\hat{Y}) = 6.41\%$.
- 11.3 105%.
- 11.4 $\text{Eff}(\hat{Y}_r | \hat{Y}_R) = 108\%$; $\text{Eff}(\hat{Y}_r | \hat{Y}) = 141\%$.
- 11.5 $E\{V(\hat{Y}_r)\} < E\{V(\hat{Y}_R)\}$.
- 11.8 $\hat{Y}_r = \hat{Y}_2 + \hat{\beta}(\hat{X}_1 - \hat{X}_2)$, $\hat{Y}_2 = \frac{G}{n_2 m} \sum_{i=1}^{n_2} \frac{a_i}{g_i} \sum_{j=1}^m y_{ij}$,
 $\hat{X}_2 = \frac{G}{n_2} \sum_{i=1}^{n_2} \frac{b_i}{g_i}$, $\hat{X}_1 = \frac{G}{n_1} \sum_{i=1}^{n_1} \frac{b_i}{g_i}$;
cf. (11.23) p. 414 and Sub-section 9.4, pp. 328-329.

11.9 $\rho^2 > 4C_1C_2/(C_1+C_2)^2$

11.11 $\hat{\bar{Y}}_r = \frac{[n'(\bar{y}_1 + \hat{\beta}(\bar{x} - \bar{x}_1)) + n_2\bar{y}_2]}{n + n_2}, \quad \frac{1}{n'} = \frac{\rho^2}{n} + \frac{1-\rho^2}{n_1};$

$$V(\hat{\bar{Y}}_r) = \frac{\sigma_y^2}{(n'+n_2)}$$

11.12 $\lambda_1 = -\beta\alpha'\delta, \quad \lambda_2 = \alpha\delta, \quad \lambda_3 = -\lambda_1, \quad \lambda_4 = 1-\lambda_2,$

where $\alpha' = 1-\alpha, \quad \delta = 1/(1-\alpha'^2\rho^2), \quad \alpha = \sqrt{1-\rho^2}/\{1+\sqrt{1-\rho^2}\};$

$$V(\hat{\bar{Y}}) = \sigma_y^2[1+\sqrt{1-\rho^2}]/2n$$

11.13 (i) $\lambda_1 = \lambda_2 = \alpha\delta', \quad \lambda_3 = \lambda_4 = \alpha(1+\rho)\delta',$

$$V(\hat{\bar{Y}}_1 + \hat{\bar{Y}}_2) = 2\sigma^2(1+\rho)/n\delta', \quad \text{where } \delta' = 1/(1+\alpha'\rho).$$

(ii) $\lambda_1 = -\alpha/\delta'', \quad \lambda_2 = -\lambda_1, \quad \lambda_3 = -\alpha'(1-\rho)\delta'', \quad \lambda_4 = -\lambda_3,$

$$V(\hat{\bar{Y}}_2 - \hat{\bar{Y}}_1) = 2\sigma^2(1-\rho)/\delta'', \quad \text{where } \delta'' = 1/(1-\alpha'\rho)$$

Chapter 12

12.1 $\hat{Y} = 311520, \quad c(\hat{Y}) = 9.47\%, \quad \hat{R} = 4.328, \quad c(\hat{R}) = 3.52\%.$

12.2 $\hat{Y} = 229750, \quad c(\hat{Y}) = 6.97\%$

12.3 (i) 88.75, (ii) and (iii) cf. table below

stra fac tum tory				stra fac tum tory				stra fac tum tory				stra fac tum tory			
1	1	6	27	2	1	35	100	3	1	31	92	4	1	3	30
2	32	15		2	27	94		2	53	118		2	21	20	
3	7	9		3	19	115		3	16	72		3	5	26	
4	45	18		4	44	120		4	12	96		4	9	31	

12.4

stratum	interval for sample village					
	1	2	3	4	5	6
1	45	65	22	42	45	40
2	113	152	129	96	154	129
3	71	45	77	67	73	71

Common inflation factor = 7000.

12.5 cf. solution to Problem 9.8 (p. 672).

