CONTRIBUTIONS TO SAMPLING THEORY AND PRACTICE USING AUXILIARY INFORMATION (INDIA).

TALAPADY. SRIVENKATARAMANA

University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Recommended Citation
SRIVENKATARAMANA, TALAPADY, "CONTRIBUTIONS TO SAMPLING THEORY AND PRACTICE USING AUXILIARY INFORMATION (INDIA)." (1982). Electronic Theses and Dissertations. 1569.
https://scholar.uwindsor.ca/etd/1569
NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

LA THÈSE A ÉTÉ MICROFILMÉE TELLE QUE NOUS L'AVONS REÇUE

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED
CONTRIBUTIONS TO SAMPLING THEORY AND PRACTICE
USING AUXILIARY INFORMATION

by

C. Talapady Srivenkataramana

A Dissertation
Submitted to the Faculty of Graduate Studies through the
Department of Mathematics in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada
1982
Dedicated to my parents

Shri T. Madhava Bhat and Shrimati T. Indiramma
ACKNOWLEDGEMENTS

The author is greatly indebted to Dr. Derrick S. Tracy for his invaluable guidance in the research work carried out. There was always a personal touch, encouragement and affection in the relationship, besides academic guidance. The benefit derived extends much beyond the preparation of this dissertation.

Thanks are due to the Bangalore University for giving leave of absence and the University of Windsor for the financial support during the work. The author is particularly grateful to the International Development Research Centre of Canada for the Thesis Research Award during 1979-82, which supported the survey work reported in Part Two of the thesis.

Special thanks are due to Mr. Narasimha Prasad for many useful discussions regarding the material in Chapters 5 and 6 and to Mr. K. Harish Chandra for the help in the field-work and computer analysis of the survey data. The author wishes to express his appreciation of the patience exhibited by his wife and children during the entire programme. The great encouragement given by his parents, brothers and sisters instilled the necessary confidence. The help obtained from the members of the Mathematics Department at the University of Windsor and, in particular, its Chairman Dr. H.R. Atkinson are sincerely acknowledged.
ABSTRACT

Part One (Chapters 2-7) of the thesis illustrates mainly the use of auxiliary information in effecting certain transformations for improving the design/estimation strategies in sample surveys. A new product-type estimator which is complementary to the usual ratio estimator, in a certain sense, is proposed in Chapter 2. A transformation which permits the use of the product method in place of the more commonly used ratio method is discussed in the next chapter. One resulting convenience is that the bias and mean squared error of the estimator have closed form expressions, unlike that for the ratio estimator. Chapter 4 develops a change of origin for the study variate in order to improve the precision of the estimates under varying probability sampling. A change of origin under ratio method of estimation and a change of scale under the difference method are also examined. Raj (1965) has proposed a two-phase pps selection scheme in the absence of information on the auxiliary variate, but when information on some other variate is available. This also requires the knowledge of a certain parameter. Chapter 5 suggests an alternative two-phase procedure not requiring the knowledge of this parameter.

Auxiliary variates in surveys are occasionally positive and negative valued. This introduces difficulties in ratio and product methods of estimation and in pps sampling. Chapter 6 outlines two simple methods of dealing with the situation: stratification by sign of the auxiliary variate and transformation by simple translation. For the purposes of illustration simple random sampling without replacement and probability
proportional to size sampling with replacement are assumed. Finally, the scope for some future work is indicated in Chapter 7.

Part Two (Chapters 8-9) of the thesis is a brief report on a sample survey conducted during 1980-81 to assess the transition of rural people to modern ways of life in the Karnataka State in South India.
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEDICATION</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>CHAPTER 1: GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>PART ONE: THEORY</td>
<td></td>
</tr>
<tr>
<td>SOME NOTATION AND ABBREVIATIONS</td>
<td>6</td>
</tr>
<tr>
<td>CHAPTER 2: A DUAL TO RATIO ESTIMATORS</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Bias and MSE of ( \hat{Y}_a )</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Efficiency Comparisons</td>
<td>11</td>
</tr>
<tr>
<td>2.4 Unbiased Estimators</td>
<td>13</td>
</tr>
<tr>
<td>2.5 Empirical Study</td>
<td>15</td>
</tr>
<tr>
<td>2.6 Optimality of the Estimators</td>
<td>19</td>
</tr>
<tr>
<td>2.7 Remarks</td>
<td>20</td>
</tr>
<tr>
<td>CHAPTER 3: EXTENDING PRODUCT METHOD OF ESTIMATION TO THE POSITIVE CORRELATION CASE</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>22</td>
</tr>
<tr>
<td>3.2 The Choice of ( L )</td>
<td>23</td>
</tr>
<tr>
<td>3.3 Case of Negative Correlation</td>
<td>27</td>
</tr>
<tr>
<td>3.4 Use of multiauxiliary Information</td>
<td>29</td>
</tr>
<tr>
<td>3.5 Empirical Performance of ( \hat{Y}_p )</td>
<td>32</td>
</tr>
</tbody>
</table>
## CHAPTER 4: VARIATE TRANSFORMATIONS AFTER SAMPLING

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>36</td>
</tr>
<tr>
<td>4.2 PPSWR Sampling</td>
<td>36</td>
</tr>
<tr>
<td>4.3 The Choice of $a$</td>
<td>37</td>
</tr>
<tr>
<td>4.4 Empirical Efficiency of $Y^*$</td>
<td>39</td>
</tr>
<tr>
<td>4.5 PPSWOR Sampling</td>
<td>43</td>
</tr>
<tr>
<td>4.6 A Modification of $Y_{HT}$</td>
<td>43</td>
</tr>
<tr>
<td>4.7 Choice of $b$</td>
<td>45</td>
</tr>
<tr>
<td>4.8 Empirical Study</td>
<td>46</td>
</tr>
<tr>
<td>4.9 Ratio Method of Estimation</td>
<td>48</td>
</tr>
<tr>
<td>4.10 Difference Method of Estimation</td>
<td>50</td>
</tr>
</tbody>
</table>

## CHAPTER 5: DOUBLE SAMPLING WITH PPS SELECTION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction</td>
<td>52</td>
</tr>
<tr>
<td>5.2 An Alternative Scheme</td>
<td>53</td>
</tr>
<tr>
<td>5.3 Efficiency of $Y^*$</td>
<td>56</td>
</tr>
<tr>
<td>5.4 Cost Function</td>
<td>57</td>
</tr>
</tbody>
</table>

## CHAPTER 6: USE OF A POSITIVE AND NEGATIVE VALUED AUXILIARY VARIATE IN SURVEYS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Introduction</td>
<td>60</td>
</tr>
<tr>
<td>6.2 The Stratification Approach</td>
<td>61</td>
</tr>
<tr>
<td>6.3 A Change of Origin</td>
<td>62</td>
</tr>
<tr>
<td>6.4 The Choice of $\theta$</td>
<td>65</td>
</tr>
<tr>
<td>6.5 PPS Sampling</td>
<td>65</td>
</tr>
<tr>
<td>CHAPTER 7: SCOPE FOR FURTHER WORK</td>
<td>Page</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.1 General Points</td>
<td>69</td>
</tr>
<tr>
<td>7.2 The Predictive Approach</td>
<td>71</td>
</tr>
<tr>
<td>7.2.1 Use of multiauxiliary information</td>
<td>75</td>
</tr>
<tr>
<td>7.2.2 Mixing estimators</td>
<td>77</td>
</tr>
<tr>
<td>7.3 Orthogonal Auxiliary Variates</td>
<td>78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PART TWO: PRACTICE - AN ACTUAL SURVEY</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 8: BACKGROUND AND SURVEY DESIGN</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1 Introduction</td>
<td>80</td>
</tr>
<tr>
<td>8.2 The Land and the People</td>
<td>81</td>
</tr>
<tr>
<td>8.3 Sample Selection</td>
<td>83</td>
</tr>
<tr>
<td>8.4 The Questionnaire</td>
<td>85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 9: FINDINGS OF THE SURVEY</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1 Quality of Data</td>
<td>86</td>
</tr>
<tr>
<td>9.2 Results of the Survey</td>
<td>87</td>
</tr>
<tr>
<td>9.2.1 Housetypes</td>
<td>87</td>
</tr>
<tr>
<td>9.2.2 Household conveniences</td>
<td>88</td>
</tr>
<tr>
<td>9.2.3 Household change</td>
<td>88</td>
</tr>
<tr>
<td>9.2.4 Outlook transition</td>
<td>89</td>
</tr>
<tr>
<td>9.2.5 Individual transition</td>
<td>90</td>
</tr>
<tr>
<td>9.2.6 Farm transition</td>
<td>92</td>
</tr>
<tr>
<td>9.2.7 Village transition</td>
<td>93</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>9.3 Factors Causing Transition</td>
<td>95</td>
</tr>
<tr>
<td>9.4 Some Limitations</td>
<td>98</td>
</tr>
<tr>
<td>9.5 Follow-Up Studies</td>
<td>99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPENDIX QUESTIONNAIRE</td>
<td>100</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>110</td>
</tr>
<tr>
<td>VITA AUCTORIS</td>
<td>113</td>
</tr>
</tbody>
</table>
GENERAL INTRODUCTION

In sample surveys emphasis is laid on the use of supplementary information for improving the precision of estimators. Such information can be incorporated either into the sampling scheme as in stratified or pps sampling, or into the estimation procedure as in the ratio and product methods or in both. Vast literature is available in standard texts and research journals in this regard. Tremendous advances have been made in methods of sampling, formation of estimators and evaluating their performance in repeated sampling which includes computer studies when the situation is difficult to be tackled directly. Attention has been paid to reducing the bias of the estimators, obtaining stable variance estimators and evolving methods which are simple for application. Among the milestones are the papers by Neyman (1934), Cochran (1942), Hansen and Hurwitz (1943), and Horvitz and Thompson (1952).

To most people in sample surveys Neyman's 1934 paper is remembered for the derivation of the formula for optimal allocation in stratified sampling. But of far greater importance to statistics in general and to sample survey theory in particular is Neyman's exposition of the logic of inference based on confidence intervals. Ratio and regression methods of estimation were introduced during the 1930s with a comprehensive account of the theory being provided by Cochran (1942). Hansen and Hurwitz (1943) introduced probability proportional to size sampling as an efficient way of
carrying out multi-stage sampling. Horvitz and Thompson (1952) provided an elegant treatment of variable probability sampling and also gave the impetus for new research at the foundations level.

Auxiliary information usually provides a basis for stratification, gives a measure of size or helps to form ratio, product or regression-type estimators. Sometimes such information is adequate for assessing the values of certain parameters or asserting that they lie in specified intervals like (1/2, 1) or (1,2). These parameters may be built into the estimators or the idea about the magnitudes of the parameters may be used to choose from a family of estimators. A number of papers have been published in this regard.

Auxiliary information may also be used for effecting certain transformations. For instance, transformations which (i) enable the use of a specific method of estimation rather than another, or (ii) make the situation satisfy certain assumptions more closely than otherwise. If these transformations are simple, they will also be well suited for large scale applications. In this context a few methods have been proposed by Srivenkataramana (1976,1980), and Srivenkataramana and Tracy (1979, 1980a,b,c). The present thesis is in the form of some contributions to sampling theory and practice using auxiliary information. Part One, Theory, compactly strings together the results in the above papers by Srivenkataramana and Tracy after filling up gaps here and there. It then tackles two new aspects. Part Two, Practice, is a brief report on an actual sample survey conducted in the villages of Karnataka State (India) during 1980-81. A brief outline of the thesis is as follows:

Chapter 2 proposes a new product-type estimator which is complementary to the common ratio estimator, for small sampling fractions.
The next chapter points out that the product method considered by Robson (1957) and Murthy (1964) has the advantage that the bias and mse of the estimator can be evaluated exactly, unlike in the case of the ratio estimator. Motivated by this, a transformation which permits the product method in place of the ratio method is discussed. Chapter 4 illustrates a few simple transformations of data after they have been collected, in order to improve the precision of estimates. Chapter 5 considers double sampling with pps selection. An alternative to the Raj (1965) scheme is suggested. The less commonly discussed situation of using positive and negative valued auxiliary variates in surveys is taken up in Chapter 6. Two methods of overcoming the associated difficulties are considered. Chapter 7 indicates some possibilities for future work.

Part Two of the thesis comprises two chapters (Chapters 8-9), giving a brief report on a sample survey conducted to assess the transition of rural people to modern ways of life in the Karnataka State in South India. The background and survey design are explained in Chapter 8. The last chapter, has a brief discussion on the findings.
PART ONE : THEORY
1. **Notation**

   - $U_1, \ldots, U_N$ : Units of a finite population.
   - $y, x$ : Variate of interest and an auxiliary variate.
   - $Y, X$ : Population totals of $y$ and $x$.
   - $\overline{Y}, \overline{X}$ : Population means of $y$ and $x$.
   - $\bar{y}, \bar{x}$ : Sample means of $y$ and $x$.
   - $\hat{Y}, \hat{X}$ : Unbiased estimators of $Y$ and $X$.
   - $V_{ij}$ : Relative central moments of the joint sampling distribution of $(\hat{Y}, \hat{X})$.
   - $\rho$ : Correlation between $\hat{Y}$ and $\hat{X}$.
   - $k : V_{11}/V_{02} = \rho \sqrt{V_{00}/V_{02}}$

2. **Abbreviations**

   - fpc : Finite population correction.
   - mse : Mean squared error.
   - ppswor : Probability proportional to size without replacement.
   - ppswr : Probability proportional to size with replacement.
   - srswor : Simple random sampling without replacement.
   - srswr : Simple random sampling with replacement.
CHAPTER 2.

A DUAL TO RATIO ESTIMATORS

[This chapter proposes a new product-type estimator which is complementary, in a certain sense, to the commonly used ratio estimator. Exact expressions for the bias and mse can be derived which is not the case for the ratio estimator. The correction of the new estimator for bias is examined.]

2.1. Introduction

Consider a finite population of \( N \) units \( U_1, \ldots, U_N \). Let the variate of interest, \( y \), and an auxiliary variate, \( x \), related to \( y \) take real values \( y_i, x_i \) on \( U_i \), \( i = 1, \ldots, N \). First assume \( y_i, x_i \geq 0 \), since survey variates are generally non-negative with only occasional exceptions like saving and profit. Section 2.7 and Chapter 6 cover such cases. Let \( \hat{Y}, \hat{X} \) be unbiased estimators of the population totals \( Y \) and \( X \) corresponding to the variates \( y \) and \( x \) respectively, based on any probability sampling design. It is assumed that \( \hat{X} \) is known and \( \hat{X}, \hat{X} \) are positive. Also let the coefficient of correlation \( \rho \) between \( \hat{Y} \) and \( \hat{X} \) be positive. Then the traditional ratio method estimates \( Y \) by

\[
\hat{Y}_r = \frac{\hat{Y}}{\hat{X}} \quad \hat{X}.
\]

The bias and mse of \( \hat{Y}_r \) are, up to second order moments
\begin{align}
\hat{B}(Y) &= Y(1-k) V_{02}, \\
M_h(Y) &= Y^2 [V_{00} + (1-2k)V_{02}], \\
\end{align}

(2.2) \quad (2.3)

where \( V_{ij} \) are the relative central moments defined by

\begin{align}
V_{ij} &= E(Y-Y)^i (X-X)^j / Y^i X^j, \\
k &= \sqrt{V_{11} / V_{02}} = \rho \sqrt{V_{20} / V_{02}}
\end{align}

(2.4) \quad (2.5)

and \( E \) denotes averaging over all possible samples.

Let \( n < N \) be the sample size. Then

\[
\hat{X} = (N - n \bar{X}) / (N - n)
\]

(2.6)

is also unbiased for \( X \), and \( \text{cor}(Y, \hat{X}^*) = - \text{cor}(Y, \hat{X}) = - \rho \). An interpretation of \( \hat{X}^* \) is given in Section 2.7. Since \( \hat{X}^* \) is negatively correlated with \( \hat{Y} \), we form a product-type estimator based on \( \hat{X}^* \).

Note that \( \hat{X}^* \) can be easily computed once \( \hat{X} \) is known. Therefore as an estimator of \( Y \) consider

\[
\hat{Y}_a = \hat{Y}(\hat{X} / \hat{X}^*)
\]

(2.7)

### 2.2 Bias and MSE of \( \hat{Y}_a \)

Write \( Y = Y(1 + e_1) \), \( X = X(1 + e_2) \) with \( E(e_1) = E(e_2) = 0 \).

Then

\[
\hat{X}^* = (N - n \bar{X}) / (N - n)
\]

\[
= [N - nX(1 + e_2)] / (N - n)
\]

\[
= X(1 - ge_2), \text{ where } g = n / (N - n)
\]
And
\[ \hat{Y}_a = \frac{Y}{x} = Y(1+c_1)(1-g_2) = Y(1+c_1 - g_2 - ge_2). \] Hence
\[ B(\hat{Y}_a) = E(\hat{Y}_a) - Y = -gYV_{11}, \] since \[ E(c_1) = E(e_2) = 0, \]
\[ = -gYV_{11} = -gkYV_{02}. \] (2.8)

Next, the 
\[ \text{mse of } \hat{Y}_a \text{ is} \]
\[ M(\hat{Y}_a) = E(\hat{Y}_a - Y)^2 \]
\[ = Y^2E(e_1^2 + g^2e_2^2 + g^2e_1e_2)^2 \]
\[ = Y^2E(e_1^2 + 2ge_1e_2 - 2g^2e_1^2 + 2g^2e_1^2 + g^2e_1^2) \]
\[ = Y^2(V_{20} + 2V_{02} - 2gV_{11} - 2gV_{21} + 2gV_{12}^2 + g^2V_{22}) \]
\[ = Y^2[V_{20} + g(g-2k)V_{02} + 2g(gV_{12} - V_{21}) + bV_{22}]. \] (2.9)

From (2.8) we see that when \( V_{11} > 0 \) the bias is always negative.

Usually \( g \leq 0.10 \) and it is seen in (2.13) that \( \hat{Y}_a \) is preferred to
\( \hat{Y}_r \) when \( k < (1+g)/2 \). In such cases (2.8) implies that
\[ |B(\hat{Y}_a)|/Y \leq 0.055V_{02}. \] Therefore the relative bias in \( \hat{Y}_a \) is likely to be
negligible, assuming \( V_{02} \) to be small.

Now in particular consider simple or varying probability sampling
with replacement or any other scheme involving independent subsamples.
Let \( \hat{Y}_t, \hat{X}_t \) be unbiased estimators of \( Y \) and \( X \) respectively, based on
the \( t \)th selection or subsample; \( t = 1, \ldots, n \). Here \( n \) is the sample
size or the number of subsamples, as the case may be. Let \( V_{i1} \) denote
the relative central moments of the joint sampling distribution of
\( \hat{Y}_t, \hat{X}_t \). These are independent of \( t \). Also the estimators \( \hat{Y}_t, \hat{X}_t, \hat{X}_s \)
are pairwise independent for \( s \neq t \), except for the pairs \( (\hat{Y}_t, \hat{X}_t) \)
and \( (\hat{Y}_s, \hat{X}_s) \). Suppose we take \( \hat{Y} = \frac{\Sigma t}{n} \hat{Y}_t / n \), \( \hat{X} = \frac{\Sigma t}{n} \hat{X}_t / n \) as the unbiased
estimators of \( Y \) and \( X \) to be used in the estimator \( \hat{Y}_a \), and \( V_{ij} \) denote
the relative central moments of the joint distribution of \( (\hat{Y}, \hat{X}) \). Then
\[
V_{20} = E \left( \frac{\Sigma \hat{Y}^2}{n} - Y^2 \right) / \Sigma t^2 = \frac{1}{n^2} \Sigma t E (Y - Y_t)^2 = \frac{\Sigma \hat{Y}^2}{n} - Y^2 / \Sigma t^2
\]
Similarly \( V_{02} = V_{02}' / n \) and \( V_{11} = V_{11}' / n \). Next
\[
V_{21} = E \left( \frac{\Sigma \hat{Y}^2}{n} - Y^2 \right) \left( \frac{\Sigma \hat{X}}{n} - X \right) / \Sigma t^2
\]
\[
= \frac{1}{n^3} \Sigma t^2 E \left[ (Y_1 - Y) + \ldots + (Y_n - Y) \right]^2 \left[ (X_1 - X) + \ldots + (X_n - X) \right]
\]
\[
= \frac{1}{n^3} \Sigma t^2 E \left( Y_t - Y \right)^2 \left( X_t - X \right), \text{ since the other terms vanish,}
\]
\[
= \frac{V_{21}}{n^2}
\]
and similarly \( V_{12} = V_{12}' / n^2 \). Finally
\[
V_{22} = E \left( \frac{\Sigma \hat{Y}^2}{n} - Y^2 \right) \left( \frac{\Sigma \hat{X}}{n} - X \right)^2 / \Sigma t^2
\]
\[
= \frac{1}{n^4} \Sigma t^2 E \left[ (Y_1 - Y) + \ldots + (Y_n - Y) \right]^2 \left[ (X_1 - X) + \ldots + (X_n - X) \right]^2
\]
\[
= \frac{1}{n^4} \Sigma t^2 E \left[ (Y_t - Y)^2 \left( X_t - X \right)^2 + \Sigma t^2 \left( \frac{Y_t - Y}{n} \right)^2 \left( \frac{X_t - X}{n} \right)^2 \right]
\]
\[
= \frac{V_{22}}{n^4}
\]

\[ + 2 \sum_{t \neq t} (\hat{Y}_t - \hat{Y})(\hat{Y}_s - \hat{Y})(\hat{X}_t - \hat{X})(\hat{X}_s - \hat{X}) \]

\[ \frac{1}{n^4} \left[ nV_{22}' + n(n-1)V_{20}'V_{02}' + 2n(n-1)V_{11}'^2 \right] \]

\[ = \frac{1}{n} \left[ V_{22}' + (n-1)(V_{20}'V_{02}' + 2V_{11}'^2) \right] / n^3. \]

