Wiley Series in Survey Methodology

Sampling Statistics

Wayne A. Fuller

WILEY

SAMPLING STATISTICS

WAYNE A. FULLER
Iowa State University

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

This Page Intentionally Left Blank

SAMPLING STATISTICS

WILEY SERIES IN SURVEY METHODOLOGY

Established in Part by WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *Robert M. Groves, Graham Kalton, J. N. K. Rao, Norbert Schwarz, Christopher Skinner*

A complete list of the titles in this series appears at the end of this volume.

SAMPLING STATISTICS

WAYNE A. FULLER
Iowa State University

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Fuller, Wayne A.
Sampling statistics / Wayne A. Fuller.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-45460-2 (cloth)
1. Sampling (Statistics) 2. Estimation theory. 3. Mathematical statistics. I. Title.
QA276.6.F84 2009
519.5'2—dc22 2009008874

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

This book developed out of a desire for a sampling course that would fit easily into a graduate program in statistics. Survey sampling is a relatively young discipline, achieving acceptance in the 1940's and 1950's, primarily for official statistics. As the discipline has matured, analytic use of survey data has increased inside and outside government. Also statistical models, such as those for nonresponse and for small area estimation, are now considered part of survey methodology. As a result, the overlap between survey sampling and other areas of statistics has increased, and the mutual dependence makes it important that survey sampling be an integral part of statistics.

Originally, survey sampling was differentiated from other areas by the size of the data sets and by the number of estimates produced. In such a setting survey statisticians prefer techniques with broad applicability, and that require a minimum of assumptions. Procedures are sought that are nearly design unbiased, but no claim of optimality is made for a particular statistic. These standard survey techniques are introduced in Chapter One. I have adopted a tenor and notation similar to statistics texts in other specialities to make the material more accessible to those with limited exposure to survey sampling. Some of the technical material in Section 1.3 can be omitted or covered at a later point. Basic sampling concepts are introduced in a way to facilitate application of model based procedures to survey samples. Likewise, models are used in constructing estimators and in discussions of designs in Chapter Two and Chapter Three, respectively. Chapter Five is devoted to procedures, such as nonresponse adjustment and small area estimation, where models play a central role. To be comfortable with the material the reader should have completed courses in the theory of statistics and in linear regression.

Survey data are now regularly used for the estimation of a parameter θ of a subject matter model. Such estimation is discussed in Chapter Six. The problem will be familiar to most statisticians, but complex survey designs complicate the analysis.

Our primary experience has been with survey samples of populations such as the residents of the state of Iowa or the land area of the United States. Thus our unconscious frame of reference will be such surveys. Likewise

the majority of our experience has been in the production of "general use" databases rather than analytic study of a limited number of characteristics.

I am indebted to a number of former students, coworkers, and friends for corrections and improvements. F. Jay Breidt made sizeable contributions to Chapter One. Emily Berg, Yu Wu, Nicholas Beyler, and Pushpal Mukhopadhyay assisted in computing examples. Chris Skinner, Jae-Kwang Kim, and Mike Hidiroglou made suggestions that led to numerous improvements. Jean Opsomer used portions of the manuscript in his class and provided valuable feed back. I am particularly grateful to Jason Legg who used the manuscript in a class, corrected a number of errors and contributed to Chapter Three and Chapter Five.

Appreciation is expressed for the support provided by the Center for Survey Statistics and Methodology and the Department of Statistics, Iowa State University. I thank Glenda Ashley and Sherri Martinez for typing the manuscript and Ozkan Zengin for editing and for technical support on LaTeX.

Wayne A. Fuller

Ames, Iowa
January 2009

List of Tables

List of Principal Results

List of Examples

CHAPTER 1

PROBABILITY SAMPLING FROM A FINITE UNIVERSE

1.1 INTRODUCTION

A large fraction of the quantitative information that we receive about our economy and our community comes from sample surveys. Statistical agencies of national governments regularly report estimates for items such as unemployment, poverty rates, crop production, retail sales, and median family income. Some statistics may come from censuses, but the majority are based on a sample of the relevant population. Less visible statistics are collected by other entities for business decisions, city planning, and political campaigns. National polls on items beyond politics are regularly reported in newspapers. These reports are so common that few reflect on the fact that almost all people believe that something interesting and (or) useful can be said about a nation of 300 million people on the basis of a sample of a few thousand. In fact, the concept that a probability sample can be so used has only been accepted by the scientific community for about 60 years. In this book we study the statistical basis for obtaining information from samples.

Sampling Statistics. By Wayne A. Fuller

In this chapter we develop a probabilistic framework for the study of samples selected from a finite population. Because the study of estimators often requires the use of large-sample approximations, we define sequences of populations and samples appropriate for such study.