12.6 $\hat{Y} = \frac{1}{n'} \left(\sum_{i=1}^n w_i \right) \left(\sum_{j=1}^{n'} y'_j \right), \quad v(\hat{Y}) = \left(\sum_{i=1}^n w_i \right)^2 \sum_{j=1}^{n'} \frac{(y'_j - \bar{y}')^2}{n'(n'-1)},$

where $\{w_i\}$ are the inflation factors and y'_j is the value of the j -th sample unit selected for tabulation.

12.7 If the interval is $l.x$, it is to be rounded off to l or $l+1$ with probabilities $P = l(1-x)/l.x$ and $1-P$ respectively.

Chapter 13

13.1 $\hat{B} = \frac{1}{n} (c-b); \quad \hat{V} = \frac{1}{n(n-1)} \{n(b+c)-(b-c)^2\}.$

13.2 $B(\hat{\bar{Y}}) = \bar{\alpha}; \quad V(\hat{\bar{Y}}) = \frac{N-n}{Nn} \sigma_s^2 + \frac{K-k}{Kk} \sigma_a^2 + \frac{1}{nk} \sigma_e^2;$

$v(\hat{\bar{Y}})$: cf. (13.21) on p. 459.

13.3 $\hat{\bar{Y}} = \frac{1}{n} (n_1 \bar{y}_1 + n_2 \bar{y}'_2), \quad V(\hat{\bar{Y}}) = \frac{\sigma^2}{n} \left\{ 1 + (k-1)Q \frac{\sigma_b^2}{\sigma^2} \right\};$
 $k = \sqrt{\left(\frac{\sigma^2}{\sigma_b^2} - Q \right) \frac{C_3}{C_1 + C_2 P}}, \quad n = \frac{\sigma^2}{V'} \left\{ 1 + (k-1)Q \frac{\sigma_b^2}{\sigma^2} \right\};$

$n = 140, \quad k = 4.2, \quad C = \text{Rs. } 160.$

13.4 k : as in solution to problem 13.3; $n = C/(C_1 + C_2 P + C_3(Q/k))$; $rse = 10\%$.

13.5 $B(p') = P_{21}(P' - P'') = P' - P_{.1}, \quad P' = P_{11}/P_{12}, \quad P'' = P_{21}/P_{22};$
 $-P_{21}(1-P') \leq B(p') \leq P_{21}P'.$

13.6 $V(\hat{Y}) = \sum_{i=1}^s \sum_{j=1}^{N_i} \left(\frac{s}{i} - 1 \right) Y_{ij}^2; \quad v(\hat{Y}) = \sum_{i=1}^s \sum_{j=1}^{N_i} d_{ij} \frac{s}{i} \left(\frac{s}{i} - 1 \right) Y_{ij}^2.$

13.7 under emmoration = 9.66%; rse = 5%.