Therefore the bias and mse of \( \hat{Y}_a \) are

\[ B(\hat{Y}_a) = -gYV_{11}' / n, \]

\[ M(\hat{Y}_a) = Y^2 \left[ (V_{20}' + g(g-2k)V_{02}') / n + 2g(gV_{12}' - V_{21}') / n^2 \right. \]

\[ \left. + g^2 (V_{22}' + (n-1)(V_{20}'V_{02}' + 2V_{11}'^2)) / n^3 \right]. \]

For srswo, (2.10) and (2.11) are applicable only if \( N \) is large enough for the fpc to be ignored. These expressions show that when \( n \) is at least moderately large the contribution of bias to mse will be small, and that the terms involving \( 1/n^2 \) or \( 1/n^3 \) in the mse can be neglected, in which case (2.11) reduces to

\[ M_1(\hat{Y}_a) = Y^2 \left[ V_{20}' + g(g-2k)V_{02}' \right] / n. \]

2.3 Efficiency Comparisons

To keep comparisons tangible, the mse is taken only up to second order moments in this section. Thus

\[ M_1(\hat{Y}_a) = Y^2 \left[ V_{20}' + g(g-2k)V_{02}' \right]. \]

The estimator \( \hat{Y}_a \) is more precise than \( \hat{Y} \) when \( M_1(\hat{Y}_a) < V(\hat{Y}) = Y^2V_{20} \).
This is the case when \( g(2k)V_{02} < 0 \), which implies \( 2k > g \). On the other hand \( \hat{Y}_n \) is to be preferred to \( \hat{Y}_r \) when \( M_1(\hat{Y}_a) < M_1(\hat{Y}_r) \) from (2.3) and (2.12) we see that this needs \( g(2k) < (1 - 2k) \). That is \( 2k(1-g) < 1 - g^2 \), or \( 2k < 1 + g \), assuming \( 1 - g > 0 \). Thus \( \hat{Y}_n \) is more efficient than \( \hat{Y} \) or \( \hat{Y}_r \) when

\[
g/2 < k < (1+g)/2, \tag*{(2.13)}
\]

regarded as a condition on \( k \). While deriving this it is assumed that \( (1-g) > 0 \), that is \( n < N/2 \) which may be taken as a typical survey situation. Also it is seen that \( \hat{Y}_r \) is more efficient than \( \hat{Y} \) when \( k > 1/2 \). \tag*{(2.14)}

Since \( g \) is usually small, (2.13) and (2.14) imply that for the most part \( \hat{Y}_a \) is superior, in terms of MSE, to \( \hat{Y} \) just when \( \hat{Y}_r \) is inferior to \( \hat{Y} \). In this sense \( \hat{Y}_a \) and \( \hat{Y}_r \) are complementary. To get a simpler idea let \( V_{20} = V_{02} \) and \( n = N/5 \). Then \( k \) reduces to \( p \) and (2.13) is satisfied by any \( p \) in \((0.125, 0.625)\) while (2.14) needs \( p > 0.5 \). In general (2.13) specifies an interval of length \( 1/2 \) for \( k \). This interval slides to the right from \((0, 1/2)\) to \((1/2, 1)\) as \( n \) is increased from \( 0 \) to \( N/2 \). In the less likely case of \( n > N/2 \), (2.13) is to be replaced by

\[
k > (1+g)/2, \tag*{(2.13a)}
\]

and \( \hat{Y}_a \) becomes an alternative to \( \hat{Y}_r \). In any case the reductions in
The expression for $M_1(Y_a)$ in (2.12) reveals an unusual type of relation between the mse of $Y_a$ and sample size. For example, under srsrwor if $Y$ and $X$ are the simple expansion estimators then $M_1(Y_a)$ decreases as $n$ increases only as long as $n < Nk/(p+k)$. Thereafter $M_1(Y_a)$ increases with $n$. In general, for any sampling design where $V_{20}, V_{11}$ and $V_{02}$ are proportional to $(N-n)/n$, the dependence of the mse on $n$ will be similar.

2.4 Unbiased Estimators

It was pointed out in Sec. 2.2 that the bias in $Y_a$ is likely to be small. However, unusual situations may exist where the coefficient of variation of $X$ is large and consequently the bias becomes significant. In such cases the use of exactly unbiased estimators may be of great advantage (Rao, 1969). Then we may consider the following alternatives:

(i) We may make $Y$ and $X$ uncorrelated, so that $Y_a$ becomes unbiased as is apparent from (2.8). But this situation is not good since there will be an unacceptable increase in variance relative to $V(Y)$.

(ii) We may draw the sample in the form of $n$ independent interpenetrating subsamples. Then let $Y_i, X_i$ be unbiased estimators of $Y$ and $X$ based on
the \(i^{th}\) subsample, \(\text{cor}(\hat{Y}_1, \hat{X}_1) > 0\) and \(\hat{X}_i = (NX-n\hat{X}_i)/(N-n)\) for \(i = 1, \ldots, n\). Consider the estimators

\[
\hat{Y}_1 = \left(\sum Y_1/n\right) \left(\sum X_1/n\right)/X,
\]
\[
\hat{Y}_2 = \sum \hat{Y}_1 \hat{X}_1/(nX).
\]

Following Murthy (1964), we can show that \(B(\hat{Y}_2) = nB(\hat{Y}_1)\) and hence that

\[
\hat{Y}_3 = (n\hat{Y}_1 - \hat{Y}_2)/(n-1)
\]

is unbiased for \(Y\). The conditions for \(\hat{Y}_3\) to be more efficient than \(\hat{Y}_1\) are similar to those given by Murthy and Nanjamma (1959) in the case of obtaining an almost unbiased ratio estimator.

In particular consider srswr. Let \(\bar{x}, \bar{y}\) be the sample means and \(\hat{x} = NX/N, \hat{y} = Ny/N\). Then (2.10) reduces to \(B(\hat{Y}_a) = -N^2S_{11}/[(N-n)X]\),

where \(S_{11}\) is the population covariance

\[
S_{11} = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})/(N-1)}{N}.
\]

This bias can be estimated in an unbiased way by \(-N^2s_{11}/[(N-n)X]\), where \(S_{11}\) is the sample covariance \(\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)\). Using this to correct \(\hat{Y}_a\) for its bias, we get the estimator \(\hat{Y}_4 = \hat{Y}_a + N^2S_{11}/(KN-n)\).

A little manipulation shows that \(\hat{Y}_3\) in fact reduces to \(\hat{Y}_4\) in this case. And for srswr \(B(\hat{Y}_a) = -NS_{11}/X\), which is seen to be independent of \(n\). The corresponding unbiased estimator is \(\hat{Y}_5 = \hat{Y}_a + NS_{11}/X\).
Srivastava et al. (1981) have shown that this method of adjusting the product-type estimator for its bias by subtracting an unbiased estimate of bias (Robson, 1957) is to be generally preferred when \( x \) and \( y \) have a joint distribution with central moments \( \mu_{12} = \mu_{21} = 0 \). In other cases we may apply the technique developed by Quenouille (1956), as in the next possibility.

(iii) Take \( n = 2m \) and split the sample at random into two subsamples of \( m \) units each. Let \( \hat{Y}_1, \hat{X}_1 \) (\( i = 1, 2 \)) be unbiased estimators of \( Y \) and \( X \) based on the subsamples and \( \hat{Y}, \hat{X} \) those based on the entire sample. Take \( \hat{X}_1^* = (NX-n\hat{X}_1)/(N-n) \) and \( \hat{X}^* = (NX-n\hat{X})/(N-n) \).

And consider the product-type estimators \( \hat{Y}_{ai} = \hat{Y}_i \hat{X}_i^*/X \), \( (i = 1, 2) \) and \( \hat{Y}_a = \hat{Y} \hat{X}^*/X \). Then it can be shown that an unbiased estimator of \( Y \) is

\[
\hat{Y}_6 = (2N-n) \hat{Y}_a/N - (N-n)(\hat{Y}_{a1} + \hat{Y}_{a2})/2N .
\]  
(2.17)

The variance of \( \hat{Y}_6 \) can also be evaluated. Details are similar to those in Sukhatme and Sukhatme (1970, pp.161-5). Interestingly enough, the variance of \( \hat{Y}_6 \) and the mse of \( \hat{Y}_a \) are equal up to second order moments.

Since \( \hat{Y}_6 \) is unbiased, while \( \hat{Y}_a \) is not, the former may be preferred to the latter.

2.5 **Empirical Study**

For illustration srs is assumed in this section and \( \hat{Y} = \bar{Y} \), \( \hat{X} = \bar{X} \). First consider a hypothetical population, \( I \), of 5 units having
(4,6), (5,5), (7,7), (8,2) and (12,10) as values of \((y,x)\). Here \(Y = 36\), \(X = 30\) and \(k = 0.44\). Assuming \(n = 2\), the 10 possible samples were listed and the biases and mse's of the estimators were computed from first principles, to avoid approximations. The results are in Table 2.1. A poor performance by \(\hat{Y}_r\) may be anticipated here, since the points \((x_i,y_i)\) do not lie near a line through the origin.

Table 2.1: Performance of estimators for population I

<table>
<thead>
<tr>
<th>Estimator</th>
<th>MSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{Y})</td>
<td>72.75</td>
<td>0</td>
</tr>
<tr>
<td>(\hat{Y}_r)</td>
<td>118.77</td>
<td>1.82</td>
</tr>
<tr>
<td>(\hat{Y}_a)</td>
<td>60.08</td>
<td>-0.75</td>
</tr>
</tbody>
</table>

Next, the estimators are applied to eight populations, seven of which have been published and used for comparative purposes in the literature. The other one is the population of livestock and cows yielding milk as noted in 50 of the villages in the survey discussed in Part Two of this thesis. Table 2.2 gives the source, nature of \(y\) and \(x\) and the value of \(k\) for these populations. Let \(\phi\) denote the ratio of \(V(\hat{Y})\) to the mse of an alternative estimator, expressed as a percentage. This can be taken as an indicator of the performance of the alternative estimator. The values of \(\phi\) and bias of the estimators are summarized in Table 2.3. Computations were done only up to second order moments as is the usual practice. The sampling fractions of \(f = 0.02, 0.05, 0.10\) and 0.15 are used to demonstrate
the increase, with \( f \), in the efficiency of \( \hat{Y}_a \) relative to that of \( \hat{Y} \).

Table - 2.2: Description of the populations II to IX

<table>
<thead>
<tr>
<th>Population</th>
<th>Source</th>
<th>( y )</th>
<th>( x )</th>
<th>( N )</th>
<th>( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>Sampford (1967)</td>
<td>Hypothetical</td>
<td>Hypothetical</td>
<td>10</td>
<td>0.23</td>
</tr>
<tr>
<td>III</td>
<td>Murthy (1967, p.228)</td>
<td>Output in a factory</td>
<td>Number of workers</td>
<td>80</td>
<td>0.35</td>
</tr>
<tr>
<td>IV</td>
<td>Murthy (1967, p.228)</td>
<td>Output in a factory</td>
<td>Fixed capital</td>
<td>80</td>
<td>0.44</td>
</tr>
<tr>
<td>V</td>
<td>Yates (1960, p.163)</td>
<td>Measured vol. of timber</td>
<td>Eye-estimated vol. of timber</td>
<td>25</td>
<td>0.57</td>
</tr>
<tr>
<td>VI</td>
<td>Cochran (1977, p.325)</td>
<td>Number of persons in a block</td>
<td>Number of rooms in a block</td>
<td>10</td>
<td>0.74</td>
</tr>
<tr>
<td>VII</td>
<td>Table 3.4</td>
<td>Number of livestock in the village</td>
<td>Number of cows yielding milk in the village</td>
<td>50</td>
<td>0.82</td>
</tr>
<tr>
<td>VIII</td>
<td>Yates (1960, p.159)</td>
<td>Total number of persons in a kraal</td>
<td>Number of persons absent from a kraal</td>
<td>43</td>
<td>0.89</td>
</tr>
<tr>
<td>IX</td>
<td>Kish (1965, p.625)</td>
<td>Number of dwellings occupied by renters</td>
<td>Number of dwellings</td>
<td>270</td>
<td>1.20</td>
</tr>
</tbody>
</table>
Table 2.3: The values of $\phi$ for $\hat{Y}$, $\hat{Y}_r$ and $\hat{Y}_a$ and values of $100|\text{Bias}|/Y$ for $\hat{Y}_r$ and $\hat{Y}_a$

<table>
<thead>
<tr>
<th>Population</th>
<th>k</th>
<th>$\hat{Y}$</th>
<th>$\hat{Y}_r$</th>
<th>$\hat{Y}_a$</th>
<th>$Y_r$</th>
<th>$Y_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>f=0.02</td>
<td>f=0.05</td>
<td>f=0.10</td>
<td>f=0.15</td>
</tr>
<tr>
<td>II</td>
<td>0.23</td>
<td>100</td>
<td>34.54</td>
<td>103.20</td>
<td>107.99</td>
<td>115.43</td>
</tr>
<tr>
<td>III</td>
<td>0.35</td>
<td>100</td>
<td>31.78</td>
<td>111.06</td>
<td>132.37</td>
<td>188.64</td>
</tr>
<tr>
<td>IV</td>
<td>0.44</td>
<td>100</td>
<td>64.91</td>
<td>109.00</td>
<td>125.79</td>
<td>167.39</td>
</tr>
<tr>
<td>V</td>
<td>0.57</td>
<td>100</td>
<td>113.54</td>
<td>102.06</td>
<td>105.33</td>
<td>111.24</td>
</tr>
<tr>
<td>VI</td>
<td>0.74</td>
<td>100</td>
<td>158.83</td>
<td>102.37</td>
<td>106.21</td>
<td>113.42</td>
</tr>
<tr>
<td>VII</td>
<td>0.82</td>
<td>100</td>
<td>337.50</td>
<td>101.10</td>
<td>102.90</td>
<td>106.10</td>
</tr>
<tr>
<td>VIII</td>
<td>0.89</td>
<td>100</td>
<td>147.68</td>
<td>101.51</td>
<td>103.92</td>
<td>108.33</td>
</tr>
<tr>
<td>IX</td>
<td>1.20</td>
<td>100</td>
<td>1178.57</td>
<td>103.28</td>
<td>108.80</td>
<td>119.98</td>
</tr>
</tbody>
</table>
For populations I - IV, cases of moderate \( k \), the estimator \( \hat{Y}_a \) is seen to be appropriate, for population V, \( k = 0.57 \), \( \hat{Y}_r \) and \( \hat{Y}_a \) are almost equally good, while for others, large \( k \), \( \hat{Y}_r \) is better. From Table 2.3, it may be noted that the relative bias of \( \hat{Y}_a \) is quite small.

Thus the empirical study indicates (i) that \( \hat{Y}_a \) is dual to \( \hat{Y}_r \), in that it performs well just when \( \hat{Y}_r \) does not, and (ii) that the general conclusions of Sections 2.2 and 2.3, based on approximations apt for large \( n \), do apply even on a scale as small as that of populations I, II or VI.

2.6 Optimality of the Estimators

Suppose that the population is divided into \( r \) classes and that in the \( i \)th class \( x = x_i \) for all the units. Let \( n \) be allocated to the different classes proportional to their sizes and sampling be simple random in each class. Then it is easy to see that the sample mean \( \bar{x} \) will always equal the population mean \( \bar{X} \). Now if \( \hat{X} = \hat{N} \bar{X} \), then the estimators \( \hat{Y}, \hat{Y}_r \) and \( \hat{Y}_a \) coincide. Therefore in a situation approximating to the above the three estimators are expected to perform equally well.

In general, for \( i = 1, \ldots, N \), define \( u_i = a - bx_i \), where \( a, b \) are positive constants and \( u_i \geq 0 \). With srswor and \( \hat{X} = \hat{N} \bar{X} \), the estimator \( \hat{Y}_a \) proposed in this chapter is the same as the product estimator \( \frac{\hat{Y}u}{\bar{U}} \) for the choice \( a = N\bar{X}/(N-n) \), \( b = n/(N-n) \). Here \( \bar{U}, \bar{u} \) are population and sample means for \( u \). The conditions under which a ratio estimator is optimal are known (Brewer, 1963; Royall, 1970; Royall and Herson, 1973).
An account of these is given by Cochran (1977, pp.158-160). The corresponding conditions for optimality of a product estimator based on arithmetic means are difficult to obtain. However, the following may be mentioned.

Suppose that the \( N \) population values \((y_i, u_i)\) are a random sample from a superpopulation in which

\[
y_i = \frac{B}{u_i} + e_i,
\]

where the \( e_i \) are independent of the \( u_i \). In arrays in which \( u_i \) is fixed, \( e_i \) has mean zero and variance proportional to \( 1/u_i \). Then the estimator \( \hat{Y}_h = \hat{Y} \bar{u}/\bar{U} \) where \( \bar{u} \) and \( \bar{U} \) are respectively sample and population harmonic means for \( u \), is optimal in the same sense as in Brewer-Royall results. But the use of harmonic means is not favoured by survey practitioners. However, in the cases where it is felt that \( \bar{u}/\bar{U} \approx \hat{u}/\hat{U} \) at least approximately, the estimator \( \hat{Y}_a \) will be nearly optimal under model (2.18). The practical relevance of this result is that it suggests conditions under which \( \hat{Y}_a \) may be the best in an entire class of estimators.

2.7 Remarks

For srswor and with \( \hat{X} = N \hat{x} \), \( \hat{X}^* \) as defined in (2.6) reduces to the simple expansion estimator of \( X \) based on the \( (N-n) \) population units not included in the sample.

The expressions (2.8) and (2.9) for bias and mse of \( \hat{Y}_a \) are exact.
These can be estimated by replacing the concerned parameters by the corresponding sample statistics.

If the auxiliary variate assumes both negative and positive values, the use of the ratio or the product method of estimation based directly on $x$ is better avoided since $\hat{X}$ or $X$ may happen to be close to zero. Chapter 6 considers this problem. In general, the results of the present chapter hold whenever $\nu_{11} > 0$, that is $\rho_{XY} > 0$. If $\rho_{XY} < 0$, a ratio estimator can be formed with $\hat{X}^*$ in the denominator. Conditions for its optimality follow easily from Brewer-Royall results.
CHAPTER 3

EXTENDING PRODUCT METHOD OF ESTIMATION

TO THE POSITIVE CORRELATION CASE

[Continuing with the idea of the previous chapter a general transformation is suggested below to permit a product method of estimation rather than a ratio method in the common situation of positive correlation between \( \hat{Y} \) and \( \hat{X} \). This leads to the advantage that the bias and mse have exact expressions. An extension to use multiauxiliary information is outlined.]

3.1 Introduction

The previous chapter proposed a product-type estimator as a dual to the ratio estimator. This was achieved through a simple transformation of \( \hat{X} \) defined in (2.6). The present chapter seeks a general transformation which allows the use of the product method even in cases appropriate for ratio estimator, since the former admits closed form expressions for the bias and mse while the latter does not. In fact the closeness of the expressions (2.2), (2.3) respectively to the actual bias and mse of the ratio estimator \( \hat{Y}_t \) depends much on the composition of the population, the sampling design and the sample size. Hence these expressions must be taken with reservation (Murthy, 1967, p.365). Motivated by this, consider simple transformations

22
that render the situation suitably for a product method instead of the
ratio method. First assume the variates to be nonnegative and \( \rho \) to be
positive. Let \( \hat{Z} = L - \hat{X} \), where \( L \) is a scalar to be chosen. Clearly
\( \hat{Z} \) is unbiased for \( Z = L - X \) and \( \text{cor}(\hat{Y}, \hat{Z}) = -\rho \). Now consider the
following estimator of \( Y \):

\[
\hat{Y}_p = \hat{Y}(\hat{Z}/\hat{Z}).
\]  

(3.1)

The bias and mse of \( \hat{Y}_p \) are, with \( \Theta = X/(L-X) \),

\[
\text{B}(\hat{Y}_p) = -\Theta V_{11} \]

(3.2)

\[
\text{M}(\hat{Y}_p) = Y^2 \left[ V_{20} + \Theta(\Theta-2k)V_{02} + 2\Theta(V_{12}^2V_{21}) + \Theta^2 V_{22} \right].
\]  

(3.3)

These are respectively the same as the expressions for bias and mse of
the estimator \( \hat{Y}_a \) of the previous chapter as given in (2.2), (2.3)
except that \( g \) is now replaced by \( \Theta \). The variance estimators for
products of estimators have been considered by Goodman (1960).

### 3.2 The Choice of \( L \)

\( \text{M}(\hat{Y}_p) \) is minimized when \( \Theta_{\text{opt}} = (kV_{02} + V_{21})/(V_{02} + 2V_{12} + V_{22}) \) and

the corresponding \( L \) is

\[
L_{\text{opt}} = X(1 + \Theta_{\text{opt}})/\Theta_{\text{opt}}
\]

\[
= X(1 + 1/k) + X(2V_{12} + V_{22} - V_{21}/k)/(kV_{02} + V_{21}).
\]  

(3.4)
Let \( \hat{Y}_p^* \) denote \( \hat{Y}_p \) for optimum \( L \). The bias and mse of \( \hat{Y}_p^* \) are obtained by replacing \( \Theta \) in (3.2) with \( \Theta_{opt} \). Ideally, one should know \( L_{opt} \), but in practice this is seldom possible. At the best an approximation to \( L_{opt} \) may be obtained. For example as \( X(1 + 1/k) \); the second term on the right side of (3.4) is likely to be unimportant since \( V_{ij} \) with \( i + j > 2 \) are generally small (Murthy, 1967, pp.380-81). The quantity \( X \) is known and it is often possible to assess the value of \( k \). This problem of assessing certain parameters has been studied among others by Murthy (1967, pp.96-99) and Reddy (1978). Let \( L_o \) denote the corresponding approximation to \( L_{opt} \). We examine below to what extent \( L_o \) may deviate from \( L_{opt} \) and yet give an estimator better than \( \hat{Y}_r \) and \( \hat{Y} \).