1.2 PROBABILITY SAMPLING

Consider a finite set of elements identified by the integers $U = \{1, 2, \ldots, N\}$. The set of identifiers, sometimes called *labels*, can be thought of as forming a list. The existence of such a list, a list in which every element is associated with one and only one element of the list, is the cornerstone of probability sampling. The list is also called the *sampling frame*. In practice, the frame takes many forms. For example, it may be a list in the traditional sense, such as the list of employees of a firm or the list of patients in a hospital. It is sometimes the set of subareas that exhaust the geographic area of a political unit such as a city or state.

Associated with the jth element of the frame is a vector of characteristics denoted by $\mathbf{y}_j$. In all of our applications, the $\mathbf{y}_j$ are assumed to be real valued. The entire set of N vectors is denoted by $\mathcal{F}$. The set is called a *finite population* or a *finite universe*. A sample is a subset of the elements. Let A denote the set of indices from U that are in the sample. In statistical sampling the interest is in the selection of samples using probability rules such that the probability characteristics of the set of samples defined by the selection rules can be established. Let $\mathcal{A}$ denote the set of possible samples under a particular probability procedure. A person who wishes to obtain information about a population on the basis of a sample must develop a procedure for selecting the sample.

The terms *random samples* and *probability samples* are both used for samples selected by probability rules. Some people associate the term *random sampling* with the procedure in which every sample has the same probability and every element in the population has the same probability of appearing in the sample.

1.2.1 Basic properties of probability samples

In this section we present some basic properties of statistics constructed from probability samples. In the methods of this section, the probabilistic properties depend only on the sampling procedure. The population from which the samples are selected is fixed. Let A be a subset of U and let $\mathcal{A}$ be the collection of subsets of U that contains all possible samples. Let $P[A = a]$ denote the probability that a, $a \in \mathcal{A}$, is selected.

Definition 1.2.1. A *sampling design* is a function $p(\cdot)$ that maps a to $[0, 1]$ such that $p(a) = P[A = a]$ for any $a \in \mathcal{A}$.

A set of samples of primary importance is the set of all possible samples containing a fixed number of distinct units. Denote the fixed size by n. Then the number of such samples is

$$\binom{N}{n} = \frac{N!}{(N-n)!n!}, \tag{1.2.1}$$

where $N! = 1 \times 2 \times \cdots \times N$.

A probability sampling scheme for samples of fixed size n assigns a probability to each possible sample. *Simple random nonreplacement sampling* assigns equal probability to each possible sample. We may occasionally refer to such samples as simple random samples. The *inclusion probability* for element i is the sum of the sample probabilities for all samples that contain element i; that is,

$$\pi_i = P(i \in A) = \sum_{a \in A_{(i)}} p(a),$$

where $A_{(i)}$ is the set of samples that contain element i.

The terms *selection probability*, *probability of selection*, and *observation probability* are also used. In simple random nonreplacement sampling, element i appears in

$$\binom{1}{1}\binom{N-1}{n-1} \tag{1.2.2}$$

samples. If every sample has equal probability, the probability of selecting element i is

$$\pi_i = \left[\binom{N}{n}\right]^{-1}\binom{N-1}{n-1} = \frac{n}{N}. \tag{1.2.3}$$

In discussing probability sampling schemes, we define indicator variables to identify those elements appearing in the sample. Let I_i be the indicator variable for element i. Then

$$\begin{aligned} I_i &= 1 \quad \text{if element } i \text{ is in the sample} \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{1.2.4}$$

Let $\mathbf{d} = (I_1, I_2, \ldots, I_N)$ be the vector of random variables. The probabilistic behavior of functions of the sample depends on the probability distribution of $\mathbf{d}$. The sampling design specifies the probability structure of $\mathbf{d}$, where the inclusion probability for element i is the expectation of I_i,

$$\pi_i = E\{I_i\}. \tag{1.2.5}$$

With this notation, the sum of characteristic y for the elements in the sample is

$$\text{sample sum} \quad = \quad \sum_{i=1}^{N} I_i y_i. \tag{1.2.6}$$

The set A is the set of indices *appearing* in the sample. Thus,

$$A \quad = \quad \{i \in U : I_i = 1\}. \tag{1.2.7}$$

Then the sample sum of (1.2.6) can be written

$$\sum_{i=1}^{N} I_i y_i = \sum_{i \in A} y_i. \tag{1.2.8}$$

The joint inclusion probability, denoted by π_{ik}, for elements i and k is the sum of sample probabilities for all samples that contain both elements i and k. In terms of the indicator variables, the joint inclusion probability for elements i and k is

$$\pi_{ik} \quad = \quad E\{I_i I_k\}. \tag{1.2.9}$$

For simple random nonreplacement sampling, the number of samples that contain elements i and k is

$$\binom{1}{1}\binom{1}{1}\binom{N-2}{n-2} \tag{1.2.10}$$

and

$$\pi_{ik} \quad = \quad [N(N-1)]^{-1} n(n-1). \tag{1.2.11}$$

The number of units in a particular sample is

$$n \quad = \quad \sum_{i=1}^{N} I_i, \tag{1.2.12}$$

and because each I_i is a random variable with expected value π_i, the expected sample size is

$$E\{n\} \quad = \quad \sum_{i=1}^{N} E\{I_i\} \quad = \quad \sum_{i=1}^{N} \pi_i. \tag{1.2.13}$$