INDEX

- Accuracy, definition, 40
Aggregate, *see* population total
Aggregate check, 455, 470
Allocation to strata, 239-248
AOYAMA, H., 271, 285
Area sampling, 197, 207, 257, 344,
Area unit, 33, 197, 207
Ascertainment errors, 451
Auxiliary information, 32-34
 systematic sampling, 160, 165, 167,
 172
 pps sampling, 184, 215, 218
 stratified sampling, 237, 264
 ratio estimator, 369
 regression estimator, 406
- BAILEY, N. T. J., 94
BALAKRISHNAN, T. R., 262, 286
Balanced differences, 155
BANERJEE, S., 152, 173, 176
BASU, D., 66, 89
BERSHAD, M. A., 462, 476
Beta distribution, 266
BHARGAVA, R. P., 134, 176
BHATTACHARJEE, G. P., 121, 123
Bias, definition, 39, 453
 in selection, 113
 of ratio estimator, 363
 of regression estimator, 454
 see also non-sampling bias
BIRNBAUM, Z. W., 450, 476, 478
Bivariate population, 31
BOSE, C., 409, 420
BOWLEY, A. L., 243, 285
Branch sampling, 193
BROOKS, E. M., 509
BRYANT, E. C., 284, 285
BUCKLAND, W. R., 52, 134, 176
- CENTRAL STATISTICAL ORGANIZATION, 513,
 565
Centrally located sample, 164
- Chain ratio estimator, 390
CHAKRAVARTHY, I. M., 279, 286
Check-sample, 469
CHEVRY, G., 9, 20
CHURCH, B. M., 428, 444
Cluster sampling, 293
 equal clusters, 294
 unequal clusters, 306
COCHRAN, W. G., 36, 45, 52, 89, 113, 123,
 134, 155, 176, 181, 214, 226, 230,
 249, 261, 275, 281, 286, 290, 291,
 292, 319, 353, 360, 362, 397, 401,
 409, 420, 422
- Coefficient of variation
 in population 27, stability, 96
 of estimator 41, *see* relative
 standard error
Collapsed strata, 292
Colour bias, 112
Combined ratio estimator, 377, 389
Combined regression estimator, 411
Complete enumeration survey, 6, 16
Component-wise product estimator, 393
Component-wise ratio estimator, 389, 393
Composite sampling design, 352
Conditional probability, 42
 expectation, 42, variance, 42
Confidence coefficient, definition, 45
 effect of bias, 45
Confidence interval, definition, 45
 srs : σ known, 81, σ unknown, 82
Confidence limits, 45
Consistency checks, 467
Consistent estimator, 39
Content errors, 469
Continuing survey, 484
Continuous population, 24
Controlled selection, 282
Control of errors, 466
 see non-sampling errors
CORNFIELD, J., 89
Correlation coefficient, 31

- Cost aspect, 13, 206,
 Cost-efficiency, definition, 44
 of pps sampling, 190
 of cluster sampling, 312
 of two stage sampling, 341
 Cost function, general, 48, 49
 in srs, 121, in pps, 190
 in stratified sampling, 240
 in cluster sampling, 301
 in two-stage sampling, 334
 in multi subject survey, 348
 in two phase sampling, 414
 in non sampling errors, 463
 in non response, 463
 Coverage errors, 469, 477, 480
 Cumulative total method, 200

DALENIUS, T., 38, 52, 261, 264, 270, 271, 272, 274, 275, 279, 286, 291
 Data collection, methods of, 485
 Data requirements, formulation of, 482
DAS, A. C., 175, 176, 213, 225
 Deep stratification, 278
 Demarcation of strata
 theoretical solutions, 261-270
 approximations, 270-273
DEMING, W E., 48, 52, 94, 171, 177, 480,
DES RAJ, 66, 89, 93, 189, 212, 215, 222, 225, 230, 351, 353, 359, 388, 397, 402, 403, 419, 420, 421, 424
 Difference, estimation of, 290, 405
 Domain of study, 25, 77, 81, 234, 482
 Double ratio estimator 390
 Double sampling, 415
DURBIN, J., 215, 225, 231, 272, 286, 383, 397, 403, 450, 476, 478

ECKLER, A. R., 9, 20
 Efficiency, definition, 44
 srs, 101, 102
 systematic sampling, 147
 pps sampling, 186, 196, 213
 stratified sampling, 248, 252
 cluster sampling, 296, 310
 two stage sampling, 326, 340
 ratio estimator, 370
 regression estimator, 407
EKMAN, G., 245, 261, 271, 286
 Elementary unit, 23
ELKIN, J. M., 402
 Empirical studies
 srs, 83
 sample size determination, 98, 100, 102
 haphazard selection 112
 systematic sampling, 149, 151, 152, 160, 170
 pps sampling 192, 193, 196, 220
 stratified sampling, 258, 261, 272
 cluster sampling, 300, 303, 304, 311
 two-stage sampling, 344
 ratio estimator, 379, 385, 396
 self weighting design, 439
 End corrections, 163
 Equal probability sampling,
 see simple random sampling
 Estimate, definition, 38
 Equi weighting design 427
 Estimate, definition, 38
EVANS, W D., 243, 286
 Expected value, 38
 Experimental data, 3
 External record check, 472
 Family living surveys, 563
FELLEGI, I. P., 211, 223
 Field work, 497
 Finite population, 24
 Finite population correction, 72
FINNEY, D J., 134, 176
FISHER, R A., 19, 20, 104, 123
FITZPATRICK, T B., 283, 286
 Fractile graphical analysis, 501
 Frequency distribution, 27, 84, 174
FULLER, W A., 292