To get specific ideas, the \( V_{ij} \) with \( i + j > 2 \) are ignored in the expressions. Thus \( \Theta_{opt} = k \) and

\[
M(\hat{Y}_p^*) = (1-p^2) Y^2 V_{20} = (1-p^2) V(\hat{Y}),
\]

(3.5)

\[
M(\hat{Y}_p) = Y^2 \left[ V_{20} + \Theta_o (\Theta_o - 2k) V_{02} \right] \\
= M(\hat{Y}_p^*) \left[ 1 + e^2 p^2/(1-p^2) \right],
\]

(3.6)

if \( \Theta = \Theta_o = k(1+e) \) when \( L = L_o \). It is seen that \( M(\hat{Y}_p^*) \) is the same as the variance of the difference estimator \( \hat{Y} - B(\hat{X} - \hat{X}) \) in the ideal case, namely when \( B \) is the coefficient of regression of \( \hat{Y} \) on \( \hat{X} \). \( M(\hat{Y}_p^*) \) is also the same as the large sample approximation to the mse of the regression estimator. From (3.6) it follows that the proportional increase in the mse
of $Y_p$ over that of $Y^*$ is less than $d$ if

$$|e| < \sqrt{\frac{d(1-\rho^2)}{\rho^2}}.$$  \hspace{1cm} (3.7)

Thus to ensure only a small relative increase in $\text{mse}$, $|e|$ must be close to 0 if $\rho$ is high, but can depart considerably from 0 if $\rho$ is just moderate. Also from (2.3) and (3.6) we get

$$M(Y_{\hat{r}}) - M(Y_p) = \gamma^2 ((k-1)^2 - (\theta_o - k)^2) \nu \text{O2} > 0$$

when

$$\theta \text{ lies between } (2k-1) \text{ and } 1.$$ \hspace{1cm} (3.8)

Similarly a necessary and sufficient condition for $M(Y_p) < V(Y)$ is

$$0 < \theta_o < 2k.$$ \hspace{1cm} (3.9)

To investigate where (3.8), (3.9) are satisfied simultaneously we distinguish between the cases $0 \leq k \leq 1$ and $k > 1$.

**Case I.** $0 \leq k \leq 1$.

Here choose $L_o > 2X$ so that $\theta_o$ is in $(0,1)$. If $k$ happens to be in $(0,0.5)$, the condition (3.8) is automatically met since $(2k-1) < 0$, but (3.9) needs

$$L_o > (1 + 1/2k)X.$$ \hspace{1cm} (3.10)

On the other hand if $k$ is in $(0.5,1)$, then (3.9) is always met since $2k > 1$, but (3.8) requires

$$L_o < [1 + 1/(2k-1)]X.$$ \hspace{1cm} (3.11)
Thus any $L_0 > 2X$ satisfying (3.10) or (3.11) as the case may be will make $\hat{Y}_p$ an improved estimator.

Case II. $k > 1$.

Here choose $L_0 < 2X$. In addition we need only that $L_0 > [1 + 1/(2k-1)]X$ for $\hat{Y}_p$ to be more precise than $\hat{Y}_r$ or $\hat{Y}$. To give a clearer idea some typical situations are presented in Table 3.1.

<table>
<thead>
<tr>
<th>$k$</th>
<th>Lower bound on $L$</th>
<th>Optimum $L$</th>
<th>Upper bound on $L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>6.00X</td>
<td>11.00X</td>
<td>$\infty$</td>
</tr>
<tr>
<td>0.2</td>
<td>3.50X</td>
<td>6.00X</td>
<td>$\infty$</td>
</tr>
<tr>
<td>0.3</td>
<td>2.67X</td>
<td>4.33X</td>
<td>$\infty$</td>
</tr>
<tr>
<td>0.4</td>
<td>2.25X</td>
<td>3.50X</td>
<td>$\infty$</td>
</tr>
<tr>
<td>0.5</td>
<td>2.00X</td>
<td>3.00X</td>
<td>$\infty$</td>
</tr>
<tr>
<td>0.6</td>
<td>2.00X</td>
<td>2.67X</td>
<td>6.00X</td>
</tr>
<tr>
<td>0.7</td>
<td>2.00X</td>
<td>2.43X</td>
<td>3.50X</td>
</tr>
<tr>
<td>0.8</td>
<td>2.00X</td>
<td>2.25X</td>
<td>2.66X</td>
</tr>
<tr>
<td>0.9</td>
<td>2.00X</td>
<td>2.11X</td>
<td>2.25X</td>
</tr>
<tr>
<td>1.0</td>
<td>2.00X</td>
<td>2.00X</td>
<td>2.00X</td>
</tr>
<tr>
<td>1.1</td>
<td>1.83X</td>
<td>1.91X</td>
<td>2.00X</td>
</tr>
<tr>
<td>1.3</td>
<td>1.63X</td>
<td>1.77X</td>
<td>2.00X</td>
</tr>
<tr>
<td>1.5</td>
<td>1.50X</td>
<td>1.67X</td>
<td>2.00X</td>
</tr>
<tr>
<td>2.0</td>
<td>1.33X</td>
<td>1.50X</td>
<td>2.00X</td>
</tr>
<tr>
<td>2.5</td>
<td>1.25X</td>
<td>1.40X</td>
<td>2.00X</td>
</tr>
<tr>
<td>3.0</td>
<td>1.20X</td>
<td>1.33X</td>
<td>2.00X</td>
</tr>
</tbody>
</table>
Interestingly the choice \( L_o = 2.25X \) covers a fairly wide range for \( k \) from 0.4 to 0.9, being actually optimum for \( k = 0.8 \).

Similarly \( L_o = 3.5X \) suits the range from 0.2 to 0.7 for \( k \) with optimum at \( k = 0.4 \). In fact the choice of \( L_o \) is very flexible when \( k \) is moderate, say in (0,0.7). This flexibility disappears in the neighbourhood of \( k = 1 \). However a value like \( L_o = 1.9X \) is safe virtually for all \( k > 1 \). Better selections can be made when \( k \) is known more precisely.

### 3.3 Case of Negative Correlation

When \( p < 0 \), take \( \hat{Z} = L + \hat{X} \) so that \( \text{cor}(\hat{Y}, \hat{Z}) = \text{cor}(\hat{Y}, \hat{X}) \).

Here it is appropriate to compare \( \hat{Y} \) with the traditional product estimator \( \hat{Y}(\hat{X}/X) \). An approximation to \( L_{opt} \) is \(-(1 + 1/k)X\). The restrictions on the choice \( L_o \) for \( L_{opt} \) can be investigated. It turns out that \( L_o = 0.25X \) covers the range \(-0.9\) to \(-0.4\) for \( k \), being the best \( k = -0.8 \), while \( L_o = 1.5X \) is suitable for \( k \) in \((-0.7, -0.2)\) being actually optimum at \( k = -0.4 \). And a choice like \( L_o = -0.10X \) is practically safe for all \( k < -1 \). A better selection can be made if \( k \) is more precisely known. Figure 3.1 presents the results in a graphical form.

The rules of thumb for choosing \( L \) when a firm guess of the value of \( k \) cannot be made, but only an interval containing \( k \) can be specified, are given in Table 3.2.
Fig. 3.1 Region where $\hat{F}_p$ has smaller variance
Table 3.2: Rules of thumb for choosing $L$

<table>
<thead>
<tr>
<th>$k &gt; 0$</th>
<th>$k &lt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case</strong></td>
<td><strong>$L$</strong></td>
</tr>
<tr>
<td>$0 &lt; k \leq \frac{1}{2}$</td>
<td>$3.50X$</td>
</tr>
<tr>
<td>$\frac{1}{2} &lt; k &lt; 1$</td>
<td>$2.25X$</td>
</tr>
<tr>
<td>$k &gt; 1$</td>
<td>$1.90X$</td>
</tr>
</tbody>
</table>

These rules are not applicable when $k$ is in the neighbourhood of either 0 or $\pm 1$. In fact if $k$ is (i) close to 0 the simple estimator $\hat{Y}$ is to be used, (ii) close to 1 either $\hat{Y}_r$ or $\hat{Y}_p$ with $L = 2X$ may be used, and (iii) close to $-1$, $\hat{Y}_p$ with $L = 0$ (which is the same as the usual product estimator) may be used.

If it is desirable to use an exactly unbiased estimator we may consider the alternatives suggested in Sec. 2.4, with the change that in the place of $\hat{X}^*$ and $X$ we now use $\hat{X}$ and $X$.

3.4 Use of Multiauxiliary Information

Frequently information on several $x$-variates may be used: for instance utilizing census data to adjust current estimates. Let $x_t$, $t = 1, \ldots, q + s$ be the auxiliary variates and $\hat{Y}, \hat{x}_t$ be unbiased estimators of the totals $Y, x_t$, based on any probability sampling design.

All values are real, nonnegative and $x_t$ are known. Also let $\text{cor}(\hat{Y}, \hat{x}_t) > 0$.
for \( t = 1, \ldots, q \) and \( \text{cor}(\hat{Y}, \hat{X}_t) < 0 \) for \( t = q+1, \ldots, q+s \). Then Rao and Mudholkar (1967) have suggested the following linear combination of estimators of ratio and product-types as an estimator of \( Y \):

\[
\hat{Y}_{RM} = \sum_{t=1}^{q} a_t (X_t \hat{X}_t) + \sum_{q+1}^{q+s} a_t (X_t \hat{X}_t)
\]

where the \( a_t \) are the weights; \( \sum a_t = 1 \). As an alternative we consider the following extension of the methods of Sections 3.1 and 3.3.

The estimator \( \hat{Z}_t = L_t + h_t \hat{X}_t \) is unbiased for \( Z_t = L_t + h_t X_t \) for each \( t \). Take \( h_t = 1 \) for \( t = 1, \ldots, q \) and \( h_t = 1 \) for \( t = q+1, \ldots, q+s \). Then an estimator of \( Y \) is

\[
\hat{Y}_p = \sum_{t=1}^{q+s} \left( \hat{Z}_t / Z_t \right) W_t
\]

where \( W = (W_1, \ldots, W_{q+s}) \), \( \sum W_t = 1 \), is a vector of weights. And

\[
\hat{B}(Y_p) = \sum_{t=1}^{q+s} \left( \hat{Y}_p \Theta \hat{Y}(t) - \sum_{q+1}^{q+s} \hat{Y}_p \Theta \hat{Y}(t) \right) \frac{1}{t \cdot l \cdot l \cdot l}
\]

and the mse, up to second order moments, is

\[
\hat{M}(Y_p) = \sum_{t=1}^{q+s} \frac{1}{t \cdot u \cdot u \cdot u} W \Theta D W
\]

where the elements of the matrix \( D \) are given by

\[
d_{tu} = V_{20} + h_t \Theta V_{10} + h_t \Theta V_{01} + h_t h_u \Theta V_{01}
\]

with \( \Theta_t = X_t / (L_t + h_t X_t) ; t, u = 1, \ldots, q+s \). Here \( V_{ij} \) stand for the
relative central moments \( V_{ij} \) defined in (2.4) with \((X-X)\) replaced by \( \hat{X}_t - X \) by \( X_t \) and

\[
V_{11}^{(tu)} = E(\hat{X}_t - X_t)(X - X)/X_t X_t
\]

As shown in Rao and Mudholkar (1967) the matrix \( D \) is positive definite if the \((q+s+1) \times (q+s+1)\) matrix of the coefficients of variation of \( \hat{Y} \) and \( \hat{Z}_t \) is positive definite.

Theoretically the \( L_t \) can be determined to minimize \( M(\hat{Y}_t) \). But a practicable alternative is to choose \( L_t \) such that \( d_{tt} \) is controlled. Thus as a rule of thumb, \( L_t \) may be 3.5\( X_t \) if \( k_t = V_{11}^{(t)}/V_{00}^{(t)} \) is positive but moderate, while \( L_t \) may be 1.5\( X_t \) when \( k_t \) is negative but moderate.

Other choices may be made as discussed in Sections 3.2 and 3.3.

Next, applying the generalized Cauchy inequality (see Olkin, 1958) the \( W_t \) optimum in the sense of minimizing \( M(\hat{Y}_t) \) for given \( D \) are provided by

\[
W_{\text{opt}} = eD^{-1}/eD^{-1}e'
\]

where \( e = (1, \ldots, 1) \). Substituting \( W_{\text{opt}} \) in (3.3),

\[
M_{\text{min}}(\hat{Y}_t) = \gamma^2/eD^{-1}e'.
\]

However, in surveys \( W_{\text{opt}} \) can rarely be computed and used since the matrix \( D \) is usually unknown. Theoretically \( W_t \) will all be equal (= 1/(q+s)) if and only if the column sums of \( D \) are equal. A hypothetical example of this occurs when the population coefficients of variation of the \( \hat{Z}_t \) are
all equal, \( Y \) is equally correlated with all \( Z_t \) and all pairs of two
different estimators \( Z_t \) have the same correlation. Usually the \( W_t \)
are selected from experience and theoretical considerations. In small
scale surveys of specialized scope it may be feasible to estimate \( W_{opt} \)
from the sample itself. The sampling error of these estimates must be
examined.

3.5 Empirical Performance of \( Y_p \)

For purposes of illustration, strswor is assumed in this section.
The following five populations are considered.

Population 1. Hypothetical population of 7 pairs of \((y,x)\) values:
\((1,6), (2,5), (4,7), (5,2), (6,4), (8,10)\) and \((9,8)\).

Population 2. In the survey of villages reported in Part Two of the
thesis, the Heads of 20 of the households who had bank accounts were
asked to state the number of withdrawals \((y)\) during July-December 1980,
number of deposits during July-December 1980 \((x_1)\) and during January-June
1980 \((x_2)\). The respondents were asked to refer their bank passbooks to
be able to get these numbers as accurately as possible. The responses are
recorded in Table 3.3 as population 2.

Table - 3.3: Number of bank deposits \((y)\) during a 6-month period and
withdrawals \((x_1, x_2)\) during two 6-month periods by 20
households in the survey of Karnataka.

\[
\begin{align*}
y & : 12 22 38 15 18 31 15 20 10 25 11 17 12 22 14 26 08 16 13 19 \\
x_1 & : 14 25 37 18 20 30 15 21 12 28 14 19 12 23 16 28 09 15 15 20 \\
x_2 & : 30 25 09 30 28 12 30 24 36 28 30 30 31 25 31 25 35 25 30 28
\end{align*}
\]
Population 3. Data on the number of workers ($x_1$), fixed capital $x_2$ and output ($y$) for 80 factories in a certain region (Murthy, 1967, p. 228).

Population 4. Data on cultivated area ($y$) and area under wheat ($x_1, x_2$) during two different years for 34 villages in a certain region (Murthy, 1967, p. 399, Table 10.6).

Population 5. Data on the number of livestock ($y$), number of cows yielding milk ($x_1$) and the number of farms ($x_2$) in 50 of the villages in a study of the Karnataka State in India, reported in Part Two of the thesis. Details are in Table 3.4.

Table 3.4. Number of livestock ($y$), milk yielding cows ($x_1$) and farms ($x_2$) in 50 villages of Karnataka.

<table>
<thead>
<tr>
<th>y</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>y</th>
<th>$x_1$</th>
<th>$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>70</td>
<td>55</td>
<td>715</td>
<td>190</td>
<td>170</td>
</tr>
<tr>
<td>630</td>
<td>160</td>
<td>155</td>
<td>845</td>
<td>360</td>
<td>335</td>
</tr>
<tr>
<td>1200</td>
<td>320</td>
<td>285</td>
<td>1016</td>
<td>235</td>
<td>219</td>
</tr>
<tr>
<td>1170</td>
<td>445</td>
<td>381</td>
<td>184</td>
<td>74</td>
<td>63</td>
</tr>
<tr>
<td>1060</td>
<td>250</td>
<td>278</td>
<td>282</td>
<td>62</td>
<td>79</td>
</tr>
<tr>
<td>823</td>
<td>120</td>
<td>112</td>
<td>195</td>
<td>71</td>
<td>60</td>
</tr>
<tr>
<td>1737</td>
<td>560</td>
<td>632</td>
<td>439</td>
<td>137</td>
<td>100</td>
</tr>
<tr>
<td>1061</td>
<td>254</td>
<td>278</td>
<td>854</td>
<td>195</td>
<td>142</td>
</tr>
<tr>
<td>360</td>
<td>102</td>
<td>112</td>
<td>821</td>
<td>260</td>
<td>265</td>
</tr>
<tr>
<td>945</td>
<td>359</td>
<td>345</td>
<td>740</td>
<td>142</td>
<td>85</td>
</tr>
<tr>
<td>470</td>
<td>110</td>
<td>100</td>
<td>750</td>
<td>142</td>
<td>190</td>
</tr>
<tr>
<td>1625</td>
<td>481</td>
<td>489</td>
<td>625</td>
<td>130</td>
<td>140</td>
</tr>
<tr>
<td>827</td>
<td>125</td>
<td>113</td>
<td>600</td>
<td>125</td>
<td>138</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>8</td>
<td>75</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>1300</td>
<td>428</td>
<td>340</td>
<td>152</td>
<td>60</td>
<td>38</td>
</tr>
<tr>
<td>337</td>
<td>78</td>
<td>82</td>
<td>65</td>
<td>40</td>
<td>35</td>
</tr>
<tr>
<td>260</td>
<td>75</td>
<td>105</td>
<td>235</td>
<td>142</td>
<td>128</td>
</tr>
<tr>
<td>186</td>
<td>45</td>
<td>28</td>
<td>280</td>
<td>162</td>
<td>190</td>
</tr>
<tr>
<td>1760</td>
<td>564</td>
<td>515</td>
<td>335</td>
<td>180</td>
<td>200</td>
</tr>
<tr>
<td>605</td>
<td>238</td>
<td>245</td>
<td>432</td>
<td>200</td>
<td>235</td>
</tr>
</tbody>
</table>

| 700 | 92    | 85    | 645  | 148   | 162   |
| 524 | 247   | 220   | 800  | 300   | 80    |
| 571 | 134   | 133   | 926  | 400   | 230   |
| 962 | 131   | 145   | 421  | 220   | 150   |
| 407 | 129   | 102   | 850  | 321   | 140   |
For populations 2, 3, 4 and 5, three subcases are studied: using $x_1$ alone, $x_2$ alone, and $x_1x_2$ as auxiliary variates. These are respectively denoted by 2a, 2b, 2c etc. In the case of population 1, all possible samples of $n = 3$ units were listed and the biases and mses were computed from first principles to avoid approximation, while in 2a the exact expressions for bias and mse were used. In the remaining cases computations were made only up to second order moments. The rules of thumb in Table 3.2 were applied for choosing $L$ for $\hat{Y}_p$, while $L$ was taken to be $(1 + 1/k)X$ or $-(1 + 1/k)X$ as the case may be for $\hat{Y}_p$. When information on $x_1$ and $x_2$ was utilized, $\hat{Y}_p$ was compared with the generalized multivariate estimator $\hat{Y}_{RM}$ discussed in Rao and Mudholkar (1967), with weights $w_1 = w_2 = 1/2$. The results are summarized in Table 3.5. For srswor the relative bias of the suggested estimator $\hat{Y}_p$ reduces to

$$B(\hat{Y}_p)/Y = \frac{n(n-1)}{n^2}S_{11}/n\bar{Y}\bar{X},$$

where $S_{11}$ is the population covariance between $Y$ and $X$. Hence

$$\min\{n/(n-n)\} \cdot B(\hat{Y}_p)/Y = S_{11}/\bar{Y}\bar{X},$$

where the right side expression is seen to be independent of the sample size $n$. To have this convenience the values of $100[n/(n-n)] \cdot \text{Bias}/Y$ are reported in Table 3.5 within parentheses.
Table 3.5: The values of $\phi$ and $100\{\text{Na/(N-n)}\} |\text{Bias}|/Y$
for different estimators.

<table>
<thead>
<tr>
<th>Population</th>
<th>$k$</th>
<th>100</th>
<th>Estimate (product) estimator</th>
<th>$\hat{Y}_p$</th>
<th>$\hat{Y}^*_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.57</td>
<td>100</td>
<td>94 (10)</td>
<td>115 (9)</td>
<td>121 (6)</td>
</tr>
<tr>
<td>2a</td>
<td>1.11</td>
<td>100</td>
<td>2433 (1)</td>
<td>4253 (16)</td>
<td>4253 (16)</td>
</tr>
<tr>
<td>2b</td>
<td>-0.50</td>
<td>100</td>
<td>100 (3)</td>
<td>109 (1)</td>
<td>109 (1)</td>
</tr>
<tr>
<td>2c</td>
<td>1.11, -0.50</td>
<td>100</td>
<td>218 (2)</td>
<td>315 (1)</td>
<td>798 (1)</td>
</tr>
<tr>
<td>3a</td>
<td>0.35</td>
<td>100</td>
<td>32 (58)</td>
<td>729 (13)</td>
<td>837 (11)</td>
</tr>
<tr>
<td>3b</td>
<td>0.44</td>
<td>100</td>
<td>65 (33)</td>
<td>1088 (10)</td>
<td>1197 (11)</td>
</tr>
<tr>
<td>3c</td>
<td>0.35, 0.44</td>
<td>100</td>
<td>45 (46)</td>
<td>246 (11)</td>
<td>1052 (11)</td>
</tr>
<tr>
<td>4a</td>
<td>0.75</td>
<td>100</td>
<td>376 (13)</td>
<td>565 (31)</td>
<td>577 (29)</td>
</tr>
<tr>
<td>4b</td>
<td>0.71</td>
<td>100</td>
<td>318 (16)</td>
<td>514 (32)</td>
<td>549 (29)</td>
</tr>
<tr>
<td>4c</td>
<td>0.75, 0.71</td>
<td>100</td>
<td>365 (15)</td>
<td>570 (32)</td>
<td>590 (29)</td>
</tr>
<tr>
<td>5a</td>
<td>0.82</td>
<td>100</td>
<td>338 (38)</td>
<td>381 (31)</td>
<td>382 (28)</td>
</tr>
<tr>
<td>5b</td>
<td>0.20</td>
<td>100</td>
<td>56 (86)</td>
<td>101 (18)</td>
<td>105 (17)</td>
</tr>
<tr>
<td>5c</td>
<td>0.82, 0.20</td>
<td>100</td>
<td>342 (42)</td>
<td>365 (28)</td>
<td>396 (30)</td>
</tr>
</tbody>
</table>

$\phi = 100V(\hat{Y})/(\text{mse of the estimator})$.