Also, the variance of the sample size is

$$V\{n\} = V\left\{\sum_{i=1}^{N} I_i\right\} = \sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_{ik} - \pi_i\pi_k),$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{N}\pi_{ik} - \left(\sum_{i=1}^{N}\pi_i\right)^2, \quad (1.2.14)$$

where $\pi_{ii} = \pi_i$. If $V\{n\} = 0$, we say that the design is a *fixed sample size* or *fixed-size design.* It follows from (1.2.14) that

$$\sum_{\substack{i=1 \\ i\neq k}}^{N}\sum_{k=1}^{N}\pi_{ik} = n^2 - n = n(n-1) \quad (1.2.15)$$

for fixed-size designs. Also, for fixed-size designs,

$$\sum_{k\in U: i\neq k}\pi_{ik} = \sum_{k=1}^{N} E\{I_iI_k\} - \pi_i = (n-1)\pi_i. \quad (1.2.16)$$

Discussions of estimation for finite population sampling begin most easily with estimation of linear functions such as finite population totals. This is because it is possible to construct estimators of totals for a wide range of designs that are unbiased conditionally on the particular finite population. Such estimators are said to be *design unbiased.*

Definition 1.2.2. A statistic $\hat{\theta}$ is *design unbiased* for the finite population parameter $\theta_N = \theta(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$ if

$$E\{\hat{\theta} \mid \mathcal{F}\} = \theta_N$$

for any vector $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$, where $E\{\hat{\theta} \mid \mathcal{F}\}$, the design expectation, denotes the average over all samples possible under the design for the finite population $\mathcal{F}$.

Probability sampling became widely accepted in the 1940s. For a number of years thereafter, sampling statisticians who considered estimation problems approached design and estimation problems by treating the N unknown values of the finite population as fixed values. All probability statements were with respect to the distribution created by the sample design probabilities. Thus, in many discussions in the sampling literature the statement that an estimator is "unbiased" means design unbiased for a parameter of the finite universe.

The concept of a linear estimator is also very important in estimation theory. Often, modifiers are required to fully define the construct. If an estimator $\hat{\theta}$ can be written as

$$\hat{\theta} = \sum_{i \in A} w_i y_i, \tag{1.2.17}$$

where the w_i are not functions of the sample y's, we say that the estimator $\hat{\theta}$ is *linear in* y. In the statistical theory of linear models, estimators of the form (1.2.17) are called *linear estimators* provided that the w_i are fixed with respect to the random mechanism generating the y values. Thus, the model specification for the random process and the set of samples under consideration define the statistical linearity property. We will have use for the concept of linearity relative to the design.

Definition 1.2.3. An estimator is *design linear* if it can be written in the form (1.2.17) or, equivalently, as

$$\hat{\theta} = \sum_{i \in U} I_i w_i y_i,$$

where the w_i are fixed with respect to the sampling design.

Observe that for a given finite population, the vector $(w_1 y_1, w_2 y_2, \ldots, w_N y_N)$ is a fixed vector and the elements of the vector are the coefficients of the random variables I_i.

The design mean and design variance of design linear estimators are functions of the selection probabilities. In Definition 1.2.2 we introduced the concept of the expectation over all possible samples for a particular finite population $\mathcal{F}$. We use $V\{\hat{\theta} \mid \mathcal{F}\}$ to denote the analogous design variance.

Theorem 1.2.1. Let $(y_1, y_2, \ldots, y_N)$ be the vector of values for a finite universe of real-valued elements. Let a probability sampling procedure be defined, where π_i denotes the probability that element i is included in the sample and π_{ik} denotes the probability that elements i and k are in the sample. Let

$$\hat{\theta} = \sum_{i \in A} w_i y_i = \sum_{i \in U} I_i w_i y_i$$

be a design linear estimator. Then

$$E\{\hat{\theta} \mid \mathcal{F}\} = \sum_{i=1}^{N} w_i \pi_i y_i \tag{1.2.18}$$

and

$$V\{\hat{\theta} \mid \mathcal{F}\} = \sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_{ik} - \pi_i\pi_k)w_iy_iw_ky_k, \tag{1.2.19}$$

where $\pi_{ik} = \pi_i$ if $i = k$.