 Gamma distribution, 29, 266
GANGULI, M., 319, 353
GHOSH, B., 419, 420
GHOSH, S P., 279, 286, 316
GLASSER, G J., 94

- GODAMBE, V. P., 38, 52, 229
 GOODMAN, L. A., 93, 381, 384, 385, 397 403
 GOODMAN, R., 282, 286
 GRAHAM, J. E., 415, 420
 Gross error, 471, 477
 GRUNDY, P. M., 211, 215, 219, 225, 226
 GURNEY, M., 261, 270, 286, 509
- HALDANE, J. B. S., 94
 HALDAR, A., 229
- HANSEN, M. H., 36, 52, 93, 171, 177, 184, 225, 244, 271, 281, 286, 300, 313, 319, 353, 401, 415, 420, 423, 427, 444, 450, 462, 464, 476, 509
- HANURAV, T. V., 38, 52, 211, 225, 229, 231
- Haphazard selection, 111
- HARTLEY, H. O., 214, 217, 225, 226, 230, 284, 285, 384, 385, 397, 403
- HASEL, A. A., 95, 123, 132, 134, 176
- HESS, I., 261, 283, 286
- Histogram, 84, 85, 174
- HODGES, J. L., 261, 271, 272, 286
- HOLLAND, D. A., 193, 196, 226
- Horvitz, D. G., 209, 211, 225
- HURWITZ, W. N., 36, 52, 93, 184, 225, 244, 271, 281, 286, 300, 313, 319, 353, 401, 415, 420, 423, 427, 444, 462, 464, 476
- Illustrative populations
 catch of fish, 180, 356
 forest data, 95, 131, 180, 315
 industry data, 98, 228, 288, 356, 398, 446
 land holdings, 30, 91
 livestock, 100, 479
 urban population, 90
 village data, 91, 95, 98, 127, 178, 227, 287, 399, 400, 447
- INDIAN STATISTICAL INSTITUTE, 509, 513, 565
- Infinite population, 24
- Inflation factor, 425
- Information, definition, 44
- Integration of surveys, 220, 504, 529
- Interpenetrating sub-samples
 in general, 47
- srs, 64
 systematic sampling, 158
 stratified sampling, 277
 multi-stage sampling, 359
 ratio estimator, 365, 368, 381
 regression estimator, 408
 non-sampling errors, 474
 in planning surveys, 495
 National Sample Survey, 515
 Family Living Surveys, 571
- Intraclass correlation coefficient, 49, 146, 296, 308, 459
- Intra-investigator correlation, 459
- Investigation zones, 494, 535
- JABINE, T. B., 36, 52
- JESSEN R. J., 193, 196, 225, 284, 285, 300, 301, 313
- Judgement sample, 37
- KEMSLEY, W. F. F., 509
- KENDALL, M. G., 52, 104, 123
- KEYFITZ, N., 221, 225, 392
- KHAMIS, H. S., 66, 89, 93
- KISH, L., 282, 286
- KOOP, J. C., 38, 52, 402
- KULLDOFF, G., 409, 420
- LABOUR BUREAU, 565
- LAHIRI, D. B., 16, 21, 134, 139, 152, 171, 173, 176, 181, 202, 218, 222, 225, 232, 319, 353, 397, 428, 444, 450, 455, 476, 479, 480
- Linear trend
 in systematic sampling, 160
 in pps sampling, 188
- List frame, 34
- MADOW, L. H., 134, 176, 181
- MADOW, W. G., 93, 134, 176, 181, 215, 225, 244, 271, 281, 286, 315, 423, 428, 444
- MAHALANOBIS, P. C., 16, 21, 36, 48, 52, 184, 225, 242, 243, 244, 261, 271, 286, 300, 301, 304, 313, 314, 319, 353, 409, 420, 450, 474, 476, 479, 502, 509