The value of $k$ ranges from -0.50 to 1.11 in Table 3.5. Here substantial gain in efficiency is seen when $\hat{Y}_p$ is used instead of the traditional estimators, in most of the cases. Also $\hat{Y}^*_p$ compares quite well with the ideal case of $\hat{Y}_p^*$. Thus the illustrations indicate that (i) the use of $\hat{Y}_p$ is desirable in practice, and (ii) the rules of thumb for choosing $L$ work well. If an unbiased estimator is preferred then the techniques outlined in Sec. 2.4 may be employed.
CHAPTER 4.

VARIATE TRANSFORMATIONS AFTER SAMPLING

[A change of origin for $y$ is illustrated for improving the precision of estimators under varying probability sampling. A change of origin under ratio method of estimation and a change of scale under the difference method are also considered.]

4.1 Introduction

It is known that ppswr sampling is very effective when the value of the study variate $y$ is proportional to the value of the measure of size $x$. In ppsw or sampling, proportionality between value of $y$ and the probability of including the unit in the sample is desirable. Sections 4.2 to 4.8 discuss a change of origin for $y$ for achieving such proportionality. Sections 4.9 and 4.10 examine changes of origin and scale for $\hat{y}, \hat{x}$ in the context of ratio and difference methods of estimation. Assume that $x$ is a positive valued variate.

4.2 PPSWR Sampling

Consider with replacement pps sampling of $n$ units, using $x$ as the measure of size. The usual unbiased estimator of $Y$ is

$$\hat{Y}_{pps} = \frac{X}{n} \sum_{i=1}^{n} y_i/x_i.$$  \hspace{1cm} (4.1)

Let $z_i = y_i - a$; $i = 1, \ldots, N$, $a$ being a scalar to be chosen. As an
estimator of $Y$ consider

$$\hat{Y}_1 = (X/n) \left[ \Sigma (y_i - a)/x_i \right] + Na . \quad (4.2)$$

It is easy to show that $\hat{Y}_1$ is unbiased for $Y$. On the other hand, if

srs wr is used an unbiased estimator of $Y$ would have been $\hat{Y} = (N/n) \Sigma y_i$.

The variances of the three competing estimators are given by

$$V(\hat{Y}_{pps}) = (x/n) \Sigma x_i^2 - \beta^2/n , \quad (4.3)$$

$$V(\hat{Y}_1) = (x/n) \Sigma (y_i - a)^2/x_i - (Y - Na)^2/n , \quad (4.4)$$

$$V(Y) = (N/n) \Sigma y_i^2 - \beta^2/n . \quad (4.5)$$

4.3 The Choice of $a$

From (4.3) and (4.4) we obtain

$$V(\hat{Y}_{pps}) - V(\hat{Y}_1) = \frac{x}{n} \left[ 2a\Sigma y_i/x_i - a^2 \Sigma (1/x_i) \right]$$

$$+ \left[ \frac{N^2 a^2}{n} - \frac{2N a Y}{n} \right]$$

$$= \frac{NX}{n} \left[ 2a\bar{X} \bar{Y} - a^2 \bar{X} \right] + \frac{N^2 a}{n} \left[ a - 2\bar{Y} \right]$$

$$= \frac{N^2 a}{nX} \left[ 2\bar{X} \bar{R} - a\bar{X} \right] + \frac{N^2 a}{nX} \left[ a\bar{X} - 2\bar{XY} \right]$$

$$= \frac{N^2 a}{nX} \left[ 2\bar{X} \bar{R} - a\bar{X} \right]$$

$$= \frac{N^2 a}{nX} \left[ 2\bar{X} (\bar{R}-R) - a(\bar{X}-\bar{R}) \right] . \quad (4.6)$$

where $X = N/\Sigma (1/x_i)$ is the harmonic mean of $x$, and $\bar{R} = (1/N)\Sigma (y_i/x_i)$.
Therefore \( V(\hat{Y}_{pps}) - V(\hat{Y}_{*}) > 0 \), and hence \( \hat{Y}_{*} \) is more precise than \( \hat{Y}_{pps} \) when both the factors in (4.6) have the same sign. Since \( X_\bar{X} X > 0 \) when at least two \( x_i \) are unequal, this needs that

\[
a > 0 \quad \text{and} \quad a < \frac{2XX(\bar{X}-X)}{(\bar{X}-\bar{X})}
\]

or

\[
a < 0 \quad \text{and} \quad a > \frac{2XX(\bar{X}-X)}{(\bar{X}-\bar{X})}.
\]

That is

\[
a \text{ lies between } 0 \text{ and } 2a^*, \quad (4.7)
\]

where

\[
a^* = \frac{XX(\bar{X}-X)}{(\bar{X}-\bar{X})}. \quad (4.8)
\]

It is interesting to note that the sign of \( a^* \) is the same as that of \( (\bar{X}-\bar{X}) \), since \( (\bar{X}-X) > 0 \). In particular the value of a minimizing \( V(\hat{Y}_{*}) \) is \( a_{opt} = a^* \), which is the midpoint of the interval for \( a \), specified in (4.7). Denote the estimator \( \hat{Y}_{*} \) for the choice \( a_{opt} = a^* \) by \( \hat{Y}_{*} \).

The expression for \( V(\hat{Y}_{*}) \) is the same as (4.4) with \( a \) replaced by \( a^* \).

To get a simpler idea consider the case \( y_i = A + Bx_i \), so that there is perfect correlation between \( y \) and \( x \). Then

\[
\bar{R} - R = \frac{1}{N} \sum_{i=1}^{N} \left( y_i/x_i \right) - \frac{1}{N} \sum_{i=1}^{N} y_i/x_i
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} \left( A + Bx_i \right)/x_i - \frac{1}{N} \sum_{i=1}^{N} \left( A + Bx_i \right)/x_i
\]

\[
= \frac{A(\bar{X} - X)}{\bar{X}X}, \quad (4.9)
\]

and hence \( a^* = A \). Thus in general when the regression of \( y \) on \( x \) is linear, \( a \) can be interpreted as an approximation to the intercept \( A \). Also
when the regression line passes through a point far from the origin (which corresponds to nonproportionality between \( y_i \) and \( x_i \)) the value of \( a^* \) will tend to be large and there will be sufficient flexibility in the choice of \( a \).

Reddy and Rao (1977) have suggested a change of origin for \( x \) in order to improve the proportionality between values of \( y \) and \( x \). The suggested estimator is

\[
\hat{y}^*_{pp} = \left( x'/n \right) \sum_{i=1}^{n} \left( y_i/x_i' \right),
\]

where \( x_i' = x_i + d\bar{x} \) are the size measures, \( x' = \sum_{i=1}^{N} x_i' \) and \( d^* = (1-k)/k \).

If we are interested in a single \( y \), this manipulation of the size measures works well. However, a transformation of \( x \) has the following practical difficulties, unlike that of \( y \):

(i) The transformation is to be made before the sample is drawn.

(ii) The transformed \( x \) has to be positive for every unit.

(iii) When multiple characteristics are being estimated from the same sample, a single transformation of \( x \) may not suit all the cases.

4.4. Empirical Efficiency of \( \hat{y}^*_1 \)

Five populations, which are the same as those used by Reddy and Rao (1977), are employed. This allows comparisons.

Population 1: Consists of data on number of workers (\( x \)) and output (\( y \)) for 80 factories in a certain region (Murthy, 1967, p.228).
Population 2: Consists of data on fixed capital \((x)\) and output \((y)\) for the factories in population 1.

The other three are the hypothetical populations A, B and C considered by Yates and Grundy (1953), with details as in Table 4.1.

The value of \(a\) and \(n\) (estimator variance) for the populations are in Table 4.2.

<table>
<thead>
<tr>
<th>Table 4.1: Populations A, B and C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4.2: Sampling variance for the five populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>
We note that for populations $A, B$ the estimator $\hat{Y}^*_1$ performs the best, it is the same as $\hat{Y}_{pp}$ for $C$, and for populations 1 and 2, it is next only to $\hat{Y}_{pp}^*$. Percentage efficiencies of $\hat{Y}, \hat{Y}_{pp}^*$ and $\hat{Y}_1^*$ relative to that of $\hat{Y}_{pp}$ are in Table 4.3, while values of $nV(\hat{Y}_1)$ for typical choices of $a$ for populations 1 and 2 are in Table 4.4.

Table 4.3: Percentage efficiencies of the different estimators

<table>
<thead>
<tr>
<th>Population</th>
<th>$\hat{Y}_{pp}$</th>
<th>$\hat{Y}$</th>
<th>$\hat{Y}_{pp}^*$</th>
<th>$\hat{Y}_1^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>318</td>
<td>1644</td>
<td>753</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>133</td>
<td>930</td>
<td>493</td>
</tr>
<tr>
<td>A</td>
<td>100</td>
<td>6</td>
<td>505</td>
<td>581</td>
</tr>
<tr>
<td>B</td>
<td>100</td>
<td>30</td>
<td>568</td>
<td>581</td>
</tr>
<tr>
<td>C</td>
<td>100</td>
<td>22</td>
<td>89</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4.4: Values of $V(\hat{Y}_1) \times 10^{-8}$ for typical choices of $a$

<table>
<thead>
<tr>
<th>$V(\hat{Y}_1) \times 10^{-8}$</th>
<th>Pop. 1</th>
<th>Pop. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>676</td>
<td>282</td>
</tr>
<tr>
<td>500</td>
<td>452</td>
<td>189</td>
</tr>
<tr>
<td>1000</td>
<td>333</td>
<td>121</td>
</tr>
<tr>
<td>1500</td>
<td>217</td>
<td>77</td>
</tr>
<tr>
<td>2000</td>
<td>139</td>
<td>58</td>
</tr>
<tr>
<td>2500</td>
<td>97</td>
<td>63</td>
</tr>
<tr>
<td>3000</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>3500</td>
<td>125</td>
<td>147</td>
</tr>
<tr>
<td>4000</td>
<td>195</td>
<td>226</td>
</tr>
</tbody>
</table>
From Table 4.4 we note that there is enough flexibility in the choice of $a$. Finally, Table 4.5 shows the performances of $\hat{Y}_{pps}^*$, $\hat{Y}_1$, Horvitz-Thompson estimator $\hat{Y}_{HT}$, symmetrized Des Raj estimator $\hat{Y}_{SD}$ and Rao-Hartley-Cochran estimator $\hat{Y}_{RHC}$ in relation to that of $\hat{Y}_{pps}$ for the populations A, B, C and $n = 2$.

Table 4.5: Performances of the different estimators

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Population A</th>
<th></th>
<th>Population B</th>
<th></th>
<th>Population C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Variance Efficency</td>
<td>Variance Efficency</td>
<td>Variance Efficency</td>
<td>Variance Efficency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{Y}_{pps}$</td>
<td>0.099 505</td>
<td>0.088 568</td>
<td>0.141 89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{Y}_1$</td>
<td>0.081 617</td>
<td>0.081 617</td>
<td>0.125 100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{Y}_{HT}$</td>
<td>0.823 61</td>
<td>0.057 877</td>
<td>0.059 212</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{Y}_{SD}$</td>
<td>0.333 150</td>
<td>0.333 150</td>
<td>0.083 151</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{Y}_{RHC}$</td>
<td>0.333 150</td>
<td>0.333 150</td>
<td>0.083 151</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{Y}_{pps}$</td>
<td>0.500 100</td>
<td>0.500 100</td>
<td>0.125 100</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is seen that $\hat{Y}_1^*$ is the best (among the estimators considered) for population A and does better than $\hat{Y}_{pps}^*$ in each of the three cases. These empirical studies, though of limited scope, throw some light on the relative performances of the estimators considered.
4.5 PPSWOR Sampling

Consider the Midzuno-Sen scheme of varying probability sampling using $x$ as the measure of size. The first sample unit is selected with probability proportional to the value of $x$, and the subsequent units are selected with equal probability and without replacement. Under this scheme the probability of selecting a specified sample is proportional to the total size of the units included in the sample.

The probability of including $U_i$ in a sample of size $n$ is

$$\pi_i = \frac{(N-n)x_i}{(N-1)x} + \frac{(n-1)}{(N-1)},$$

while the probability of including both $U_i$ and $U_j$ ($i \neq j$) is

$$\pi_{ij} = \frac{(n-1)}{(N-1)(N-2)} \left\{ (N-n)(x_i + x_j)x^{-1} + (n-2) \right\}.$$

Then the Horvitz-Thompson estimator of $Y$ is

$$\hat{Y}_HT = \sum_{i=1}^{n} \left( \frac{y_i}{\pi_i} \right),$$

which is unbiased and

$$V(\hat{Y}_HT) = \sum_{i>j=1}^{N} \left( \pi_{ij} - \pi_{ji} \right) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.$$  \hspace{1cm} (4.13)

An advantage with the Midzuno-Sen scheme is that the Yates-Grundy estimate of $V(\hat{Y}_HT)$ is never negative.

4.6 A Modification to $\hat{Y}_HT$

Expression (4.13) shows that $V(\hat{Y}_HT)$ will be small if $y_i/\pi_i$ is
nearly the same for all the population units. However, even when
\( y_i = Bx_i \) for all \( i \), \( y_i / \pi_i \) will not be a constant because of the
second term on the right-hand side of (4.11). On the other hand,

\[ [y_i + BX(n-1)/(N-n)] \pi_i^{-1} \] will be a constant. Motivated by this, consider
the transformation

\[ z_i = y_i + (n-1)b / (N-n), \quad (i=1, \ldots, N) \] (4.14)

where \( b \) is a scalar to be chosen. Then an unbiased estimator of \( Y \) is

\[ \hat{Y}_2 = \sum_{i=1}^{n} (z_i / \pi_i) - N(n-1)b/(N-n). \] (4.15)

It is seen that

\[
V(\hat{Y}_2) = \sum_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij})(z_i / \pi_i - z_j / \pi_j)^2 \\
= \sum_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij})(y_i / \pi_i - y_j / \pi_j) + \frac{(n-1)b}{N-n} (1/\pi_i - 1/\pi_j)^2 \\
= V(\hat{Y}_{HT}) + b^2 D_1 + 2b D_2, \] (4.16)

where

\[
D_1 = \{(n-1)/(N-n)\}^2 \sum_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij})(1/\pi_i - 1/\pi_j)^2; \\
D_2 = \{(n-1)/(N-n)\} \sum_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij})(y_i / \pi_i - y_j / \pi_j)(1/\pi_i - 1/\pi_j). \\

\]

And \( V(\hat{Y}_2) \) is minimized for the choice \( b_{\text{opt}} = -D_2/D_1 \) with corresponding

\[
V_{\min}(\hat{Y}_2) = V(\hat{Y}_{HT}) - D_2^2/D_1.
\]
Result (4.16) demonstrates that $\hat{Y}_2$ will have a variance smaller than that of $\hat{Y}_{HT}$ when the last two terms on the right-hand side add up to a negative number. That is $b^2D_1 + 2bD_2 < 0$. Since $D_1 > 0$, this will happen if and only if the roots of the equation $b^2D_1 + 2bD_2 = 0$ are real and distinct and $b$ lies between them. Here these roots are real and distinct, and they are $b = 0$ and $b = -2D_2/D_1 = 2b_{opt}$.

Hence for $\hat{Y}_2$ to be more efficient than $\hat{Y}_{HT}$ we obtain the condition that $b$ lies between 0 and $2b_{opt}$.

Thus when $D_2$ is large relative to $D_1$, $b_{opt}$ will be large and hence there will be considerable flexibility in choosing $b$ such that $\hat{Y}_2$ is more efficient than $\hat{Y}_{HT}$. Note that $D_2$ is a quantity that reflects the nonconstancy of $y_i/\pi_i$.

4.7 Choice of $b$

Since in practice $y_i$ is not known for all the units in the population, $b_{opt}$ can not be computed. However, a reasonable choice of $b_{opt}$ can be made. Two methods for doing this are outlined below.

Method I. Consider the case $y_i = Bx_i$ for all $i$. Then

$$D_2 = \frac{1}{N} \sum_{i > j = 1}^N (\pi_i \pi_j - \pi_{ij}) (y_i/\pi_i - y_j/\pi_j)(1/\pi_i - 1/\pi_j)$$

$$= B \frac{1}{N} \sum_{i > j = 1}^N (\pi_i \pi_j - \pi_{ij}) (x_i/\pi_i - x_j/\pi_j)(1/\pi_i - 1/\pi_j)$$
\[- B X [(n-1)/(N-n)]^2 \sum_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij})(1/\pi_i - 1/\pi_j)^2,\]

since, in view of the expression (4.11) for the first order inclusion probabilities, we have

\[ (x_i/\pi_i - x_j/\pi_j) = - (n-1)/(N-n)(1/\pi_i - 1/\pi_j) X. \]

Hence \( b_{opt} \) reduces to \( B X \) in this case and the range for a suitable \( b \) is 0 to 2\( B X \). An estimate of \( B \) may be obtained by plotting \( y_i/\pi_i \) against \( x_i/\pi_i \) for the sample units and gauging the slope of the best fitting line.

**Method II.** Computing \( b \) with the summations in the expressions for \( D_1 \) and \( D_2 \) taken over only the sample units and the terms weighted by \( 1/\pi_{ij} \). That is

\[ b = - \frac{(n-n)}{(n-1)} \frac{\sum(\pi_i \pi_j - \pi_{ij})(y_i/\pi_i - y_j/\pi_j)(1/\pi_i - 1/\pi_j)\pi_{ij}^{-1}}{\sum(\pi_i \pi_j - \pi_{ij})(1/\pi_i - 1/\pi_j)^2 \pi_{ij}^{-1}}, \]

where the summations are over \( i > j = 1, \ldots, n \). This method computes \( b \) from current data. The sampling variability introduced by this factor is not investigated here.

**4.8 Empirical Study**

To compare the efficiency of the estimator \( \hat{Y}_2 \) with that of the usual estimator under the Midzuno-Sen scheme, five populations are considered.
Population I consists of the number of cattle, \( y \), and the number of farms, \( x \), in 13 village blocks as noted in the Dakshina Kannada district of Karnataka during the survey discussed in Part Two of the thesis. Details are in Table 4.6. Population II is the population considered by Sampford (1978, p.37, Table 2), where the concepts of predictive estimation and internal congruency are presented. The other three are populations A, B and C considered by Yates and Grundy (1953), as in Table 4.1.

Table 4.6: Number of cattle, \( y \), and the number of farms, \( x \), in 13 village blocks of the Dakshina Kannada district of Karnataka (Popn. I).

| \( x \) | 19 28 28 30 31 46 51 55 56 61 64 83 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| \( y \) | 168 326 396 360 331 697 586 739 914 930 619 784 906 |

For the above populations, values of \( b_{opt} \), \( V(\hat{Y}_{HT}) \), \( V(\hat{Y}_{2}) \) and percentage efficiency of \( \hat{Y}_{2} \) relative to \( \hat{Y}_{HT} \), that is \( 100V(\hat{Y}_{HT})/V(\hat{Y}_{2}) \), are given in Table 4.7.

Table 4.7: Efficiency of \( \hat{Y}_{2} \)

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample Size</th>
<th>( b_{opt} )</th>
<th>( V(\hat{Y}_{HT}) )</th>
<th>( V(\hat{Y}_{2}) )</th>
<th>Efficiency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2</td>
<td>8434.8803</td>
<td>( 2227.56 \times 10^{3} )</td>
<td>( 932.27 \times 10^{3} )</td>
<td>239</td>
</tr>
<tr>
<td>II</td>
<td>4</td>
<td>303.5651</td>
<td>( 10.96 \times 10^{3} )</td>
<td>( 6.20 \times 10^{3} )</td>
<td>177</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>9.6781</td>
<td>2.8841</td>
<td>0.0510</td>
<td>5665</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>3.3324</td>
<td>0.3839</td>
<td>0.0530</td>
<td>724</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2.2959</td>
<td>0.2416</td>
<td>0.0821</td>
<td>294</td>
</tr>
</tbody>
</table>
These empirical studies indicate that \( \hat{Y}_2 \) can be considerably more efficient than \( \hat{Y}_{HT} \). For population II the variance of the biased internally congruent estimator suggested by Sampford (1978) was computed and found to be 6948 which is higher than \( V(\hat{Y}_2) \). Finally Table 4.8 shows the effect of deviations from the optimum value of \( b \) on the efficiency of \( \hat{Y}_2 \).

Table 4.8: Sensitivity of efficiency of \( \hat{Y}_2 \) to departures from the optimum choice of \( b \).

| \( 100|1 - b/b_{opt} | \) | Value of efficiency for populations |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| I | II | A | B | C |
| 0 | 239 | 177 | 5655 | 724 | 294 |
| 20 | 226 | 171 | 1755 | 586 | 273 |
| 40 | 200 | 160 | 644 | 389 | 231 |
| 60 | 159 | 139 | 269 | 224 | 173 |
| 80 | 126 | 118 | 152 | 143 | 130 |
| 100 | 100 | 100 | 100 | 100 | 100 |

4.9 Ratio Method of Estimation

Let \( \hat{Y}, \hat{X} \) be unbiased estimators of \( Y \) and \( X \), based on any probability sampling design. Consider the following modification to the ratio estimator \( \hat{Y}_r = \hat{Y}(\hat{X}/\hat{X}) \).

\[
\hat{Y}_3 = (\hat{Y} - Na)(\hat{X}/\hat{X}) + Na
\] \hspace{1cm} (4.17)

The bias and mse of \( \hat{Y}_3 \) are given approximately by
\[ B(Y_3) = Y(\Theta - k)V_{02} \]  
\[ M(Y_3) = Y^2(V_{20} - 20V_{11} + \Theta^2V_{02}) \]

where the \( V_{ij} \) are the relative central moments defined in (2.4),
\[ \Theta = 1 - \frac{N_a}{Y}, \quad \Theta_{opt} = \frac{V_{11}}{V_{02}} = k, \quad \text{and the corresponding} \quad a_{opt} = (1-k)\bar{Y}. \]

For this choice of \( a \), \( B(Y_3) = 0 \). In general there will be reduction in absolute bias relative to that of \( \hat{Y}_r \) when \(|0 - k| < |1 - k|\), that is, \( k \) is closer to \( \Theta \) than to \( 1 \).