If $V(n) = 0$, then $V\{\hat{\theta} \mid \mathcal{F}\}$ can be expressed as

$$V\{\hat{\theta} \mid \mathcal{F}\} = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_i\pi_k - \pi_{ik})(w_iy_i - w_ky_k)^2. \tag{1.2.20}$$

Proof. Because $E\{I_i\} = \pi_i$ and because w_iy_i, $i = 1, 2, \ldots, N$, are fixed, we have

$$E\{\hat{\theta} \mid \mathcal{F}\} = \sum_{i=1}^{N} E\{I_i \mid \mathcal{F}\}w_iy_i = \sum_{i=1}^{N}\pi_iw_iy_i$$

and (1.2.18) is proven. In a similar manner, and using $E\{I_iI_k\} = \pi_{ik}$, we have

$$\begin{aligned} V\{\hat{\theta} \mid \mathcal{F}\} &= V\left\{\sum_{i=1}^{N} I_iw_iy_j \mid \mathcal{F}\right\} \\ &= E\left\{\left(\sum_{i=1}^{N} I_iw_iy_i\right)^2 \middle| \mathcal{F}\right\} - \left(\sum_{i=1}^{N}\pi_iw_iy_i\right)^2 \\ &= \sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_{ik} - \pi_i\pi_k)w_iy_iw_ky_k \end{aligned}$$

and (1.2.19) is proven. Also see Exercise 1.

To prove (1.2.20) for fixed-size designs, expand the square in (1.2.20) to obtain

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_i\pi_k - \pi_{ik})(w_i^2y_i^2 - 2w_iy_iw_ky_k + w_k^2y_k^2) \\ &\qquad = \sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_i\pi_k - \pi_{ik})w_i^2y_i^2 \\ &\qquad\quad - \sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_i\pi_k - \pi_{ik})w_iy_iw_ky_k. \end{aligned}$$

The result follows because $\Sigma_{k=1}^{N}(\pi_{ik} - \pi_i\pi_k) = 0$ for fixed-size designs. See (1.2.14). ■

We have stated Theorem 1.2.1 for scalars, but the results extend immediately to vectors. If $\mathbf{y}_i$ is a column vector and

$$\hat{\boldsymbol{\theta}} = \sum_{i \in A} w_i \mathbf{y}_i,$$

the covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$V\{\hat{\boldsymbol{\theta}} \mid \mathcal{F}\} = \sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_{ik} - \pi_i\pi_k)w_i\mathbf{y}_i w_k\mathbf{y}_k'.$$

Two finite population parameters of particular interest are the finite population total,

$$T_y = \sum_{i \in U} y_i = \sum_{i=1}^{N} y_i, \tag{1.2.21}$$

and the finite population mean,

$$\bar{y}_N = N^{-1}T_y. \tag{1.2.22}$$

If $\pi_i > 0$ for all i, the design linear estimator of the total,

$$\hat{T}_y = \sum_{i \in A} \pi_i^{-1} y_i, \tag{1.2.23}$$

is design unbiased. The estimator (1.2.23) is known as the *Horvitz–Thompson estimator* and is sometimes called the π *estimator*. See Horvitz and Thompson (1952) and Narain (1951). The corresponding design-unbiased estimator of the mean is

$$\bar{y}_{HT} = N^{-1}\hat{T}_y \tag{1.2.24}$$

The properties of the Horvitz–Thompson estimator follow from Theorem 1.2.1.

Corollary 1.2.1.1. Let the conditions of Theorem 1.2.1 hold, let $\pi_i > 0$ for all i, and let the design linear estimator of T_y be $\hat{T}_y$ of (1.2.23). Then

$$E\{\hat{T}_y \mid \mathcal{F}\} = T_y \tag{1.2.25}$$

and

$$V\{\hat{T}_y - T_y \mid \mathcal{F}\} = \sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_{ik} - \pi_i\pi_k)\pi_i^{-1}y_i\pi_k^{-1}y_k. \quad (1.2.26)$$

If $V\{n\} = 0$, then $V\{(\hat{T}_y - T_y) \mid \mathcal{F}\}$ can be expressed as

$$\sum_{i=1}^{N}\sum_{k=1}^{N}(\pi_{ik} - \pi_i\pi_k)(\pi_i^{-1}y_i - n^{-1}T_y)(\pi_k^{-1}y_k - n^{-1}T_y) \quad (1.2.27)$$

or as

$$\frac{1}{2}\sum_{\substack{i=1\\ i\neq k}}^{N}\sum_{k=1}^{N}(\pi_i\pi_k - \pi_{ik})(\pi_i^{-1}y_i - \pi_k^{-1}y_k)^2. \quad (1.2.28)$$

Proof. Results (1.2.25), (1.2.26), and (1.2.28) follow from (1.2.18), (1.2.19), and (1.2.20), respectively, by substituting $w_i = \pi_i^{-1}$.