- Mail enquiry, 487
 Map frame, 34, 35
 MARKS, E S, 476
 MATTHAI, A, 28, 46, 52, 104, 109, 123, 124
 MAULDIN, W P, 476
 McHUGH, R B, 124
 Mean square error, definition, 40
 Measures of error, 40
 MICKEY, M R, 384, 397, 409, 420
 Middle class survey, 582
 MIDZUNO, H, 218, 226, 230, 387, 397
 MITRA, S K, 28, 46, 52, 104, 124
 MOSTELLER, F, 36, 52
 Multi phase sampling, definition, 41
 for pps sampling, 351
 for stratification, 360
 for ratio estimator, 394
 for regression estimator, 412
 Multiple characteristics
 systematic sampling, 167
 ratio estimator, 392, 404
 regression estimator, 419, 424
 product estimator, 392
 multiple sampling *see*
 multi phase sampling
 Multiple stratification, 278
 Multiplier, 425, 548
 Multi stage sampling, 41, 317, 321
 two stage sampling, 317, 319, 322
 three stage sampling, 345
 Multi subject survey, 182, 347, 504, 513
 Multi variable product estimator, 392
 Multi variable ratio estimator, 392
 Multi variable regression estimator, 419
 MURTHY, M N, 38, 39, 52, 83, 89, 119,
 124, 213, 215, 226, 230, 277, 286, 290,
 359, 381, 383, 386, 397, 409, 420 444
 446 448, 450 473, 476
 NAIR, K R, 134, 176
 NANJAMMA, N S, 383, 386, 397
 NARAIN, R D, 211, 226
 NATIONAL SAMPLE SURVEY (NSS), 30, 140,
 166, 169, 215, 319, 511, 563
 Nested sampling, 319
 Net error, 471, 477
 NEYMAN, J, 242, 286
 NIETO DE PASCUAL, J, 385, 397
 NISSELSON, H, 415, 420
 Non normal distribution, 83
 Non random sample, 37
 Non response error, 463
 Non sampling bias 44 452, 454
 Non sampling errors, 11, 43, 449
 Non sampling variance, 44, 452, 456
 Normal deviate, 45
 Normal distribution, 27, 81, 82, 266
 Not at home, problem of, 463, 466, 478
 Notations, 50-51
 Number bias, 111
 Number of strata,
 determination of, 274
 OLKIN, I, 389, 392, 397
 Optimum cluster size, 299, 303
 Optimum sample size, 15
 Optimum size of fsu, 340
 Optimum stratification,
 see demarcation of strata
 Ordered estimator, 213
 PANSE, V G, 134, 177, 180, 428 444
 Parameter, definition, 25
 Parent distribution, 81
 PARTHASARATHY, G, 417, 420
 PATHAK, P K, 66, 89, 93, 199, 226, 230
 PATTERSON, H D, 409 420, 423
 Payroll sampling method, 580
 PEARCE, S C, 193, 196, 226
 Periodic variation, 169
 Permanent survey organization, 507
 Permissible error, 8, 113, 483
 Personal interview, 486
 Pilot survey, 496, 572, 586
 Plane systematic sampling, 175
 POLITZ, A V, 466, 476
 Pooling of estimates, 86, 88
 Population: definition, 6, 24
 Population mean, 26