In order to have an increase in precision relative to the estimators \( \hat{Y}_r \) or \( \hat{Y}_r' \) it is necessary and sufficient that \( \Theta \) lies between \( 0 \) and \( 2k \), and between \( 2k - 1 \) and \( 1 \). The situation is similar to that in Sec. 3.2.

The implication of these conditions on the choice of \( a \) are given for typical values of \( k \) in Table 4.9.

Table 4.9: Optimum \( a \) and lower and upper bounds on the choice of \( a \) for typical values of \( k \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>Lower bound on ( a )</th>
<th>Optimum ( a )</th>
<th>Upper bound on ( a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.8Y</td>
<td>0.9Y</td>
<td>Y</td>
</tr>
<tr>
<td>0.2</td>
<td>0.6Y</td>
<td>0.8Y</td>
<td>Y</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4Y</td>
<td>0.7Y</td>
<td>Y</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2Y</td>
<td>0.6Y</td>
<td>Y</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.5Y</td>
<td>Y</td>
</tr>
<tr>
<td>0.6</td>
<td>0</td>
<td>0.4Y</td>
<td>0.8Y</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0.3Y</td>
<td>0.6Y</td>
</tr>
<tr>
<td>0.8</td>
<td>0</td>
<td>0.2Y</td>
<td>0.4Y</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>0.1Y</td>
<td>0.2Y</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.2</td>
<td>-0.4Y</td>
<td>-0.2Y</td>
<td>0</td>
</tr>
<tr>
<td>1.4</td>
<td>-0.8Y</td>
<td>-0.4Y</td>
<td>0</td>
</tr>
<tr>
<td>1.6</td>
<td>-1.2Y</td>
<td>-0.6Y</td>
<td>0</td>
</tr>
<tr>
<td>1.8</td>
<td>-1.6Y</td>
<td>-0.8Y</td>
<td>0</td>
</tr>
<tr>
<td>2.0</td>
<td>-2.0Y</td>
<td>-1.0Y</td>
<td>0</td>
</tr>
</tbody>
</table>
This table indicates that there is reasonable flexibility in the choice of $a$, except when $k$ is close to 0 or 1. In these cases the simple estimator $\hat{Y}$ and the ratio estimator $\hat{Y}_r$ may be used respectively.

4.10 Difference Method of Estimation

Define $\hat{W} = \frac{\hat{X}}{\hat{X}}$, $\hat{Z} = \frac{\hat{Y}}{\hat{Y}_k}$ where $\hat{Y}_k$ is a scalar to be chosen.

Consider the difference estimator

$$\hat{Y}_4 = \hat{Y}_k[(\hat{Z} - \hat{W}) + N].$$

(4.20)

This is unbiased for $Y$ since

$$E(\hat{Y}_4) = E\hat{Y}_k[(\hat{Z} - \hat{W}) + N] = \hat{Y}_k[(Y/Y_k - X/X_k) + N] = Y.$$

Further

$$V(\hat{Y}_4) = Y^2\{V_{20} - 2\theta^2V_{11} + \theta^2V_{02}\},$$

(4.21)

with $\theta = Y_k/\sqrt{Y}$. The restrictions on $\theta$ in order to have increased precision relative to that of the ratio estimator are the same as those in the previous section. The implications of these on the choice of $Y_k$ are summarized in Table 4.10.
Table 4.10: Optimum $Y_k$ and lower and upper bounds on the choice of $Y_k$ for typical values of $k$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>Lower bound on $Y_k$</th>
<th>Optimum $Y_k$</th>
<th>Upper bound on $Y_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0</td>
<td>$0.1\bar{Y}$</td>
<td>$0.2\bar{Y}$</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>$0.2\bar{Y}$</td>
<td>$0.4\bar{Y}$</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>$0.3\bar{Y}$</td>
<td>$0.6\bar{Y}$</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>$0.4\bar{Y}$</td>
<td>$0.8\bar{Y}$</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>$0.5\bar{Y}$</td>
<td>$1.0\bar{Y}$</td>
</tr>
<tr>
<td>0.6</td>
<td>$0.2\bar{Y}$</td>
<td>$0.6\bar{Y}$</td>
<td>$\bar{Y}$</td>
</tr>
<tr>
<td>0.7</td>
<td>$0.4\bar{Y}$</td>
<td>$0.7\bar{Y}$</td>
<td>$\bar{Y}$</td>
</tr>
<tr>
<td>0.8</td>
<td>$0.6\bar{Y}$</td>
<td>$0.8\bar{Y}$</td>
<td>$\bar{Y}$</td>
</tr>
<tr>
<td>0.9</td>
<td>$0.8\bar{Y}$</td>
<td>$0.9\bar{Y}$</td>
<td>$\bar{Y}$</td>
</tr>
<tr>
<td>1.0</td>
<td>$1.0\bar{Y}$</td>
<td>$1.0\bar{Y}$</td>
<td>$\bar{Y}$</td>
</tr>
<tr>
<td>1.2</td>
<td>$\bar{Y}$</td>
<td>$1.2\bar{Y}$</td>
<td>$1.4\bar{Y}$</td>
</tr>
<tr>
<td>1.4</td>
<td>$\bar{Y}$</td>
<td>$1.4\bar{Y}$</td>
<td>$1.8\bar{Y}$</td>
</tr>
<tr>
<td>1.6</td>
<td>$\bar{Y}$</td>
<td>$1.6\bar{Y}$</td>
<td>$2.2\bar{Y}$</td>
</tr>
<tr>
<td>1.8</td>
<td>$\bar{Y}$</td>
<td>$1.8\bar{Y}$</td>
<td>$2.6\bar{Y}$</td>
</tr>
<tr>
<td>2.0</td>
<td>$\bar{Y}$</td>
<td>$2.0\bar{Y}$</td>
<td>$3.0\bar{Y}$</td>
</tr>
</tbody>
</table>

Again there is sufficient flexibility in the choice of $Y_k$ except when $k$ is close to 0 or 1. The particular case of srs with design for the estimators $\hat{Y}_3$, $\hat{Y}_4$ has been discussed by Srivenkataramana (1978), and Srivenkataramana and Srinath (1982).
CHAPTER 5.

DOUBLE SAMPLING WITH PPS SELECTION

[Raj (1965) has proposed a pps selection scheme for estimating the population total in the absence of information on the auxiliary variate, but when information on some other variate is available. This requires the knowledge of a certain parameter h. But in practice the value of h may not easily be known. Accordingly this chapter proposes an alternative scheme which does not need the knowledge of h.]

5.1 Introduction

Let \( y \) and \( x \) be the study and auxiliary variates; \((y_i, x_i)_i\), \(i = 1, \ldots, N\) being real values, which are however unknown. Let \( z \) be another variate for which the value is known for each unit in the population. It is assumed that \( x_i, z_i > 0 \) for \( i = 1, \ldots, N \). For instance, in estimating the total yield of wheat in a village having \( N \) farms, the yield and area under the crop in each farm are likely to be unknown. But the total area in each farm may be known from village records or may be obtained at a low cost. Then \( y, x \) and \( z \) are respectively yield, area under wheat and area under cultivation.
In this context Raj (1965) has proposed the following scheme.

**Scheme I**

Suppose an initial sample of \( n' \) units is selected with probabilities

\[
p_i = z_{i1}/Z; \quad i = 1, \ldots, N,
\]

where \( Z = \sum z_{i1} \), and information on \( x \) is collected. And a subsample of size \( n \) is selected from the initial sample with equal probabilities (wor) and information on \( y \) is collected.

**Theorem 5.1** Under scheme I, if

\[
\hat{Y} = \frac{1}{n} \sum_{i=1}^{n} (y_{i1}/p_i) - \frac{h}{n} \sum_{i=1}^{n} (x_{i1}/p_i) + \frac{h}{n'} \sum_{i=1}^{n'} (x_{i1}/p_i)
\]

then \( \hat{Y} \) is unbiased for \( Y \), and

\[
V(\hat{Y}) = S_1^2/n + (1/n - 1/n') (h^2 S_1^2 - 2hd S_1 S_2)
\]

where

\[
S_1^2 = \sum_{i=1}^{N} (y_{i1}/p_i - Y)^2 p_i,
\]

\[
S_2^2 = \sum_{i=1}^{N} (x_{i1}/p_i - X)^2 p_i,
\]

\[
d = [\sum (y_{i1}/p_i - Y)(x_{i1}/p_i - X) p_i]/S_1 S_2
\]

and \( h \) is a scalar. \( V(\hat{Y}) \) is a minimum when \( h_{\text{opt}} = d S_1/S_2 \), and

\[
\min V(\hat{Y}) = S_1^2 (1-d^2)/n + S_2^2 d^2/n'.
\]

5.2 **An Alternative Scheme**

In order to compute \( \hat{Y} \) from (5.1), we have to guess the value of \( h \).
When this is difficult the following alternative scheme is suggested which avoids the necessity of \( h \).

**Scheme II**

Select the initial sample, \( s' \), of size \( n' \) with probability proportional to \( z \) and collect information on \( x \) as in scheme I. And then select a subsample, \( s \), of size \( n \) with probability proportional to \( x_i / z_i \), with replacement.

**Theorem 5.2** Under scheme II, if \( Z = \sum z_i \), and

\[
\hat{Y}^* = \left( Z / nn' \right) \sum y_i / x_i \cdot \sum x_i / z_i
\]

then \( \hat{Y}^* \) is unbiased for \( Y \), and

\[
V(\hat{Y}^*) = S_1^2 / n' + S_3^2 \left( n' - 1 \right) / nn'
\]

where

\[
S_3^2 = \sum \left( y_i / x_i - Y \right)^2 / x_i / X
\]

An unbiased estimate of \( V(\hat{Y}^*) \) is provided by

\[
v(\hat{Y}^*) = \left( \sum \omega_i^2 / p_i^2 - nn' \hat{Y}^2 \right) / nn'(n' - 1)
\]

\[+ \sum \left( \omega_i / p_i^2 - \sum \omega_i / np_i^2 \right)^2 / (n' - 1) nn' (n' - 1)
\]

where \( \omega_i = y_i / p_i \) and \( p_i^* = x_i / z_i \cdot \sum (x_i / z_i) \).

**Proof:** Consider \( Y^* = \left( Z / nn' \right) \sum y_i / x_i \cdot \sum x_i / z_i \)

\[= \left( Z / nn' \right) \sum \left( y_i / z_i \right) / \sum \left( x_i / z_i \right)
\]

\[= \frac{1}{nn'} \sum \left( y_i / z_i \right) / \sum \left( x_i / z_i \right), \text{ since } \frac{z_1}{Z} = p_i
\]

\[= \left( 1 / nn' \right) \sum \omega_i / p_i^*
\]
where \( \omega_i = y_i / p_i \) and \( p^\ast_i = x_i / \left\{ \sum_{i=1}^{n'} (x_i/z_i) \right\} \).

Therefore,
\[
E(Y^\ast) = E_1 E_2 (Y^\ast | s') \\
= \frac{N}{n'} E_1 \{ \Sigma \omega_i \} / n' = \frac{N}{n'} E_1 \{ \sum_{i=1}^{n'} y_i / p_i \} / n' = \Sigma y_i = \Sigma y_i = Y.
\]

Hence \( Y^\ast \) is unbiased for \( Y \). Next recall the standard result
\[
V(Y^\ast) = V_1 E_2 (Y^\ast | s') + E_1 V_2 (Y^\ast | s').
\]

Consider
\[
V_1 E_2 (Y^\ast | s') = \frac{N}{n'} \{ \sum_{i=1}^{n'} y_i / p_i - Y \} p_i / n' = \frac{S_1^2}{n'}, \tag{5.6}
\]

and
\[
V_2 (Y^\ast | s') = \frac{N}{n'} \{ \sum_{i=1}^{n'} (\omega_i / p_i - \omega_i^\ast) p_i / n' \}
= (1 / n n') ^{2} \{ \sum_{i=1}^{n'} \omega_i^2 / p_i - (\omega_i^\ast)^2 \}.
\]

Writing
\[
\sum_{i=1}^{n'} \omega_i^2 / p_i = \sum_{i=1}^{n'} (y_i^2 / x_i z_i) \sum_{i=1}^{n'} (x_i / z_i)
= \sum_{i=1}^{n'} \sum_{j=1}^{n'} (y_i^2 / x_i p_i) (x_j / p_j) + \sum_{i=1}^{n'} y_i^2 / p_i,
\]

and \( \Sigma \omega_i = n' E_2 (Y^\ast) \) we have
\[
E_1 V_2 (Y^\ast | s') = \frac{n'(n'-1)}{n'} \sum_{i=1}^{n'} \sum_{j=1}^{n'} (y_i^2 / x_i z_i) / x_i
+ \frac{n'}{n'} \sum_{i=1}^{n'} (y_i^2 / p_i) - \frac{n'}{n'} Y - \frac{n'}{n'} V_1 E_2 (Y^\ast) / n n'^2
\]
\[
= \frac{n'(n'-1)}{n'} \sum_{i=1}^{n'} (y_i^2 / x_i) + n' \sum_{i=1}^{n'} (y_i^2 / p_i) - n'^2 y^2
- \frac{n'}{n'} \sum_{i=1}^{n'} (y_i^2 / p_i) + n' y^2 / n n'^2
\]
\[
= (n'-1) \frac{S_2^2}{n n'}. \tag{5.7}
\]
Hence from (5.6) and (5.7), \( V(\hat{Y}^*) = \frac{S_1^2}{n'} + (n'-1) \frac{S_3^2}{nn'} \), which establishes (5.4).

Next we have to show that the expression given in (5.5) is an unbiased estimator of \( V(\hat{Y}^*) \). Denote the first term on the right side of (5.5) by \( v_1 \) and the second by \( v_2 \). Then

\[
E(v_2) = E(\sum_{i=1}^{n} \frac{\omega_i^2}{p_i^* + \eta_i I_i^*} - \frac{nn'\hat{Y}^*}{nn'})/(nn' - 1) \\
= \left( E(\sum_{i=1}^{n} \frac{\omega_i^2}{p_i^* + \eta_i I_i^*}) - E(\frac{\hat{Y}^*}{\hat{Y}^*}) \right)/(nn' - 1) \\
= \left( \sum_{i=1}^{n} \frac{\omega_i^2}{p_i} - \frac{\hat{Y}^*}{\hat{Y}^*} - \frac{V(\hat{Y}^*)}{V(\hat{Y}^*)} \right)/(nn' - 1) , \text{ using } V(\hat{Y}^*) = E(\frac{\hat{Y}^*}{\hat{Y}^*}) - \left( E(\frac{\hat{Y}^*}{\hat{Y}^*}) \right)^2 \\
= \left( S_1^2 - V(\hat{Y}^*) \right)/(nn' - 1) \\
= \frac{S_1^2}{n'} - \frac{S_3^2}{nn'} . \quad (5.8)
\]

Finally, \( V_2(\hat{Y}^*|s') \) is estimated by

\[
T = \sum_{i=1}^{n} \left( \frac{\omega_i}{p_i^*} - \frac{\omega_i}{n\eta_i} \right)^2 / n(n-1)n'2
\]

and therefore \( T \) also estimates \( E_1 V_2(\hat{Y}^*|s') = (n'-1)S_3^2/nn' \). Hence

\[
v_2 = \frac{n'T}{(n'-1)} \text{ estimates } S_3^2/n . \quad (5.9)
\]

From (5.8) and (5.9) it follows that \( v_1 + v_2 \) provides an unbiased estimator of \( V(\hat{Y}^*) \). This completes the proof.

5.3 Efficiency of \( \hat{Y}^* \)

Consider

\[
E = \min \frac{V(\hat{Y})}{V(\hat{Y}^*)} = \frac{(1-d^2)n' + d^2n}{n + (n'-1)d}
\]

where \( D = \frac{S_3^2}{S_1^2} \). This reflects the efficiency of \( \hat{Y}^* \) over the estimator \( \hat{Y} \).
In particular when \( \frac{y_i}{x_i} \) is a constant for all the population units,

\[ S^2_3 = 0 \] (hence \( D = 0 \)) and \( d = 1 \) so that \( E = 1 \). In general denoting \( n'/n \) by \( r \), the expression for \( E \) can be written as

\[
E = \frac{(1-d^2)rn + d^2n}{(n + (rn-1)D)}
\]

\[ \approx \frac{(1-d^2)r + d^2}{(1+rD)} \]  \( (5.10) \)

The values of \( E \) computed from \((5.10)\) for typical values of \( r \) and \( D \) are given in Table 5.1.

Table 5.1: Values of \( E \) for typical values of \( r, D \) and \( d \).

<table>
<thead>
<tr>
<th>( d )</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>0.1</td>
<td>1.17</td>
<td>1.63</td>
<td>2.03</td>
<td>2.37</td>
<td>2.67</td>
<td>1.35</td>
<td>2.12</td>
<td>2.79</td>
<td>3.38</td>
<td>3.88</td>
</tr>
<tr>
<td>0.2</td>
<td>-</td>
<td>1.22</td>
<td>1.52</td>
<td>1.78</td>
<td>2.00</td>
<td>-</td>
<td>1.41</td>
<td>1.86</td>
<td>2.25</td>
<td>2.58</td>
</tr>
<tr>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>1.22</td>
<td>1.42</td>
<td>1.60</td>
<td>-</td>
<td>1.06</td>
<td>1.40</td>
<td>1.69</td>
<td>1.94</td>
</tr>
<tr>
<td>0.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.01</td>
<td>1.19</td>
<td>1.33</td>
<td>-</td>
<td>-</td>
<td>1.12</td>
<td>1.35</td>
</tr>
<tr>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.02</td>
<td>1.14</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.13</td>
</tr>
<tr>
<td>0.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.11</td>
</tr>
</tbody>
</table>

- denotes values of \( E \) less than 1.

For small values of \( d \) and \( D \), the estimator \( \hat{Y} \) performs well. Precisely in this case guessing \( h_{opt} \) is difficult and \( \hat{Y} \) can be used rather than \( \hat{Y} \).

5.4 Cost Function

If \( c' \) and \( c \) denote the unit costs of collecting information on
x and y respectively, the total cost under scheme I or II would be

$$C = c'n + cn.$$  \hspace{1cm} (5.11)

On the other hand if a straight probability-proportional to z sample
is taken for y (Scheme III) the sample size for the same total cost
will be

$$n_0 = (c'n + cn)/c$$

and variance of the estimator of the
population total will be

$$S^2/n_0.$$  \hspace{1cm} (5.12)

Comparing (5.4) with (5.12), the condition that scheme II is better than
scheme III under cost function (5.11) becomes

$$D < \frac{n}{n^* - 1} \left( \frac{c(n^* - n) - c'n}{c'n + cn} \right).$$

Roughly this reduces to

$$D < \frac{e^{(2g)} - a}{r^2 + ra}, \text{ where } a = c/c'.$$

As an example, let $r = 10, a = 40$ then $D$ should be less than 0.7,
that is $S^2 < 0.7S^2_1$.

**Numerical Illustration.** In order to compare the efficiency of $\hat{Y}$ with
that of $\hat{Y}$ computed from (5.1) with $h_{opt}$, the data relating to 34
villages given in Murthy (1967) were treated as making up the population,
where $y$, $x$ and $z$ respectively denote area under wheat in 1964, in 1963
and cultivated area in 1961. For this population the values of the
required parameter $E(z)$ given in Table 5.2. The values of $E(z)$ for
typical values of $z$ are given in Table 5.2.
Table 5.2: Parameters of the population in the illustration

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{1}^{2}$</td>
<td>$4421.81 \times 10^{3}$</td>
</tr>
<tr>
<td>$S_{3}^{2}$</td>
<td>$826.06 \times 10^{3}$</td>
</tr>
<tr>
<td>$S_{2}^{2}$</td>
<td>$2165.27 \times 10^{3}$</td>
</tr>
<tr>
<td>d</td>
<td>0.6404</td>
</tr>
<tr>
<td>D</td>
<td>0.1868</td>
</tr>
</tbody>
</table>

Table 5.3: Values for $E$ for typical $r$ for the population in the illustration

<table>
<thead>
<tr>
<th>r</th>
<th>E (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>116</td>
</tr>
<tr>
<td>3</td>
<td>140</td>
</tr>
<tr>
<td>4</td>
<td>159</td>
</tr>
<tr>
<td>5</td>
<td>174</td>
</tr>
<tr>
<td>6</td>
<td>186</td>
</tr>
</tbody>
</table>

It is seen that $E$ increases as subsampling rate decreases. Therefore the proposed scheme will be advantageous especially when the information on $y$ is expensive as compared to information on $x$. 
CHAPTER 6.

USE OF A POSITIVE AND NEGATIVE
VALUED AUXILIARY VARIATE IN SURVEYS

Auxiliary variates in surveys are occasionally positive and negative valued. This introduces difficulties in ratio and product methods of estimation and in pps sampling. Two simple methods of dealing with the situation are outlined: stratification by sign and transformation by simple translation. For illustration purpose srswor and ppswr designs are assumed.