To show that (1.2.27) is equal to (1.2.26) for fixed-size designs, observe that

$$\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_i\pi_k(\pi_i^{-1}y_i - n^{-1}T_y)(\pi_k^{-1}y_k - n^{-1}T_y)$$

$$= \left(\sum_{i=1}^{N}(y_i - n^{-1}\pi_i T_y)\right)^2 = 0.$$

From (1.2.16), $\Sigma_{k=1}^{N}\pi_{ik} = n\pi_i$. Thus,

$$\sum_{i=1}^{N}(\pi_i^{-1}y_i - n^{-1}T_y)\sum_{k=1}^{N}\pi_{ik}n^{-1}T_y = T_y\sum_{i=1}^{N}(y_i - n^{-1}\pi_i T_y) = 0$$

and (1.2.27) is equal to

$$\sum_{i=1}^{N}\sum_{k=1}^{N}\pi_{ik}\pi_i^{-1}y_i\pi_k^{-1}y_k - T_y^2.$$

■

The Horvitz–Thompson estimator is an unbiased estimator of the total, but it has some undesirable features. The estimator is scale invariant but not

location invariant. That is, for real α, β not zero,

$$\sum_{i \in A} \pi_i^{-1} \beta y_i = \beta \sum_{i \in A} \pi_i^{-1} y_i,$$

but

$$\sum_{i \in A} \pi_i^{-1}(y_i + \alpha) = \sum_{i \in A} \pi_i^{-1} y_i + \alpha \sum_{i \in A} \pi_i^{-1}. \tag{1.2.29}$$

The second term of (1.2.29) is $N\alpha$ for many designs, including equal-probability fixed-sample-size designs. However, $\sum_{i \in A} \pi_i^{-1}$ is, in general, a nondegenerate random variable.

The lack of location invariance restricts the number of practical situations in which the Horvitz–Thompson estimator and unequal probability designs are used. One important use of unequal probability sampling is the situation in which the π_i are proportional to a measure of the number of observation units associated with the sampling unit.

Example 1.2.1. Assume that one is interested in the characteristics of households in Des Moines, Iowa. A recent listing of the city blocks and the number of dwelling units in each block is available. On the presumption that the number of households is strongly correlated with the number of dwelling units, we might select a sample of blocks with probability proportional to the number of dwelling units. Assume that all households in the block are observed. In this situation, the fact that the Horvitz–Thompson estimator is not location invariant is relatively unimportant because we are interested in the properties of households, not in the properties of linear functions of blocks. It was in a context such as this that unequal probability sampling was first suggested. See Hansen and Hurwitz (1943). ■■

The fact that the Horvitz–Thompson estimator is not location invariant has another consequence. Associated with each sampling unit is the characteristic, which is always 1. The population total for this characteristic is the number of sampling units in the population. The Horvitz–Thompson estimator of the population size is the coefficient of α in (1.2.29),

$$\hat{N} = \hat{T}_1 = \sum_{i \in A} \pi_i^{-1} \tag{1.2.30}$$

with variance

$$V\{\hat{T}_1 \mid \mathcal{F}\} = \sum_{i=1}^{N} \sum_{k=1}^{N} (\pi_{ik} - \pi_i \pi_k) \pi_i^{-1} \pi_k^{-1}. \tag{1.2.31}$$

Although there are situations in which N is unknown, in many situations N is known. Therefore, the fact that the estimator of the population size is not equal to the true size suggests the possibility of improving the Horvitz–Thompson estimator. We pursue this issue in Section 1.3 and Chapter 2.

Under the conditions that $\pi_i > 0$ for all i and $\pi_{ik} > 0$ for all i and k, it is possible to construct a design-unbiased estimator of the variance of a design linear estimator. Designs with the properties $\pi_i > 0$ for all i and $\pi_{ik} > 0$ for all ik are sometimes said to be *measurable*.

Theorem 1.2.2. Let the conditions of Theorem 1.2.1 hold with $\pi_{ik} > 0$ for all $i,\ k,\ \in U$. Let $\hat{\theta}$ be a design linear estimator of the form (1.2.17). Then

$$\hat{V}\{\hat{\theta} \mid \mathcal{F}\} = \sum_{i,k\in A}\sum \pi_{ik}^{-1}\left(\pi_{ik} - \pi_i\pi_k\right) w_i y_i w_k y_k \qquad (1.2.32)$$

is a design-unbiased estimator of $V\{\hat{\theta} \mid \mathcal{F}\}$. If $V\{n\} = 0$,

$$\tilde{V}\{\hat{\theta} \mid \mathcal{F}\} = \frac{1}{2}\sum_{i,k\in A}\sum \pi_{ik}^{-1}(\pi_i\pi_k - \pi_{ik})(w_i y_i - w_k y_k)^2 \qquad (1.2.33)$$

is a design-unbiased estimator of $V\{\hat{\theta} \mid \mathcal{F}\}$.