- Population total, 26, estimation of, srs, 76
 pps sampling, 185, 209
 stratified sampling, 235
 two-stage sampling, 322
 three-stage sampling, 345
 ratio estimator, 369
 regression estimator, 406
 Post-cluster sampling, 316
 Post-enumeration survey, 455, 469
 Post-stratification, 280, 292
 Post-stratified ratio estimator, 389
 POTR, J., 152, 173, 176
 Precision, definition, 40, 44
 Probability proportional to size sampling
 with replacement, 185
 without replacement, 209
 systematically, 215
 extension of pps, 223
 in stratified sampling, 255
 in cluster sampling, 309
 in two-phase sampling, 351
 in two-stage sampling, 328
 in three-stage sampling, 345
 in ratio estimation, 375
 in self-weighting design, 438
 Probability proportional to total size, 218
 Probability sample, 36
 Probability sampling scheme, 36
 Processing of survey data, 499
 Product method of estimation, 380
 Proportion, 26, estimation of,
 srs, 78
 confidence interval, 84
 sampling distribution, 85
 sample size, 122
 systematic sampling, 172
 stratified sampling, 253
 cluster sampling, 305
 changes in, 417
 non-sampling errors, 461
 Purposive sample, 37
 Quality control techniques, 472
- QUENOUILLE, M. H., 175, 176, 181, 383, 397, 403
 Questionnaire, 488
- Raising factor, 425
 RAND CORPORATION, 104, 124
 Random group method, 214
 Random inspection, 498
 Random number tables, 103, 657
 Random sample, 36
 Random start, 133
 Random variable, 38
 Randomized rounded-off weights, 437, 448
 RANGARAJAN, R., 336, 353, 358
 RAO, C. R., 28, 46, 52, 104, 124
 RAO, J. N. K., 214, 217, 225, 226, 230, 336, 353, 359, 415, 420
 RAO, T. J., 291, 385, 397, 403
 Rare item, sampling for, 94
 Rate of change, estimation of, 417
 Ratio, 31
 Ratio-cum-product estimator, 393
 Ratio estimator, 361
 bias of, 363
 almost unbiased, 381
 unbiased, 386
 Ratio method of estimation, 362, 369
 Ratio-type estimator, 383
 Recall error, 469, 473
 Reconciliation check survey, 455, 470
 Reference period, 490
 Registration, method of, 487
 REGISTRAR GENERAL, 9, 21
 Regression coefficient, 31, 411
 Regression estimator, 407,
 bias, 408
 variance, 409
 Relative standard error, 41
 Relativ variance, 27, 41
 Repetitive survey, 484
 Reporting unit, 23
 Reports, preparation of, 502
 Reporting period, 490, 514
 Response variance, 458
 RIEDEL, D. C., 283, 286

- ROBSON, D S , 384, 389, 393, 397
 Root mean square error, 40
 Ross, A , 384, 397
 Rotation sampling, 415
 Rounding off of weights, 436
 ROY, J , 319, 354
 ROY CHOWDHURY, D K , 213, 223, 226
- SAMPFORD, M R , 231
 Sample, definition, 10, 36
 Sample check, 468
 Sample design, definition, 37
 Sample observation, 37
 Sample registration, 488
 Sample size, definition, 12, 37
 determination of, 99, 113 122, 173,
 522, 589, 585
 Sample survey, need for, 6, 10, 16
 Sample unit, 37
 Sampling design, choice of, 493
 Sampling distribution, 45, 81, 85, 174
 Sampling efficiency, *see* efficiency
 Sampling error, 11, 73, 101
 Sampling fraction, 37
 Sampling frame, 17, 32
 specimens, 34, 35
 changes in, 110, 221
 minimal frame, 492
 master frame, 493
 Sampling interval, 133
 fractional, 141
 varying, 177
 Sampling on successive occasions, 415
 Sampling procedure
 srs, 103
 systematic sampling, 142
 pps sampling, 200
 selection in stages, 208
 Sampling unit, definition, 31
 Sampling variance, definition, 40
 srs, 61, 70, 79
 systematic sampling 144, 148
 pps sampling, 185, 210, 214
 stratified sampling, 239, 246, 250, 253,
 256
- cluster sampling, 295, 305, 307, 309
 two phase sampling, 351,
 two stage sampling, 323, 325, 329
 three stage sampling, 346
 ratio estimator, 369, 369, 376, 377
 regression estimator, 408, 411
 self weighting design, 438
- SASTRY, K V R , 134, 177, 180
 Scatter diagram, 161, 194
 Schedule, 488
 SEAL, K C , 36, 52, 409, 421
 SEARLS, D T , 93
 SEKHAR, C C , 480
 Selection, procedure,
 see sampling procedure
 Self validation, 478
 Self weighting design, 425
 field stage, 427, 428
 tabulation stage, 436
 SEN, A R , 211, 218, 226, 229, 387, 397
 SENG, Y P , 11, 21
 SENGUPTA, J M , 301, 313, 450, 476
 Separate ratio estimator, 376
 Separate regression estimator, 411
 Serpentine method, 222, 232
 SETH, G R , 38, 52, 292, 450, 476
 SETHI V K , 166, 177, 188, 215, 226, 261,
 262, 265, 266, 267, 269, 271, 272, 275,
 286, 386, 397, 444, 446, 448
 SHUKLA, N D , 230
 SIMMONS, W R , 466, 476
 Simple random sampling, 55
 with replacement, 59, 60
 without replacement 67, 69
 stratified sampling, 246, 250
 cluster sampling, 295, 305, 306
 two stage sampling, 322, 328
 ratio estimator, 371, 387
 regression estimator, 410
 non sampling error, 457
 non response, 463
 SINGH, D , 319, 354, 360
 SINGH, M P , 393, 394, 397, 398, 404
 SIRKEN, M G , 450, 476, 478
 SMITH, B B , 104, 123