6.1 Introduction

In Chapters 2 to 5 it was mostly assumed that the auxiliary variate was non-negative since survey variates are generally so. But exceptions do occur. Rates of change, elasticities of supply and demand, and variates arising as differences, for example, saving (= income - expenditure) and profit (= revenue - cost) are some cases in point. Such variates can not be used directly as (i) the auxiliary variate in ratio and product methods of estimation, or (ii) the measure of size in pps sampling. The former is due to that the population mean and/or the sample mean of the variate may be close to zero which could easily happen when the distribution of the variate in the population is nearly symmetric.
The other difficulty is because of the non-positive values. This chapter suggests (i) a stratification of the population according to the sign of the value of the auxiliary variate, and as an alternative (ii) a change of origin for the auxiliary variate to make it positive and a corresponding change for the study variate in order to have proportionality between the values of the two variates.

6.2 The Stratification Approach

The values $x_i, y_i$ are assumed to be real, but they may be positive, negative or zero. The $x_i$ are known. Consider a stratification of the population according to the sign of $x_i$. That is, unit $i$ is allotted to stratum 1 if $x_i > 0$; otherwise to stratum 2. The sample size $n$ is allocated to these strata, for example, by proportional allocation. Let $n_h$ denote the size of the sample and $\bar{x}_h, \bar{y}_h$ denote the sample means for stratum $h$ under a srs with replacement scheme. Assume that the selections from the two strata are independent. Let $\bar{x}_h, \bar{y}_h$ be the means of $x$ and $y$ in stratum $h$. Then an estimator of $Y_h$, the stratum total for $y$, is

$$t_h = n_h \bar{y}_h \left( \frac{\bar{x}_h}{x_h} \right).$$

And $Y$ may be estimated by

$$Y_1 = t_1 + t_2.$$  (6.1)

This is the usual 'separate' ratio estimator. If the $n_h$ are at least moderate then the mse of $t_h$ is given approximately by

$$M(t_h) = n_h (N_h - n_h) \left( S_y^2 + \frac{R_h^2 S_x^2}{n_h} - 2R_h S_{y|xh} \right) / n_h.$$  (6.2)
where $S^2_y$, $S^2_x$, $S_{yx}$ are variances and covariance in stratum $h$ and

$$R_h = \frac{\bar{y}_h}{\bar{x}_h}.$$  And

$$M(Y_1) = M(t_1) + M(t_2). \quad (6.3)$$

It is known that a ratio estimator $t_h$ performs well if

$$(B_h/R_h) > 1/2$$

where $B_h$ is the coefficient of regression of $y$ on $x$ in stratum $h$. It is assumed above that this is in fact the case. On the other hand if $(B_h/R_h) < -1/2$ a product-type estimator $t_h = N_h \bar{y}_h (\bar{x}_h/\bar{X}_h)$ may be formed in that stratum. And if $B_h/R_h$ is in $[-1/2, 1/2]$ the simple estimator $N_h \bar{y}_h$ itself may be used as $t_h$. Of course, the expression for $M(t_h)$ in (6.2) should be modified accordingly. Thus the stratification approach allows a choice among ratio, product and simple estimators in each of the two strata. This is handy when the nature and extent of dependence of $y$ on $x$ differ in the two cases: $x_i \geq 0$ and $x_i < 0$.

If stratification by sign is difficult, a strategy of post stratification might be appropriate, particularly as the problem of zero stratum sample sizes is unlikely to occur.

6.3 A Change of Origin

As an alternative to stratification consider the following approach. Suppose that a srswor of $n$ units is drawn from the entire population, the value of $x$ and $y$ obtained for these units and the sample means $\bar{x}, \bar{y}$ computed. Then as an estimator of $Y$ the ratio method uses

$$\hat{Y}_r = Ny(\bar{X}/\bar{x}), \quad (6.4)$$
while the product method uses
\[ \hat{Y}_p = Ny(x/X) . \] 

(6.5)

The estimator \( \hat{Y}_r \) is an improvement over the simple estimator \( \bar{Y} \) if \( (B/R) > 1/2 \), where \( B \) is the coefficient of regression of \( y \) on \( x \) and \( R = \bar{Y}/\bar{X} \). The estimator \( \hat{Y}_p \) may be used when \( B/R < -1/2 \). And when \( B/R \) is in \([-1/2, 1/2]\) the estimator \( \bar{Y} \) itself is preferred. However, it is apparent from (6.4) and (6.5) that the use of \( \hat{Y}_r \) or \( \hat{Y}_p \) should be avoided when \( \bar{x} \) and/or \( \bar{x} \) is likely to be close to zero. This can easily be the case when \( x \) assumes both positive and negative values.

In this context consider the transformation
\[ w = x + c , \quad z = y + \Theta . \] 

(6.6)

The scalar \( c \) is chosen such that \( w \) is a positive valued variate. For instance, let \( x_1 \) be the smallest \( x \)-value and assume that it is negative. Then \( c \) must satisfy \( |x_1| < c < \infty \). Next, the choice of \( \Theta \) is made so as to control the mse of the estimator. This is discussed in Sec. 6.4.

Now let \( \bar{w} = \bar{x} + c \), \( \bar{z} = \bar{y} + \Theta \), \( \bar{w} = \bar{x} + c \), \( \bar{z} = \bar{y} + \Theta \). Then the usual ratio estimator of the total \( Z \) is
\[ \hat{Z}_r = N\bar{w} (\bar{w}/\bar{w}) = N(y+\Theta)((\bar{x}+c)/(\bar{x}+c)) , \] 

(6.7)

and hence \( Y \) may be estimated by
\[ \hat{Y}_r = \hat{Z}_r - N\Theta . \] 

(6.8)

The transformations (6.6), being only changes of origin, leave the
variances, covariances and correlations unchanged. Also the standard theory for ratio estimators applies to \( \hat{Z}_r \), except that in the place of \( R \) we now have \( R_1 = (\bar{Y} + \Theta)/(\bar{X} + c) \). Thus the bias and mse of \( \hat{Y}_2 \) as an estimator of \( Y \) are, up to second order moments

\[
B(\hat{Y}_2) = (N - n) \left( R_1 S_x^2 - S_{xy}^2 \right) / n (\bar{X} + c),
\]

and

\[
M(\hat{Y}_2) = N(N - n) \left( S_y^2 + R_1^2 S_x^2 - 2 R_1 S_{xy} \right) / n,
\]

where \( S_y^2 \) is the population variance of \( Y \), etc.

In order to use \( \hat{Y}_2 \) in practice we expect that it is more precise than \( NY \). This will be the case when \( (B/R_1) > 1/2 \). Also the introduction of the parameter \( \Theta \) will be justified only if the estimator is more precise than that without \( \Theta \), that is than that with \( \Theta = 0 \). It can be shown that these two considerations require that \( \Theta \) lies

between \( \bar{Y} \) and \( D - \bar{Y} \),

and

between \( 0 \) and \( D - 2\bar{Y} \),

where \( D = 2B (\bar{X} + c) \). The second condition ensures also a simultaneous reduction in the absolute bias relative to the \( \Theta = 0 \) case. For a given \( c \), the value of \( \Theta \) minimizing \( M(\hat{Y}_2) \) is provided by \( R_1 = B \) which implies \( \Theta_{opt} = (D/2) - \bar{Y} \). In this case \( \hat{Y}_2 \) is almost unbiased for \( Y \).

And

\[
M(\hat{Y}_2) = N(N - n) S_y^2 \left( 1 - \rho^2 \right) / n.
\]

This mse is the same as the large sample approximation to the mse of
the linear regression estimator. In the special case \( y_i = Bx_i \) for all \( i \), \( \theta_{\text{opt}} = Bc \) and \( \hat{z}_r \) reduces to \( NB(\bar{x}+c) \) so that \( \hat{y}_2 = NB\bar{x} \). Hence \( \hat{y}_2 \) estimates \( Y = NB\bar{x} \) without any error just like \( \hat{y}_r \) in this situation.

6.4 The Choice of \( \theta \)

The optimum \( \theta \) is \( B(\bar{x}+c) - \bar{y} \). Here \( \bar{x} \) and \( c \) are known, but generally \( B \) and \( \bar{y} \) are not. A geometric interpretation of \( \theta_{\text{opt}} \) is that it is the distance between the \( x \) and \( w \) axes such that the population regression of \( z \) on \( w \) is through the origin (See fig. 6.1). Past experience and data from a pilot study may be used in finding approximations to \( \theta_{\text{opt}} \) as commonly done in such situations. Also since \( \theta \) is needed only at the estimation stage, the knowledge of \( \bar{y} \) and a scatter diagram for at least a part of the sample data on \( y \) and \( x \) may be helpful in this regard. In general (6.11) and (6.12) specify two intervals in which \( \theta \) should lie in order that \( \hat{y}_2 \) is more precise than \( \hat{y}_r \) or \( \hat{y}_2 \) with \( \theta = 0 \). The midpoint of these intervals is \( \theta_{\text{opt}} \), and the widths are \( |D| \) and \( |D-2\bar{y}| \) respectively.

If \( (B/R_1) \leq 1/2 \), a product-type estimator
\[
\hat{y}_3 = N(\bar{y}+\theta)\left[(\bar{x}+c)(\bar{x}+c)\right] - N\theta
\]  
(6.13)

may be used. Here \( \theta_{\text{opt}} = -B(\bar{x}+c) - \bar{y} \).

6.5 PPS Sampling

Again consider the transformation (6.6), which in the first place makes \( w \) a positive valued variate. Therefore it can be used as a
Fig. 6.1 Geometric interpretation of $\Theta_{opt}$
measure of size. With this suppose a ppswr sample of \( n \) units is drawn. Then an unbiased estimator of \( Z \) is
\[
\hat{Z}_{\text{pps}} = (W/n) \sum_{i=1}^{n} (z_i/\omega_i)
\]
and hence an unbiased estimator of \( Y \) is
\[
\hat{Y}_4 = (W/n) \sum_{i=1}^{n} \frac{(z_i/\omega_i)}{\omega_i} - NO = \left\{ \frac{N(\bar{X}+c)}{n} \right\} \frac{N}{n} \sum_{i=1}^{n} \frac{(y_i+O)/(x_i+c)}{(\bar{X}+O)} - NO .
\]
(6.14)

And
\[
V(\hat{Y}_4) = \frac{N(\bar{X}+c)}{n} \frac{N}{n} \sum_{i=1}^{n} \left( \frac{(y_i+O)/(x_i+c)}{(\bar{X}+O)} \right)^2 - \frac{N^2}{n} \left( \frac{(\bar{Y}+O)}{(\bar{X}+O)} \right)^2 .
\]

It is natural to expect that \( \hat{Y}_4 \) should be more precise than either \( \hat{Y}_4 \) or \( \hat{Y}_4 \) with \( O = 0 \). For a given \( c \), the former requires (Raj, 1968, p.50)
\[
\frac{N}{n} \sum_{i=1}^{n} \frac{(\omega_i - \bar{W})z_i^2/\omega_i}{\bar{W}} > 0
\]
(6.15)
while the latter needs
\[
O \text{ to lie between } 0 \text{ and } 2\Theta^* .
\]
(6.16)
where
\[
\Theta = \frac{\bar{W} - \bar{W}}{R_0} = \frac{\bar{W}}{\bar{W}}
\]
(6.17)
with \( R_0 = \frac{\bar{Y} - \bar{W}}{\bar{W} - \bar{W}} \).

The condition (6.15) implies that \( w \) and \( z^2/\omega \) are positively correlated.
The other condition (6.16) for \( O \) is the same as (4.7) for 'a'. And \( V(\hat{Y}_4) \) is minimized for given \( c \) when \( O = \Theta^* \). When \( y_i = A + Bx_i \) for all \( i \), \( \Theta^* \) reduces to \( \Theta_{\text{opt}} \) of the previous sections. Thus when the regression of \( y \) on \( x \) is linear, \( \Theta^* \) can be interpreted as an approximation
to $\theta_{\text{opt}}$. Refer figure 6.1.

If the choice of $\theta$ is felt to be difficult, one may use the stratification suggested in Sec. 6.2, and use $|x_i|$ as the measure of size provided no $x_i = 0$. The strata totals for $y$ may be estimated separately and then added to give an estimate of $Y$. 
CHAPTER 7.

SCOPES FOR FURTHER WORK

7.1 General Points

1. A product-type estimator has been suggested either as a dual or as an alternative to the commonly used ratio estimator in Chapters 2 and 3. The conditions for the optimality of the latter are known. The corresponding conditions for the product estimator based on harmonic means are given in Section 2.6. However, the optimality of a product estimator based on arithmetic means is to be investigated.

2. The performance of different estimators proposed in Chapter 2 to 6 may be studied by imposing on $y$ a superpopulation dependent on $x$. Specific assumptions regarding the distribution of $x$ may be made. For example, we may postulate that

$$y_i = \lambda + B x_i + e_i$$

(7.1)

where $E(e_i | x_i) = 0$, and the $x_i$ have independent and identical gamma distributions. For instance, following Srivastava’s (1978) work, Chaudhuri and Adhikary (1979) have examined the use of variate transformations in improving the efficiencies of some sampling strategies involving selections with varying probabilities. They have assumed regression models with and without specified forms of distributions for auxiliary variates.

3. The different estimators suggested in Chapters 2 to 6 may be generalized to incorporate auxiliary information on several variates. In
order to be practicable, these methods should be simple for application. The generalization suggested in Sec. 3.4 of forming a weighted combination of estimators, each using auxiliary information on a single $x$, may not be always effective.

4. The use of auxiliary information in estimating proportions and other features of interest may be investigated. Some work in this regard has been reported. For example Wynn (1976), Rao (1977) and Das (1979).

5. The Midzuno-Sen scheme was employed in Sections 4.5 to 4.8 to demonstrate a change of origin for the study variate for improving the efficiency of the Horvitz-Thompson estimator. Generalizations applying to any ppswor scheme may be attempted. One line of approach is the following. Consider probability proportional to $x$ sampling with $p_i^* = x_i/X, i = 1, \ldots, N$, as initial probabilities where $X = \sum x_i$. Let the first and second order inclusion probabilities be denoted by $\pi_i$ and $\pi_{ij} (i \neq j)$. Then $\pi_i$ can be written as

$$\pi_i = p_i + p_i^* ; i = 1, \ldots, N,$$

(7.2)

where $p_i^*$ is the probability of selecting $i^{th}$ unit in the second or subsequent selections. The familiar Horvitz-Thompson estimator of $Y$ is

$$\hat{Y}_{HT} = z' E,$$

(7.3)

where $z' = (y_1/\pi_1, \ldots, y_N/\pi_N)$, $E' = (d_1, \ldots, d_N)$ and $d_i = 1$ if unit $i$ is included in the sample and zero otherwise. The variance of $\hat{Y}_{HT}$ is

$$V(\hat{Y}_{HT}) = z' W z,$$

(7.4)
where \( W = (\omega_{ij})_{N \times N} \) with \( \omega_{ii} = \pi^*_i (1 - \pi^*_i) \); \( \omega_{ij} = \pi^*_i x_j \); \( i \neq j = 1, \ldots, N \).

If \( y^*_i \) is proportional to \( \pi^*_i \) for all \( i \) then the estimator \( \hat{\pi}^*_{HT} \) reduces to a constant and hence \( \hat{V}(\hat{\pi}^*_{HT}) = 0 \). But from (7.2) we note that if \( p^*_i \) is not proportional to \( p^*_i \) for at least one \( i \), then \( y^*_i \) is not proportional to \( \pi^*_i \) even if \( y^*_i = \beta x_i \) for all \( i \). In this context consider the transformation

\[
y^*_i = y^*_i + b p^*_i ; \quad i = 1, \ldots, N
\]

where \( b \) is a scalar to be chosen.

Then an unbiased estimator of \( Y \) is

\[
\hat{Y}^*_{HT} = \bar{Z}^* E - b \sum_{i=1}^{N} p^*_i = \bar{Z}^* E - b \,(n-1) ,
\]

since for any scheme \( \sum_{i=1}^{N} p^*_i = \sum_{i=1}^{N} (\pi^*_i - p^*_i) = n - 1 \).

Here \( \bar{Z}^* = (y^*_1/\pi^*_1, \ldots, y^*_N/\pi^*_N) \). And

\[
\hat{V}(\hat{Y}^*_{HT}) = \bar{Z}^* W \bar{Z}^* = \bar{Z}'WZ + 2bZ'WP + b^2P'WP ,
\]

where \( P' = (p^*_1/\pi^*_1, \ldots, p^*_N/\pi^*_N) \). This variance is a minimum when

\[
b_{opt} = - (Z'WP)/(P'WP) ,
\]

and

\[
\min \hat{V}(\hat{Y}^*_{HT}) = \bar{Z}'WZ - (Z'WP)^2/(P'WP).
\]

Comparing (7.4) with (7.7), the restriction on \( b \) in order that \( \hat{Y}^*_{HT} \) is at least as precise as \( \hat{Y}^*_{HT} \) is

\[
0 \leq b \leq 2b_{opt} .
\]
The choice of \( b \) may be made on the lines suggested in Sec. 4.7.

6. The auxiliary variates occasionally assume negative and positive values, as pointed out in Chapter 6. Two problems introduced by this were mentioned and two ways of overcoming these problems were outlined. Other methods of tackling these problems may be developed.

7. The auxiliary information may be incomplete in several respects. For example, with multiple auxiliary variates, it is quite likely that the values of these are not available for some of the population units. Some genuine techniques for utilizing the available information to the extent possible may be developed. Some efforts in this direction are in Han (1973) and Singh (1977).

8. The auxiliary information is generally used for improving the estimates of population mean, total, ratio etc. The possibility of using such information for obtaining better estimates of variance of these estimates may be investigated.

7.2 The Predictive Approach

Suppose \( s \) denotes a sample of \( n \) units from the finite population under consideration. Then any sampling procedure divides the population into a completely known (in respect of the character \( y \) being observed), and a completely unknown part. If we write

\[
Y = \sum_{i \in s} y_i + \sum_{i \notin s} y_i = y^{(1)} + y^{(2)} \tag{7.11}
\]

then the first term \( y^{(1)} \) of (7.11) is known. Therefore, essentially a predictor of \( y^{(2)} \) is needed for estimating \( Y \).
With the above premises Srivenkatarama and Tracy (1979) have suggested four methods of estimating the population total \( Y \), given a simple random sample of \( n \) units without replacement from that population. Two of these methods are suited for the case of positive correlation between \( y \) and an auxiliary variate \( x \), and the other two for the case of negative correlation.

Let \( X^{(1)} \), \( X^{(2)} \) denote the totals of the \( x \)-variante values for the sample and nonsample units respectively and

\[
\hat{Y} = \frac{NY^{(1)}}{n}, \quad \hat{X} = \frac{NX^{(1)}}{n} \tag{7.12}
\]

be the (unbiased) simple expansion estimators of \( Y \) and \( X \) respectively, constructed from the sample. Then the suggested estimators are the following.

(i) **Case of positive correlation**

\[
\hat{Y}_1 = Y^{(1)} + (\hat{Y} - Y^{(1)}) X^{(2)} / gX = Y^{(1)}(1 + X^{(2)} / fX), \tag{7.13}
\]

\[
\hat{Y}_2 = Y^{(1)} + (\hat{Y} - Y^{(1)}) \hat{X} / \hat{X} = Y^{(1)}(1 + gX / X^{(1)}). \tag{7.14}
\]

(ii) **Case of negative correlation**

\[
\hat{Y}_3 = Y^{(1)} + (\hat{Y} - Y^{(1)}) gX / X^{(2)} = Y^{(1)}(1 + gX / fX^{(2)}), \tag{7.15}
\]

\[
\hat{Y}_4 = Y^{(1)} + (\hat{Y} - Y^{(1)}) \hat{X} / X = Y^{(1)}(1 + gX^{(1)} / f^2 X^{(1)}). \tag{7.16}
\]

Here \( f = n/N \) is the sampling fraction and \( g = 1 - f \).

The expressions on the extreme right in (7.13) and (7.14) exhibit the near duality of the estimators \( \hat{Y}_1 \) and \( \hat{Y}_2 \). Here the unknown \( Y^{(2)} \) has
been predicted to be \((\hat{Y} - Y^{(1)})\) modified by the factor \(X^{(2)}/gX\) or \(X/\hat{X}\). Similar is the case with \(\hat{Y}_3\) and \(\hat{Y}_4\). The guidelines for choosing among the estimators \(\hat{Y}, \hat{Y}_1, \hat{Y}_2\) and the usual ratio estimator \(\hat{Y}_r\) are in Table 7.1. And the guidelines in the case of negative correlation for choosing among \(\hat{Y}, \hat{Y}_3, \hat{Y}_4\) and the product estimator \(\hat{Y}_p\) are in Table 7.2.