Proof. Let $g(y_i,\ y_k)$ be any real-valued function of $(y_i,\ y_k)$. Because $\pi_{ik} > 0$ for all $(i,\ k)$, it follows by direct analogy to (1.2.18) that

$$E\left\{\sum_{i,k\in A} \pi_{ik}^{-1} g(y_i,\ y_k) \mid \mathcal{F}\right\} = \sum_{i=1}^{N}\sum_{k=1}^{N} g(y_i,\ y_k). \qquad (1.2.34)$$

Result (1.2.32) is obtained from (1.2.34) and (1.2.19) by setting

$$g(y_i,\ y_k) = (\pi_{ik} - \pi_i\pi_k) w_i y_i w_k y_k.$$

Result (1.2.33) follows from (1.2.34) and (1.2.20) by setting

$$g(y_i,\ y_k) = (\pi_{ik} - \pi_i\pi_k)\ (w_i y_i - w_k y_k)^2.$$

■

The estimator (1.2.32) for estimator (1.2.19) is due to Horvitz and Thompson (1952), and the estimator (1.2.33) was suggested by Yates and Grundy (1953) and Sen (1953) for estimator (1.2.20).

Theoretically, it is possible to obtain the variance of the estimated variance. The squared differences in (1.2.33) are a sample of all possible differences.

If we consider differences $(w_i y_i - w_k y_k)^2$, for $i \neq k$ there is a population of $N(N-1)$ differences. The probability of selecting any particular difference is π_{ik}. The variance of the estimated difference is a function of the π_{ik} and of the probability that any pair of pairs occurs in the sample. Clearly, this computation can be cumbersome for general designs. See Exercise 12.

Although design unbiased, the estimators of variance in Theorem 1.2.2 have the unpleasant property that they can be negative. If at least two values of $\pi_i^{-1} y_i$ differ in the sample, the variance must be positive and any other value for an estimator is unreasonable.

The Horvitz–Thompson variance estimator also has the undesirable property that it can give a positive estimate for an estimator known to have zero variance. For example, if y_i is proportional to π_i, the variance of $\hat{T}_y$ is zero for fixed-size designs, but estimator (1.2.32) can be nonzero for some designs.

Theorem 1.2.2 makes it clear that there are some designs for which design-unbiased variance estimation is impossible because unbiased variance estimation requires that $\pi_{ik} > 0$ for all (i, k). A sufficient condition for a design to yield nonnegative estimators of variance is $\pi_{ik} < \pi_i \pi_k$. See (1.2.33).

For simple random nonreplacement sampling, $\pi_i = N^{-1} n$ and

$$\pi_{ik} = [N(N-1)]^{-1} n(n-1) \text{ for } i \neq k.$$

Then the estimated total (1.2.23) is

$$\hat{T}_y = N n^{-1} \sum_{i \in A} y_i = N \bar{y}_n, \tag{1.2.35}$$

where

$$\bar{y}_n = n^{-1} \sum_{i \in A} y_i.$$

Similarly, the variance (1.2.26) reduces to

$$\begin{aligned} V\{(\hat{T}_y - T_y) \mid \mathcal{F}\} &= N(N-n) n^{-1} S_{y,N}^2, \\ &= N^2 (1 - f_N) n^{-1} S_{y,N}^2, \end{aligned} \tag{1.2.36}$$

where

$$S_{y,N}^2 = (N-1)^{-1} \sum_{j=1}^{N} (y_j - \bar{y}_N)^2$$

and $f_N = N^{-1} n$. The quantity $S_{y,N}^2$, also written without the N subscript and with different subscripts, is called the *finite population variance*. A few texts define the finite population variance with a divisor of N and change the definition of $V\{\hat{T}_y \mid \mathcal{F}\}$ appropriately. The term $(1 - f_N)$ is called the *finite*

population correction (fpc) or *finite correction term.* It is common practice to ignore the term if the sampling rate is less than 5%. See Cochran (1977, p. 24).

The estimated variance (1.2.33) reduces to

$$\hat{V}\{\hat{T}_y \mid \mathcal{F}\} = Nn^{-1}(N-n)s_{y,n}^2 \tag{1.2.37}$$

for simple random sampling, where

$$s_{y,n}^2 = (n-1)^{-1}\sum_{j\in A}(y_j - \bar{y}_n)^2.$$

The quantity $s_{y,n}^2$ is sometimes called the *sample variance* and may be written with different subscripts. The results for simple random sampling are summarized in Corollary 1.2.2.1.