- SMITH, H. F., 300, 313
 SOM, R. K., 428, 444, 474, 476
 Specification errors, 451
 SRIKANTAN, K. S., 54
 Stages of randomization, 41
 Standard deviation, definition, 27
 Standard error, definition, 41
 Statistic, 38
 Statistical analysis, 499
 Statistical information, 1
 Statistical quality control techniques, 472
 STEIN, C., 120, 124
 STEINBERG, J., 415, 420
 STEPHAN, F. F., 171, 177
 STEVENS, W. L., 230
 Stratification, need for, 233
 Stratification variable, 233, 237, 264
 Stratified sampling, 233
 srs, 246
 pps sampling, 255
 ratio estimator, 376, 387
 regression estimator, 411
 self-weighting design, 428
 STUART, A., 242, 287
 Student's *t*-distribution, 46
 Sub-frame, 182, 234, 495
 Subject analysis, 499
 Sub-population, *see* domain of study
 SUBRAHMANYA, M. T., 230
 Sub-units, 187, 217
 SUKILATME, P. V., 38, 52, 134, 177, 180,
 243, 287, 300, 301, 313, 315, 401,
 428, 444, 450, 476, 477
 Summarization of data, 499
 Super-population, 181, 188, 229, 290, 291,
 401, 402
 Survey data, 3
 Survey period, 490
 SWAIN, A. K. P. C., 374, 398
 Systematic sampling
 linear, 133, 144
 circular, 139, 155
 balanced, 165
 use in forest survey, 149
 use in census, 151
 for catch of fish, 180
 with pps, 216
 Tabulation errors, 451
 Target population, 25
 Tenement sampling, 574, 588
 THOMPSON, D. J., 209, 211, 225
 TIKKIWAL, B. D., 409, 421
 TIPPETT, L. H. C., 104, 124
 Transcription from records, 488
 TURKEY, J. W., 36, 52
 Two-dimensional systematic
 sampling, 175, 182
 Types of data, 3
 Unbiased estimator, 38
 Unconditional probability, 42
 Uncorrelated response variance, 458
 Unit, 23
 Unit of analysis, 24
 Unitary check, 455, 470
 UNITED NATIONS, 20, 21, 477, 502, 504, 510
 UNITED STATES BUREAU OF THE
 CENSUS, 473, 477, 510
 Universe, 6, 10, 24
 Unordered estimator, 213
 Variance, *see* sampling variance
 Variance components in two-stage
 sampling, 333
 Variance estimator, *see* Appendix 2, p. 659
 Variance function
 general, 48, 49
 stratified sampling, 239
 cluster sampling, 299
 two-stage sampling, 329
 see sampling variance
 Varying probability sampling, *see*
 probability proportional to size
 sampling
 VIJAYAN, K., 229
 VITHYASAI, C., 389, 393, 397

- WATSOV, D J , 113, 123
Weight, 425
WILLIAMS, W H , 281, 287, 384, 398, 403
Working class survey, 566
- YATES, F , 16, 21, 36, 52, 104, 113, 123,
124, 134, 155, 164, 177, 211, 219, 226,
392, 398, 403
- ZARKOVICH, S S , 16, 21, 100, 124, 189,
226, 409, 421, 450, 477, 510
-