Table 7.1: Guidelines in the positive correlation case

<table>
<thead>
<tr>
<th>Interval for (k = V_{11}/V_{02})</th>
<th>Preferred estimator</th>
<th>Interval for (k)</th>
<th>Preferred estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0 \leq k \leq f/2)</td>
<td>(\hat{Y})</td>
<td>(0 \leq k \leq g/2)</td>
<td>(\hat{Y})</td>
</tr>
<tr>
<td>(f/2 &lt; k \leq \frac{1}{2})</td>
<td>(\hat{Y}_1)</td>
<td>(g/2 &lt; k \leq \frac{1}{2})</td>
<td>(\hat{Y}_2)</td>
</tr>
<tr>
<td>(\frac{1}{2} &lt; k \leq 1 - f/2)</td>
<td>(\hat{Y}_2)</td>
<td>(\frac{1}{2} &lt; k \leq 1 - g/2)</td>
<td>(\hat{Y}_1)</td>
</tr>
<tr>
<td>(1 - f/2 &lt; k)</td>
<td>(\hat{Y}_r)</td>
<td>(1 - g/2 &lt; k)</td>
<td>(\hat{Y}_r)</td>
</tr>
</tbody>
</table>

Table 7.2: Guidelines in the negative correlation case

<table>
<thead>
<tr>
<th>Interval for (k' = -V_{11}/V_{02})</th>
<th>Preferred estimator</th>
<th>Interval for (k')</th>
<th>Preferred estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0 \leq k' \leq f/2)</td>
<td>(\hat{Y})</td>
<td>(0 \leq k' \leq g/2)</td>
<td>(\hat{Y})</td>
</tr>
<tr>
<td>(f/2 &lt; k' \leq \frac{1}{2})</td>
<td>(\hat{Y}_3)</td>
<td>(g/2 &lt; k' \leq \frac{1}{2})</td>
<td>(\hat{Y}_4)</td>
</tr>
<tr>
<td>(\frac{1}{2} &lt; k' \leq 1 - f/2)</td>
<td>(\hat{Y}_4)</td>
<td>(\frac{1}{2} &lt; k' \leq 1 - g/2)</td>
<td>(\hat{Y}_3)</td>
</tr>
<tr>
<td>(1 - f/2 &lt; k')</td>
<td>(\hat{Y}_p)</td>
<td>(1 - g/2 &lt; k')</td>
<td>(\hat{Y}_p)</td>
</tr>
</tbody>
</table>
If we call the values of \( k(k') \) in the four different intervals specified in columns 1 and 3 of Tables 7.1 and 7.2 to be very low, moderately low, moderately high and very high respectively, a quick summary of the results is as follows. When the sampling fraction is less than or equal to \( 1/2 \) and the value of \( k(k') \) is (i) very low it is preferable to use the simple estimator \( \hat{Y}_1 \), (ii) moderately low it is better to use \( \hat{Y}_1(\hat{Y}_3) \), (iii) moderately high it is efficient to use \( \hat{Y}_2(\hat{Y}_4) \) and (iv) very high it is good to use \( \hat{Y}_4(\hat{Y}_3) \) as an estimator of the population total \( Y \). In the contrary case of \( f > 1/2 \) the roles of \( \hat{Y}_1(\hat{Y}_3) \) and \( \hat{Y}_2(\hat{Y}_4) \) are interchanged.

### 7.2.1 Use of multiauxiliary information

One fairly general approach is the following. Assume that the auxiliary information is available on \( p \) variates \( x_1, \ldots, x_p \). Let \( \rho_{yt} \) be the coefficient of correlation between \( y \) and \( x_t \), \( C_t \) be the coefficient of variation of \( x_t \) in the population and \( k_t = \rho_{yt} C_t / C_t \). Suppose we can make a good guess of the sign and magnitude of \( k_t \) for each \( t \). Assume \( f \leq 1/2 \). We partition the set \( S \) of all the \( x \)-variates into seven subsets \( S_t, t = 0,1,\ldots,6 \) as follows: \( S_0 \) consists of all \( x_t \) for which
\[
-1/2 \leq k_t \leq 1/2 ;
\]
\[
S_1 : f/2 \leq k_t \leq 1/2 ; \quad S_2 : 1/2 \leq k_t \leq 1 - f/2 ;
\]
\[
S_3 : 1 - f/2 \leq k_t ; \quad S_4 : -1/2 \leq k_t \leq -f/2 ;
\]
\[
S_5 : f/2 - 1 \leq k_t < 1/2 ; \quad S_6 : k_t < f/2 - 1 .
\]
Denote \( S_1 \cup S_2 \cup S_4 \cup S_5 \) by \( A \). Thus \( A \) is the subset of all \( x \)-variate values having moderately low or moderately high positive or negative values of \( k_t \). In fact \( A : f/2 < |k_t| < (1-f/2) \). Let \( x_t, x_t^{(2)}, x_t^{\hat{}} \) denote respectively the population total, total for the non-sample units and the simple expansion estimator of the population total for the variate \( x_t, t = 1, \ldots, p \).

We may consider the following estimator of \( Y \) based on a without replacement simple random sample of size \( n \) from the population.

\[
\hat{Y}^* = \hat{Y} \sum_{S_0} w_t + \hat{Y}_1 \sum_{S_1} w_t (1 + x_t^{(2)}/x_t^{\hat{}}) \\
+ \hat{Y}_2 \sum_{S_2} w_t (1 + g x_t^{(1)}/x_t^{\hat{}}) + \hat{Y}_3 \sum_{S_3} w_t x_t^{\hat{}}/x_t^{\hat{}} \\
+ \hat{Y}_4 \sum_{S_4} w_t (1 + g^2 x_t^{(2)}/x_t^{\hat{}}) \\
+ \hat{Y}_5 \sum_{S_5} w_t (1 + g x_t^{(1)}/x_t^{\hat{}}) + \hat{Y}_6 \sum_{S_6} w_t x_t^{\hat{}}/x_t^{\hat{}} \tag{7.17}
\]

where \( w_1, \ldots, w_p \), \( \sum_{S_0} w_t = 1 \), is a weight function and \( \sum_{S_0} \) indicates summation over all \( x \)-variate values belonging to \( S_0 \) and so on. For computation purposes \( \hat{Y}^* \) can be recast as

\[
\hat{Y}^* = \hat{Y} \sum_{S_0} w_t + \hat{Y}_1 \sum_{S_1} w_t x_t^{(2)}/x_t^{\hat{}} + \hat{Y}_2 \sum_{S_2} w_t x_t^{\hat{}}/x_t^{\hat{}} \\
+ \sum_{S_3} w_t x_t^{\hat{}}/x_t^{\hat{}} + g^2 \sum_{S_4} w_t x_t^{(2)}/x_t^{\hat{}} + g \sum_{S_5} w_t x_t^{\hat{}}/x_t^{\hat{}} + \sum_{S_6} w_t x_t^{\hat{}}/x_t^{\hat{}}. \tag{7.18}
\]
The approximate bias and mse of the estimators and the best weight function have been given in Srivenkataramana and Tracy (1979).

7.2.2 Mixing estimators

Following the above work, Vos (1980) has considered mixing of direct, ratio and product method estimators. He introduces two classes of estimators:

\[ \hat{Y}_{M1} = W \hat{Y} + (1-w) \hat{Y}_p \quad (7.19) \]

\[ \hat{Y}_{M2} = W \hat{Y} + (1-w) \hat{Y}_r \quad (7.20) \]

It is shown that, for a suitable choice of the weights, each of these estimators has smaller variance than the direct as well as the product or ratio method estimators.

More generally one may consider pooled estimators of the type

\[ \hat{Y}_w = w(t) T_1 + (1 - w(t)) T_2 \quad (7.21) \]

where \( T_1, T_2 \) are any two estimators of \( Y \) and \( w(t) \) is a function of the statistic \( t \) used to test a relevant hypothesis; e.g. \( H_0 : B = B_0 \) against the alternative \( H_1 : B \neq B_0 \) where \( B \) is the coefficient of regression of \( y \) on \( x \) in the population. Grimes and Sukhatme (1980) have examined the special case restricting the choice of \( w(t) \) to functions of the type

\[ w(t) = 1 \quad \text{if} \quad |t| \leq t_0 \]

\[ = 0 \quad \text{if} \quad |t| > t_0 \]

and with difference and regression estimators as \( T_1, T_2 \) respectively.
They call the resulting estimator $\hat{Y}_d$ the sometimes regression estimator.

The more general cases are to be investigated.

7.3 **Orthogonal Auxiliary Variates**

The case of two auxiliary variates $x_1, x_2$ and a difference method of estimation are assumed for illustration. Raj (1965a) has suggested an estimator of the form

$$\hat{Y}_d = \hat{Y} - \omega_1 k_1 (\hat{x}_1 - \bar{x}_1) - \omega_2 k_2 (\hat{x}_2 - \bar{x}_2)$$  \hspace{1cm} (7.22)

which has, under srsor, variance given by

$$V(\hat{Y}_d) = \frac{N(N-n)}{n} \left\{ \frac{\omega_1^2 k_1^2 S_{11}}{\omega_2^2 k_2^2 S_{22}} + \omega_2^2 k_2^2 S_{22} - 2 \omega_1 k_1 S_{12} \right\}$$  \hspace{1cm} (7.23)

where $S_{00}$ is the population variance of $y$; etc. It is clear that a positive $S_{12}$ inflates the sampling variance (7.23). Therefore transformations which diffuse the correlation between the auxiliary variates will be helpful. Thus we may consider replacing the set $(x_1, x_2)$ of auxiliary variates by variates which are orthogonal or nearly so. For example, replace $(x_1, x_2)$ by $(x_1, x_2 - B_{21} x_1)$. The implication of this idea for the case of two or more auxiliary variates and practicable ways of effecting the transformation are to be studied.
PART TWO : PRACTICE

A SURVEY OF RURAL TRANSITION IN KARNATAKA, INDIA.

1980 – 81
CHAPTER - 8

BACKGROUND AND SURVEY DESIGN

8.1 Introduction

The rural life in India has been undergoing a visible transition to modern ways during the recent years. In the post-independence era (1947 - ), particularly of late, a number of levelling forces have been operative on the Indian village scene. For example, villages are now uniformly administered by statutory Panchayats (Village councils) elected through universal adult franchise. Schools, hospitals, post offices and cooperative societies are found at least in the vicinity of almost every village. Transport and communication have been greatly accelerated through roads, railways, post and telegraph, newspapers and radio. Illiteracy is being steadily reduced. Statute law has replaced customary law. Absentee landlordism is almost extinct. The land tenure of tenants has become secure. There is electricity supply to almost all villages in many of the states. A few case studies which examine some of these aspects have been made once in a while. However there is scope and need for conducting more socio-economic studies of the villages. The present survey was carried out in the Karnataka State in South India, with the objective of identifying the major factors causing the village transition and assessing the magnitude of change in the rural parts. Household amenities, farming practices, individual life-style and outlook, and transition of the village as a unit are studied. Since no comparable earlier study exists for Karnataka, it is hoped that this will serve as a benchmark.
8.2 The Land and the People

The Karnataka State is situated in the south-western part of India. The total area is 191,791 sq km. According to the 1981 census the population is 37 million, showing a 26.4% increase over the 1971 figure of 29.3 million. In terms of area and population Karnataka is the eighth among the states and union territories of India. An idea of the largeness of the population may be obtained by noting that it is greater than that of Canada (23.7 million).

The Karnataka State came into existence on November 1, 1973 under the States' Reorganization Act (1971). Initially it was known as Mysore State. The name was changed to Karnataka State on November 1, 1973 to signify the fact that the state comprises the Kannada land of olden times which was spread over different administrative units before the reorganization. Accordingly we note five regions in the state: (i) Old-Mysore Region; (ii) Kodagu Region; (iii) Madras-Karnataka Region; (iv) Bombay-Karnataka Region and (v) Hyderabad-Karnataka Region. Kannada is the official language of the state, but English continues to be used for official purposes and as medium of instruction in higher education. Several other Indian languages are spoken by segments of the population. For administrative purposes Karnataka has been divided into 19 districts. Table 8.1 gives some related details.

About 76% of the people in Karnataka live in rural areas spread over 29,553 villages of which 26,826 are inhabited. Thus Karnataka is a land of villages. The rural economy is mainly based on agriculture and allied activities like animal husbandry and fishing. About 55% of the villages
Table - 8.1: Districts, their area, population and its density in Karnataka†

<table>
<thead>
<tr>
<th>Region</th>
<th>District</th>
<th>Area (sq km)</th>
<th>1981 Population (million)</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Old-Mysore</td>
<td>Bangalore</td>
<td>8005</td>
<td>4.9</td>
<td>615</td>
</tr>
<tr>
<td></td>
<td>Chikkamagalur</td>
<td>7201</td>
<td>0.9</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>Chitradurga</td>
<td>10852</td>
<td>1.8</td>
<td>164</td>
</tr>
<tr>
<td></td>
<td>Hassan</td>
<td>6814</td>
<td>1.4</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>Kolar</td>
<td>8223</td>
<td>1.9</td>
<td>231</td>
</tr>
<tr>
<td></td>
<td>Mandya</td>
<td>4961</td>
<td>1.4</td>
<td>285</td>
</tr>
<tr>
<td></td>
<td>Mysore</td>
<td>11954</td>
<td>2.6</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>Shimoga</td>
<td>10553</td>
<td>1.7</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>Tumkur</td>
<td>10598</td>
<td>2.0</td>
<td>186</td>
</tr>
<tr>
<td>II. Kodaū</td>
<td>Kodaū</td>
<td>4102</td>
<td>0.5</td>
<td>112</td>
</tr>
<tr>
<td>III. Madras-Karnataka</td>
<td>Dakshina Kannada</td>
<td>8441</td>
<td>2.4</td>
<td>281</td>
</tr>
<tr>
<td>IV. Bombay-Karnataka</td>
<td>Belgaum</td>
<td>13415</td>
<td>3.0</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>Dharwad</td>
<td>13738</td>
<td>2.9</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td>Uttara Kannada</td>
<td>10291</td>
<td>1.1</td>
<td>104</td>
</tr>
<tr>
<td>V. Hyderabad-Karnataka</td>
<td>Bellary††</td>
<td>9885</td>
<td>1.5</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>Bidar</td>
<td>5448</td>
<td>1.0</td>
<td>182</td>
</tr>
<tr>
<td></td>
<td>Bijapur</td>
<td>17069</td>
<td>2.4</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>Gulbarga</td>
<td>16224</td>
<td>2.1</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Raichur</td>
<td>14017</td>
<td>1.8</td>
<td>127</td>
</tr>
</tbody>
</table>

† Source: Census of India, 1981, Series 9 Karnataka, Provisional population totals.

†† Though Bellary was in Madras State before 1953, it is included in region V because of its proximity to the other districts in this region and distance from the district in region III.

are small, each having less than 500 inhabitants, 39% are medium sized (having 500 to 2000 inhabitants) while the others are large.
8.3 Sample Selection

The units of interest in the survey are individuals, households, farms and villages. A restricted random selection ensuring an uniform physical coverage of the state was considered desirable, since differences exist between one region and another. Therefore, it was decided to include several rural points from each district, in the sample. Accordingly the following sample design was adopted.

In each district the larger towns were identified and grouped into five sets: one comprising all those at the centre of the district and the other four in the four directions from the centre - east, south, west and north. From each set one town was selected at random as the reference point. This ensured that the five points were well spread out in the district (see Map 8.1). A pair of investigators was assigned to each district. The pair was familiar with the district allotted to it and had a good knowledge of the local language. Around each reference point the investigators were to visit at least two villages separated by a minimum of 10 km. Houses of all types - huts, small houses and large houses were included. In order to assess individual transition, two persons were selected at random from each selected household, excluding children aged below 10 years. In all, information from 184 villages, 431 households, 379 farms, and 823 individuals was recorded. Initially, the more objective method of listing the villages near a reference point and selecting two from them and then listing the households in the selected villages for picking the sample households was tried. But this was seen to be too time consuming since ready lists were rarely available and
Map 8.1: Reference Points in the Districts for Sample Selection.
was abandoned in favour of the above, though less objective, alternative.

The data were collected during December 1980 and January 1981. During these months of the year the field conditions are generally favourable, in the sense of being free from floods or other extreme climatic conditions. The investigators filled up a specially designed questionnaire (Refer Appendix), which was pretested.

8.4 The Questionnaire

This had six sections - basic information, house, farm, outlook, individual and village transitions. The majority of the questions was of multiple choice type. The questionnaire was in English. A translation was provided in the provincial language, Kannada, to facilitate understanding by a larger section of the people.

The information on household and farm was obtained by setting up a dialogue with that member who was able to respond most clearly. He (she) was allowed to be assisted by the others, since the information was on the household or farm and not on the individual. Direct observation by the interviewer was used for factors like housetypes and condition of surroundings. A household was defined as a group of people living together and eating from the same kitchen.

Information on outlook and individual transitions was obtained from the individuals selected. Information on village transition was obtained by talking to person(s) well informed about the village, like the village Panachayat President or Accountant and also through direct observation,
CHAPTER 9

FINDINGS OF THE SURVEY

9.1 Quality of Data

The investigators in the survey were the staff and students of the Department of Statistics of Bangalore University. To begin with, they were familiarized with the objectives of the project. A small scale pilot study was also conducted. In the main survey the investigators actually visited the rural points in the sample and obtained quite a bit of information through direct observation. Thus the nonsampling errors are expected to be minimal.

The Indian villagers generally mistake the investigators to be revenue officials of the government, collecting information for tax purposes and hence hesitate to talk to them. In order to overcome this problem the field workers carried with them an identification card and a statement that the project was for an academic purpose. On reaching the village they contacted the Panchayat President or an important local person and acquainted him of their objective. He usually sent word around the village, instructing the villagers to cooperate. The investigators knew the language spoken by the local populace. All this helped to win the confidence of the respondents. This reduced the nonresponse and improved the quality of the data. When the completed questionnaires arrived in Bangalore, a specially trained staff scrutinized them according to instructions prepared beforehand and coded the relevant entries. The data were then transferred to punched cards and tabulated.
9.2 Results of the Survey

The data from the survey are used to examine: (i) house types; (ii) household conveniences; (iii) household change; (iv) outlook transition; (v) individual transition; (vi) farm transition; and (vii) village transition in the rural parts of Karnataka. Since the responses were mainly categorical, the results are mostly expressed as percentages. In each case the sample size and the method used to obtain the data are stated.

9.2.1 House types

The statements are based on the data obtained from the 431 houses in the sample, through direct observation by the investigators. There are three main types of dwelling units that one sees in the villages of Karnataka - independent houses (58%), huts (26%) and part of a building (16%). Generally the walls are constructed with clay puddle (56%) or locally made brick and stone (40%). Sometimes we come across the use of bamboo poles or wooden planks (4%) for this purpose. The use of flat or pan tiles for the roof is most common (54%). Straw, grass, Palmyrah or coconut palm is used instead in the huts. The use of reed and mud (16%), asbestos or galvanized steel sheets (3%) is also occasionally seen. The flooring is made of mud plaster (60%) or brick and stone. The percentage of houses having own wells as source of water is about 25, while the majority (71%) depend on community wells. About 60% of the houses have their own angala (open-air courtyard).

About 56% of the houses have 3 or fewer rooms for an average of 8 members. Therefore per force the same room is to be used for a number of different purposes. Ventilation is often poor (36%) in the case of the older structures.
9.2.2 Household conveniences

The conclusions here are partly based on direct observation by the interviewer and partly on the information obtained by getting-up a dialogue with a member of the household who was able to respond most clearly. All the 431 households in the sample are covered. It is noted that a large number (62%) of the households have a separate space as kitchen. Aluminum (45%) and earthen (41%) cooking vessels are the most common. In the upper class families copper and brass vessels were common earlier, now making room for steelware. In the poorer sections, the change is from earthenware to aluminum vessels. Kerosene lamps continue to be the main source of light (95%), but there is a progressive switch-over to electric lamps. This is aided by the fact that almost all the villages in the state are supplied with electricity. More than 50% of the houses have some furniture.

One of the most significant changes is the trend with respect to medical care. People have changed over from native medicines to government hospitals (69%) or private medical practitioners (27%). Improved roads and bus transport to urban centres have provided access to the hospitals. However, medical aid is generally sought only in serious cases. The situation of veterinary care is similar. The habit of seeing movies for recreation is noted in the villages near urban centres. Tours are undertaken generally as yatra (pilgrimage) rather than for sightseeing.

9.2.3 Household change

The respondents were asked to what extent their family had changed in respect to education, dress pattern, household amenities, medical aid,
traditions and customs and the status of women during the past decade. The percentage distribution of the responses from the 431 families is shown in Table 9.1.

The percentage of families reporting moderate or significant change with respect to the different factors is quite high. An exception is the case of traditions and customs where half the number of households reported no change. The trends in respect to education and medical aid are particularly noteworthy.

Table 9.1 Percentage distribution according to extent of change in the families (n = 431).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Extent of change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nil</td>
</tr>
<tr>
<td>Education</td>
<td>24</td>
</tr>
<tr>
<td>Dress pattern</td>
<td>28</td>
</tr>
<tr>
<td>Household amenities</td>
<td>40</td>
</tr>
<tr>
<td>Medical aid</td>
<td>23</td>
</tr>
<tr>
<td>Traditions &amp; customs</td>
<td>50</td>
</tr>
<tr>
<td>Status of women</td>
<td>39</td>
</tr>
</tbody>
</table>

9.2.4 Outlook transition

A typical villager has religious and caste beliefs and has at least a fair awareness of the village-level political situation (77%), though less so at the provincial (53%) and national (44%) levels. The outlook on certain other factors is summarized in Table 9.2.
A high percentage of people in favour of women's education, equal rights for women, family planning and a high percentage against the dowry system in marriage is noted.

Table 9.2 Percentage distribution by outlook on certain factors in Karnataka (n = 431).

<table>
<thead>
<tr>
<th>Factor</th>
<th>For</th>
<th>Against</th>
<th>Indifferent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Religious &amp; caste beliefs</td>
<td>58</td>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td>Untouchability</td>
<td>25</td>
<td>58</td>
<td>17</td>
</tr>
<tr>
<td>Dowry system</td>
<td>15</td>
<td>67</td>
<td>18</td>
</tr>
<tr>
<td>Women's education</td>
<td>77</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>Equal rights for women</td>
<td>71</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Family planning</td>
<td>70</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Superstitions</td>
<td>32</td>
<td>39</td>
<td>29</td>
</tr>
</tbody>
</table>

9.2.5 Individual transition

Two individuals were selected at random from each selected household for studying this aspect. Children below 10 were excluded. There were 470 male and 353 female respondents. Their distribution by education level is in Table 9.3.