Corollary 1.2.2.1. Let $U = \{1, 2, \ldots, N\}$ and let $\mathcal{F} = (y_1, y_2, \ldots, y_N)$ be the values of a finite universe. Let a simple random sample of size n be selected from $\mathcal{F}$, let $\bar{y}_n$ be defined by (1.2.35), and let $s_{y,n}^2$ be as defined for (1.2.37). Then

$$\begin{aligned} E\{\bar{y}_n \mid \mathcal{F}\} &= \bar{y}_N, \\ V\{\bar{y}_n \mid \mathcal{F}\} &= N^{-1}(N-n)n^{-1}S_{y,N}^2, \end{aligned} \tag{1.2.38}$$

and

$$E\{\hat{V}(\bar{y}_n \mid \mathcal{F}) \mid \mathcal{F}\} = V\{\bar{y}_n \mid \mathcal{F}\}, \tag{1.2.39}$$

where $\bar{y}_N$ is defined in (1.2.22), $S_{y,N}^2$ is defined in (1.2.36), and

$$\hat{V}\{\bar{y}_n \mid \mathcal{F}\} = (1-f_N)n^{-1}s_{y,n}^2.$$

Proof. For simple random nonreplacement sampling, $\pi_i = N^{-1}n$ and $\pi_{ik} = [N(N-1)]^{-1}n(n-1)$ for $i \neq k$. Thus, by (1.2.25) of Corollary 1.2.1.1,

$$E\{\hat{T}_y \mid \mathcal{F}\} = E\{N\bar{y}_n\} = N\bar{y}_N$$

and we have the first result. The result (1.2.38) is obtained by inserting the probabilities into (1.2.26) to obtain

$$\begin{aligned} V\{\hat{T}_y - T_y \mid \mathcal{F}\} &= \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} n^{-1}(N-n)(N-1)^{-1}(y_i - y_k)^2 \\ &= Nn^{-1}(N-n)(N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2 \\ &= N^2n^{-1}(1-f_N)S_{y,N}^2. \end{aligned}$$

By the same algebraic argument, the estimator (1.2.33) for the estimated total is

$$\hat{V}\{\hat{T}_y \mid \mathcal{F}\} = Nn^{-1}(N-n)s_{y,n}^2$$

with expectation $Nn^{-1}(N-n)S_{y,N}^2$ by Theorem 1.2.2. ■

In some situations, such as the investigation of alternative designs, it is useful to consider the finite population to be generated by a stochastic mechanism. For example, the $\{y_i\}$, $i = 1, 2, \ldots, N$, might be independent identically distributed (iid) random variables with a distribution function $F(y)$. We then say that the finite population is a sample from the *superpopulation* $F(y)$.

A simple and useful specification is that of Theorem 1.2.3. The combination of a sample design and an estimator is called a *strategy*.

Theorem 1.2.3. Let $\{y_1, y_2, \ldots, y_N\}$ be a set of $iid(\mu, \sigma^2)$ random variables. Let the sample design have probabilities π_i, $\pi_i > 0$, and π_{ik} such that $\Sigma_{i=1}^N \pi_i = n$. Assume that the vector $\mathbf{d}$ of selection indicators is independent of $\{y_1, y_2, \ldots, y_N\}$. Let the estimated total be

$$\hat{T}_y = \sum_{i \in A} \pi_i^{-1} y_i. \tag{1.2.40}$$

Then $\pi_i = N^{-1}n$ and $\pi_{ik} = [N(N-1)]^{-1}n(n-1)$ for $i \neq k$ minimize the variance of $\hat{T}_y - T_y$.

Proof. Under the assumptions, $E\{\hat{T}_y - T_y \mid \mathcal{F}\} = 0$ and d is independent of the y_i. Therefore, the unconditional variance of $\hat{T}_y - T_y$ is the expectation of the conditional variance. Using $E\{y_i^2\} = \sigma^2 + \mu^2$ and $E\{y_i y_k\} = \mu^2$ for $i \neq k$,

$$\begin{aligned} V\{\hat{T}_y - T_y\} &= E\left\{\sum_{i=1}^N \sum_{k=1}^N C\{I_i, I_k\} \pi_i^{-1} y_i \pi_k^{-1} y_k\right\} \\ &= \mu^2 \sum_{i=1}^N \sum_{k=1}^N (\pi_{ik} - \pi_i \pi_k)\pi_i^{-1}\pi_k^{-1} \\ &\quad + \sigma^2 \sum_{i=1}^N (\pi_i - \pi_i^2)\pi_i^{-2}, \end{aligned} \tag{1.2.41}$$

where $C\{x, z\}$ is the covariance of x and z.