A typical individual in the village is traditionally dressed (56% for males, 69% for females) or shows a change towards modern type (36% for males, 24% for females) of dress. Men traditionally wear dhoties, while women wear saris. The practices as regards the use of certain items in daily life are summarized by the percentages in Table 9.4.
Table 9.3 Distribution by education level and sex of respondent

<table>
<thead>
<tr>
<th>Education level</th>
<th>Percentage</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>Nil</td>
<td>28</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Up to grade 5</td>
<td>19</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Grades 6 through 10</td>
<td>37</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>16</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Table 9.4 Use of certain items in daily life by individuals

<table>
<thead>
<tr>
<th>Item</th>
<th>Male (%)</th>
<th>Female (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Footwear</td>
<td>42</td>
<td>58</td>
</tr>
<tr>
<td>Cosmetics</td>
<td>78</td>
<td>22</td>
</tr>
<tr>
<td>Toilet soap</td>
<td>35</td>
<td>65</td>
</tr>
<tr>
<td>Toothpaste/powder</td>
<td>76</td>
<td>24</td>
</tr>
<tr>
<td>Hair oil</td>
<td>57</td>
<td>43</td>
</tr>
<tr>
<td>Wrist watch</td>
<td>54</td>
<td>46</td>
</tr>
<tr>
<td>Safety razor</td>
<td>53</td>
<td>47</td>
</tr>
<tr>
<td>Vanity bag</td>
<td>N.A.</td>
<td>82</td>
</tr>
</tbody>
</table>

It is interesting to note that 42% of the men and 61% of the women do not use any footwear. Also only 65% of the men and 59% of the women use toilet soap. The 66% of women using no cosmetics or the 82% using no vanity bag are indicative of a traditional type of rural women folk.
9.2.6 Farm transition

A typical (57%) farm is a small landholding (less than 5 acres), growing food crops only (53%), like paddy and wheat, and an additional 41% growing both food and cash crops (like cotton or sugarcane). The mode of ploughing land is mainly by animal-drawn wooden or metal plough (96%). The use of tractors is just 4%. The farmer depends heavily (66%) on rain for irrigation; 15% have water pumps and another 15% have water supply through canal systems.

The practice of using certain important factors in agriculture is indicated by the percentages in Table 9.5.

Table 9.5 Percentages showing the use of certain factors in farming in Karnataka (n = 379).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Extent of use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never</td>
</tr>
<tr>
<td>Compost</td>
<td>26</td>
</tr>
<tr>
<td>Chemical fertilizers</td>
<td>24</td>
</tr>
<tr>
<td>Chemical insecticides</td>
<td>32</td>
</tr>
<tr>
<td>Hybrid &amp; high yielding varieties</td>
<td>45</td>
</tr>
<tr>
<td>Banking facility</td>
<td>46</td>
</tr>
<tr>
<td>Farm cooperative</td>
<td>45</td>
</tr>
</tbody>
</table>

These percentages are almost equally distributed among the three columns. Exceptions are the case of compost, where usage is frequent, and, to some extent, the case of banking facility, where the situation is not very satisfactory. The mode of storing farm produce is quite poor (39%) and is just satisfactory in another 57% of the cases. Again,
the mode of marketing the produce is unsatisfactory in about half
the number of cases. Only a moderate change in farming practices
during the last 10 years was reported in 57% of the cases, with
another 23% reporting no change at all.

9.2.7 Village transition

A village in Karnataka has, on the average, about 200 households
with a population of about 750. The average distances of certain
facilities are given in Table 9.6. By and large these distances are
moderate, considering the importance of the facility for a village,
except for railway stations and banks.

<table>
<thead>
<tr>
<th>Facility</th>
<th>Average distance (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Railway station</td>
<td>33.4</td>
</tr>
<tr>
<td>Bus station</td>
<td>2.8</td>
</tr>
<tr>
<td>Tar road</td>
<td>1.9</td>
</tr>
<tr>
<td>Post office</td>
<td>1.4</td>
</tr>
<tr>
<td>Elementary school</td>
<td>0.5</td>
</tr>
<tr>
<td>High school</td>
<td>4.1</td>
</tr>
<tr>
<td>College</td>
<td>12.4</td>
</tr>
<tr>
<td>Government hospital</td>
<td>5.3</td>
</tr>
<tr>
<td>Doctor's shop</td>
<td>4.0</td>
</tr>
<tr>
<td>Fair price shop</td>
<td>2.7</td>
</tr>
<tr>
<td>Bank</td>
<td>5.9</td>
</tr>
<tr>
<td>Cinema</td>
<td>7.9</td>
</tr>
</tbody>
</table>
The condition of primary and secondary schools is found to be
good or tolerable in 83% of the villages (Table 9.7). The government-
sponsored adult education programme does not seem to work satisfactorily,
with more than 60% of the villages having very little activity of this
type. The primitive practices of barter, bonded labour and untouchability
are on their way out (Table 9.8).

Table 9.7  Condition of certain facilities for villages
in Karnataka (Percentage of villages, n = 184).

<table>
<thead>
<tr>
<th>Facility</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>School</td>
<td>33</td>
</tr>
<tr>
<td>College</td>
<td>13</td>
</tr>
<tr>
<td>Adult education</td>
<td>10</td>
</tr>
<tr>
<td>Medical aid</td>
<td>16</td>
</tr>
<tr>
<td>Bus transport</td>
<td>40</td>
</tr>
<tr>
<td>Electricity</td>
<td>46</td>
</tr>
</tbody>
</table>

Table 9.8  Prevalence of certain practices in villages
in Karnataka (Percentage of villages, n = 184).

<table>
<thead>
<tr>
<th>Practice</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nil</td>
</tr>
<tr>
<td>Barter system</td>
<td>48</td>
</tr>
<tr>
<td>Bonded labour</td>
<td>60</td>
</tr>
<tr>
<td>Untouchability</td>
<td>41</td>
</tr>
</tbody>
</table>
About 50% of the villages exhibit an overall moderate change in village life-style during the last decade, while 33% showed a substantial change (Table 9.9).

Table 9.9  Extent of change in village life during 1971-81 in Karnataka (n = 184)

<table>
<thead>
<tr>
<th>Extent</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nil</td>
<td>8</td>
</tr>
<tr>
<td>Moderate</td>
<td>59</td>
</tr>
<tr>
<td>Great</td>
<td>33</td>
</tr>
</tbody>
</table>

9.3 Factors Causing Transition

An attempt was made to order the following factors according to their influences on village life-style:

(a) Schooling

(b) Roads and means of communication

(c) Rural electricity

(d) Government programmes like community development

(e) Newspapers and radio.

The village was considered as a unit and persons well informed about the village were asked to state the impact of the above factors on rural transition in general. Their responses are summarized by the percentages in Table 9.10.

Roads and means of communication have contributed the most towards village transition, schooling comes next and then we have newspapers and radio, government programmes and rural electricity. The observed influence of these is briefly outlined next.
Table 9.10 Impact of certain factors on rural transition
(Percentage of responses, n = 184).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nil</td>
</tr>
<tr>
<td>Schooling</td>
<td>13</td>
</tr>
<tr>
<td>Roads &amp; means of</td>
<td>18</td>
</tr>
<tr>
<td>communication</td>
<td></td>
</tr>
<tr>
<td>Rural electricity</td>
<td>32</td>
</tr>
<tr>
<td>Government programmes</td>
<td>30</td>
</tr>
<tr>
<td>Newspapers &amp; radio</td>
<td>18</td>
</tr>
</tbody>
</table>

Roads passing through or nearby a village have a visible impact on the life around. This can be easily seen by comparing the life-style in a village adjacent to a road with that in a village about 10 km away from the nearest road.

In the former, a bunch of shops on or near the road where people engage themselves in petty trades is common. Bullock carts and bicycles ply around transporting people and goods. This wider accessibility sets up a network and makes well-mixed economic activity possible.

Children make use of the road for reaching the school, walking along it or taking the bus that may ply a couple of times during the day. This helps them beat the rivulets putting up hurdles during the rainy months. Farmers and worker groups also find similar utility from the roads. This has greatly improved the income from farm and dairy produce and also wage rates.

Easy access to the cities allows the village folk to observe frequently and directly the life-style in urban areas. This has resulted in such
style being copied in terms of dress, food habits and practices.
Visits to cinema and coffee houses change the attitudes of rustic people.

With an activity spot like the road-side bus stand or intersection
of roads a reversal in the mode of availability of certain services and
facilities is seen. The village barber, for example, used to go from
doors to door to make his service available. These days he prefers to
set up a petty shop and wait for the customers to walk in. Similar
observations can be made regarding the seller of bangles, earthen pots
or fish.

On the other hand, in a village far removed from roads, the
activities are geographically very much restricted. The hills, hillocks,
rivers and rivulets delimit the area of activity of a person, setting
up natural barriers to mobility. This is especially true of rivers during
the rainy months.

The ability to read and write greatly increases the sources of
information for a person, and improves his capacity to examine a
situation critically. The quality of his life improves and he develops
his own values in life. The general picture in Karnataka is one of
pronounced increase in literacy rate over the years. It was 25.4% in
1961, 31.5% in 1971 and now it is 38.4%. This trend has been maintained
by both males and females. However there is enormous scope for improvement
in female literacy. Among men, literacy improved from 41.6% in 1971 to
48.6% in 1981, while the corresponding percentages for women were only
21.0 and 27.8. The increase in literacy is the result of the expansion of
primary and secondary education in the state, free education up to
secondary level and the cumulative effect of investments made during the
past decades.

Newspapers and radio supply the people with current information constantly. Special programs and features meant for the villagers are included. These media assist in the quick and continuous spread of relevant information as a routine matter.

Special governmental programs, like community development, food for work and adult education, have helped to fill the gaps in rural development. They have also helped people to realize the benefit of cooperatives and other collective efforts.

Almost all villages in Karnataka are supplied with electricity. This has improved irrigation and has replaced some manual work by machine work. Electricity is also being used progressively for lighting.

9.4 Some Limitations

The present study was a multipurpose one and covered a large geographical area. Because of an inadequate network of roads and poor transport facilities, certain rural pockets of land were inaccessible. Further the standard sampling techniques and methods of data collection used in the developed countries cannot be applied in the rural Indian situation. Here the lists on which sampling frames can be based are inadequate. The majority of the people are illiterate and have a suspicion of outsiders, the women folk particularly shying away from strangers. The present study was planned with an awareness of these handicaps. Accordingly the survey design was adapted to suit the field conditions. Also since care was taken to deputize well trained interviewers, knowing the local language and other peculiarities, the quality of the data was reasonable. The results obtained are in general agreement with other comparable results.
like that from the census. The study sheds light on the ways of
life in a society about which investigations of this type have been
rare. It also serves as a benchmark and has implications on the larger
rural Indian context.

9.5 Follow-up Studies

A study of transition essentially tries to measure the extent of
change that has taken place during a period of time. Therefore it is
necessary to have comparable studies at the two points of time concerned.
In the present case, no comparable earlier study exists. Thus it may
be interesting to use the present study as a benchmark and conduct
a similar study after some time, say five years, and assess the
transition that occurs meanwhile. On a smaller scale, such studies
may be conducted in greater depth by considering individual districts.
Case studies of a few villages can also be taken up.

The Karnataka State was under the British before 1947. The
neighbouring union territory of Goa on the west coast was under the
Portuguese while, on the east coast, Pondicherry was under the French.
A comparison of rural transition in Karnataka with that in Goa and
Pondicherry will be worthwhile.
QUESTIONNAIRE
BANGALORE UNIVERSITY
BANGALORE

DEPARTMENT OF STATISTICS

QUESTIONNAIRE

A SURVEY ON RURAL TRANSITION
IN KARNATAKA, INDIA

1980-81

(WITH ASSISTANCE FROM IDRC, CANADA)
A Survey on Transformation of Rural People to Modern ways of Life in Karnataka, India.

<table>
<thead>
<tr>
<th>Serial No</th>
<th>Stratum No</th>
<th>Interviewers</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>i)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ii)</td>
<td></td>
</tr>
</tbody>
</table>

1. **BASIC INFORMATION**

1. Village
2. Taluk
3. District
4. Household size
5. Joint family? Yes/No

6. **Particulars of the members:**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Age</th>
<th>Relation to the Head</th>
<th>Level of education</th>
<th>Currently studying Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>i)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iv)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vi)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>viii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ix)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xi)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xiii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xiv)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xv)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7. Average monthly household income Rs.

8. Average monthly household expenditure Rs.

9. Two main sources of income i) ii)

10. Number of couples in the household
1. **House**
   - i) Own
   - ii) not own

2. **Plinth area (Sq. Ft)**

3. **House type**
   - i) hut
   - ii) part of a building
   - iii) Independent house
   - iv) flat

4. **Roof type**
   - i) straw, grass, palmyrah or coconut palm
   - ii) reed and mud
   - iii) flat or pan tiles
   - iv) asbestos sheets
   - v) galvanized steel sheets
   - vi) R. C. C.

5. **Wall type**
   - i) bamboo poles etc.
   - ii) clay puddle
   - iii) wooden plank
   - iv) brick or stone

6. **Floor type**
   - i) mud
   - ii) brick, stone
   - iii) cement

7. **Condition of structure**
   - i) bad
   - ii) moderately good
   - iii) good

8. **Period since built (years)**

9. **Period since last major repair undertaken (years)**

10. **Condition of surroundings**
    - i) bad
    - ii) moderately good
    - iii) good

11. **Number of living rooms**

12. **Number of other rooms**

13. **Ventilation**
    - i) bad
    - ii) tolerable
    - iii) good

14. **Kitchen**
    - i) not separate
    - ii) separate

15. **Bathing place**
    - i) no specific arrangement
    - ii) sharing with others
    - iii) own
    - iv) near water pond, river etc.
16. Type of lavatory
   i) no arrangement
   ii) services
      iii) septic tank
      iv) drainage

17. Source of water for house
   i) community well (or bore well) etc.
   ii) own well
   iii) water supply

18. Main cooking vessels
   i) earthen
   ii) aluminium
   iii) brass, copper or bronze
   iv) stainless steel

19. Main cooking fuel
   i) fire wood, dung cake
   ii) charcoal, saw dust
   iii) kerosene
   iv) gas
   v) electricity

20. Main lighting fuel
   i) kerosene
   ii) electricity
   iii) others

21. Household furniture (number)
   i) chairs
   ii) cots
   iii) benches
   iv) tables
   v) almirahs
   vi) stools

22. Vehicles
   i) cart
   ii) bicycle
   iii) scooter (moped)
   iv) car (van)

23. Other articles
   i) clock (watch)
   ii) radio
   iii) pen (pencil)

24. Main transport mode within village limits
   i) walking
   ii) cart
   iii) bicycle
   iv) public transport
   v) rickshaw
   vi) scooter (moped)
   vii) car

25. Medical care
   i) native medicine
   ii) govt. hospital
   iii) private doctor.
26. Distance to the nearest medical centre in kms

27. Do you take medical help only in serious cases? yes/no

28. Veterinary care
   i) native medicine
   ii) govt. hospital
   iii) private doctor

29. Distance to the nearest veterinary centre in kms

30. Do you take veterinary help only in serious cases?

Subscription to periodicals

31. dailies
   a) English Yes/No
   b) Others Yes/No

32. magazines
   a) English Yes/No
   b) Others Yes/No

33. Does the family regularly listen to radio news? Yes/No

34. Does the family go to movies regularly? Yes/No

35. Does the family go on picnics or tours frequently? Yes/No

36. Does the family like a city type of living? Yes/No

To what extent did your household change with respect to the following, during the last years?

<table>
<thead>
<tr>
<th>Item</th>
<th>Extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>37. Education</td>
<td>i) nil</td>
</tr>
<tr>
<td></td>
<td>ii) Moderate</td>
</tr>
<tr>
<td></td>
<td>iii) significant</td>
</tr>
<tr>
<td>38. Dress pattern</td>
<td>i)</td>
</tr>
<tr>
<td></td>
<td>ii)</td>
</tr>
<tr>
<td></td>
<td>iii)</td>
</tr>
<tr>
<td>39. Household amenities</td>
<td>i)</td>
</tr>
<tr>
<td></td>
<td>ii)</td>
</tr>
<tr>
<td></td>
<td>iii)</td>
</tr>
<tr>
<td>40. Medical aid</td>
<td>i)</td>
</tr>
<tr>
<td></td>
<td>ii)</td>
</tr>
<tr>
<td></td>
<td>iii)</td>
</tr>
<tr>
<td>41. Traditions and customs</td>
<td>i)</td>
</tr>
<tr>
<td></td>
<td>ii)</td>
</tr>
<tr>
<td></td>
<td>iii)</td>
</tr>
<tr>
<td>42. Status of women</td>
<td>i)</td>
</tr>
<tr>
<td></td>
<td>ii)</td>
</tr>
<tr>
<td></td>
<td>iii)</td>
</tr>
</tbody>
</table>
III. FARM TRANSITION

1. Livestock (number)
   i) cows and oxen
   ii) buffaloes
   iii) sheep and goat
   iv) poultry

2. Area under cultivation
   i) Nil
   ii) ________________________ acres

3. Land
   i) own
   ii) not own

4. Two main items grown
   i) ________________________
   ii) ________________________

5. Mode of ploughing land
   i) animal drawn wooden plough
   ii) animal drawn metal plough
   iii) tractor

6. Main mode of irrigation
   i) rain and natural water
   ii) lift system
   iii) tube well
   iv) water pump
   v) canal

Use of the following

<table>
<thead>
<tr>
<th>Compost</th>
<th>never</th>
<th>occasional</th>
<th>frequently</th>
<th>If (i) specify reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8. Chemical fertilizers
   i)    
   ii)   
   iii)  

9. Chemical insecticides
   i)    
   ii)   
   iii)  

10. Hybrid and high yielding varieties
    i)    
    ii)   
    iii)  

11. Banking facility
    i)    
    ii)   
    iii)  

12. Farm co-operative
    i)    
    ii)   
    iii)  
13. Mode of storing the produce,
   i) unsatisfactory    ii) satisfactory    iii) good

14. Mode of marketing the produce
   i) unsatisfactory    ii) satisfactory    iii) good

15. To what extent your farming practices have changed during the last ten years?
   i) nil    ii) moderate    iii) significant

(Ssections IV and V for selected individuals)

IV. OUTLOOK TRANSITION

Political awareness

1. Local
   i) good    ii) fair    iii) poor

2. Provincial
   i)    ii)    iii)

3. National
   i)    ii)    iii)

4. Religious and caste beliefs
   i) for    ii) against    iii) indifferent

5. Untouchability
   i)    ii)    iii)

6. Dowry system
   i)    ii)    iii)

7. Women's education
   i)    ii)    iii)

8. Equal rights for women
   i)    ii)    iii)

9. Family planning
   i)    ii)    iii)

10. Superstitions
    i) strong    ii) moderate    iii) nil
V. INDIVIDUAL TRANSITION

<table>
<thead>
<tr>
<th>Item</th>
<th>Respondent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1. Name</td>
<td></td>
</tr>
<tr>
<td>2. Sex</td>
<td>M/F</td>
</tr>
<tr>
<td>3. Age (years)</td>
<td></td>
</tr>
<tr>
<td>4. Level of education</td>
<td></td>
</tr>
<tr>
<td>5. Type of dress</td>
<td>i) traditional</td>
</tr>
<tr>
<td></td>
<td>ii) modern</td>
</tr>
<tr>
<td></td>
<td>iii) urbanized</td>
</tr>
<tr>
<td></td>
<td>iv) Ultramodern</td>
</tr>
<tr>
<td>Use of the following</td>
<td></td>
</tr>
<tr>
<td>6. Safety razor (for men only)</td>
<td>yes/no</td>
</tr>
<tr>
<td>7. Footwear</td>
<td></td>
</tr>
<tr>
<td></td>
<td>i) nil</td>
</tr>
<tr>
<td></td>
<td>ii) chappals</td>
</tr>
<tr>
<td></td>
<td>iii) shoes</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Cosmetics</td>
<td>yes/no</td>
</tr>
<tr>
<td>9. Toilett soap</td>
<td>yes/no</td>
</tr>
<tr>
<td>10. Tooth paste/powder</td>
<td>yes/no</td>
</tr>
<tr>
<td>11. Hair oil</td>
<td>yes/no</td>
</tr>
<tr>
<td>12. Wrist watch</td>
<td>yes/no</td>
</tr>
<tr>
<td>13. Vanity bag (for women only)</td>
<td>yes/no</td>
</tr>
</tbody>
</table>
VI. VILLAGE TRANSITION

1. District ....................................
2. Taluk ....................................
3. Village ...................................
4. No. of Household in the village ..........
5. Population of the village .............

*Distance from the nearest following places*

<table>
<thead>
<tr>
<th>Place</th>
<th>Distance (k.ms.)</th>
<th>Place</th>
<th>Distance (k.ms.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Railway station</td>
<td>...</td>
<td>Veterinary hospital</td>
<td>...</td>
</tr>
<tr>
<td>Bus station</td>
<td>...</td>
<td>Fertilizer depot</td>
<td>...</td>
</tr>
</tbody>
</table>
| Tar road               | ...              | Repair place for pump sets etc, | ...
| Post office            | ...              | Flour mill             | ...              |
| Elementary school      | ...              | Fair price shop (ration shop) | ...
| High school            | ...              | Shopping centre        | ...              |
| College                | ...              | Bank                   | ...              |
| Govt. hospital         | ...              | Cinema                 | ...              |
| Doctor's shop          | ...              | Library & reading room | ...              |
| Primary health centre  | ...              |                        |                  |

*The condition of the following facilities for the village*

25. School
   i) good
   ii) tolerable
   iii) bad

26. College
   i) ...
   ii) ...
   iii) ...

27. Adult education
   i) ...
   ii) ...
   iii) ...
REFERENCES


VITA AUCTORIS

1945
Born on the 18th of May at Mangalore, India.

1959
Matriculated from Anandashrama High School,
Dakshina Kannada, India.

1963
Graduated with a B.Sc. degree from Government
College, Mangalore, affiliated to University of
Mysore, India.

1965
Received M.Sc. degree in Statistics from Karnataka
University, Dharwar, India.
Appointed Lecturer in Statistics, Department of
Collegiate Education, Karnataka, India.

1969
Appointed Assistant Professor in Statistics,
Bangalore University, India.

1975
Won Bangalore University Research Award in Statistics.

1976
Admitted to Graduate Programme at the University of
Windsor, Canada.

1978
Won the Pierre Robillard Award of the Statistical
Society of Canada.

1979
Received M.Sc. degree in Mathematics from the University
of Windsor, Canada. Advanced to Ph.D. candidacy.
Won a Ph.D. Thesis Research Award of the International
Development Research Centre of Canada.