Given that $\Sigma_{i=1}^N \pi_i = n$, the second term of (1.2.41) is minimized if all π_i are equal because π_i^{-1} is convex in π_i. The first term of (1.2.41) is nonnegative

because it equals $\mu^2 V\{\Sigma_{i=1}^{N} I_i \pi_i^{-1}\}$. Therefore, the minimum possible value for the first term is zero, which is attained if the π_i are equal to $N^{-1}n$ and $\pi_{ik} = [N(N-1)]^{-1}n(n-1)$ for all $i \neq k$. ■

The formulation of Theorem 1.2.3 deserves further discussion. The result pertains to a property of an estimator of the finite population total. However, the property is an average over all possible finite populations. The result does not say that simple random sampling is the best procedure for the Horvitz–Thompson estimator for a particular finite population. We cannot find a design-unbiased procedure that is minimum variance for all fixed unknown finite populations because the design variance is a function of the N unknown values. See Godambe (1955), Godambe and Joshi (1965), and Basu (1971). On the other hand, if our information about the finite population is such that we are willing to act as if the finite population is a set of *iid* random variables, simple random sampling is the best sampling strategy for the Horvitz–Thompson estimator, where "best" is with respect to the superpopulation specification. If the finite population is assumed to be a sample from a normal distribution, the sample mean is optimal for the finite population mean.

Theorem 1.2.4. Let $\{y_1, y_2, \ldots, y_N\}$ be a set of normal independent random variables with mean μ and variance σ^2, denoted by $NI(\mu, \sigma^2)$ random variables. In the class of sample selection procedures for samples of size n that are independent of $\{y_1, y_2, \ldots, y_N\}$, the procedure of selecting a simple random nonreplacement sample of size n from U and using the estimator $\bar{y}_n$ to estimate the finite population mean, $\bar{y}_N$, is an optimal strategy in that there is no strategy with smaller mean square error.

Proof. By Theorem 1.3.1, the n elements in the sample are $NI(\mu, \sigma^2)$ random variables. Therefore, the sample mean is the minimum mean square error estimator of μ. See, for example, Stuart and Ord (1991, p. 617). Furthermore, the minimum mean square error predictor of $\bar{y}_{N-n}$ is $\bar{y}_n$. Thus,

$$\bar{y}_n = N^{-1}\left[n\bar{y}_n + (N-n)\bar{y}_n\right]$$

is the best predictor of $\bar{y}_N$. ■

Because the elements of the original population are identically distributed, any nonreplacement sampling scheme that is independent of $\{y_1, y_2, \ldots, y_N\}$ would lead to the same estimation scheme and the same mean square error. However, probability sampling and the Horvitz–Thompson estimator are robust in the sense that the procedure is unbiased for any set $\{y_1, y_2, \ldots, y_N\}$.

If the normality assumption is relaxed, the mean is optimal in the class of linear unbiased estimators (predictors).

Corollary 1.2.4.1. Let $\{y_1, y_2, \ldots, y_N\}$ be a set of iid random variables with mean μ and variance σ^2. Then the procedure of selecting a simple random nonreplacement sample of size n from N and using the estimator $\bar{y}_n$ is a minimum mean square error procedure for $\bar{y}_N$ in the design-estimator class composed of designs that are independent of $\{y_1, y_2, \ldots, y_N\}$ combined with linear estimators.

Proof. Under the iid assumption, the sample mean is the minimum mean square error linear predictor of $\bar{y}_{N-n}$ and the result follows. See Goldberger (1962) and Graybill (1976, Section 12.2) for discussions of best linear unbiased prediction.

■

The consideration of unequal probabilities of selection opens a wide range of options and theoretical difficulties. The very fact that one is able to associate unequal π_i with the elements means that we know something that differentiates element i from element j. It is no longer reasonable to treat the elements as exchangeable, that is, as a set for which the joint distribution does not depend on the indexing. However, it may be possible to transform the observations to achieve exchangeability. For example, it might be possible to define π_i such that it is reasonable to treat the $y_i\pi_i^{-1}$ as exchangeable. The nature of auxiliary information and the manner in which it should enter selection and estimation is the subject of survey sampling.

1.2.2 Poisson sampling

A sample design with simple theoretical properties is that in which samples are created by conducting N independent Bernoulli trials, one for each element in the population. If the result of the trial is a success, the element is included in the sample. Otherwise, the element is not part of the sample. The procedure is called *Poisson sampling* or *Bernoulli sampling* or *sieve sampling*.

Theorem 1.2.5. Let $(y_1, y_2, \ldots, y_N)$ be a finite universe of real-valued elements, and let $(\pi_1, \pi_2, \ldots, \pi_N)$ be a corresponding set of probabilities with $\pi_i > 0$ for all $i \in U$. For Poisson samples,

$$V\{\hat{T}_y - T_y \mid \mathcal{F}\} = \sum_{i=1}^{N} \pi_i^{-1}(1-\pi_i)y_i^2, \tag{1.2.42}$$

where T_y is the total defined in (1.2.21) and $\hat{T}_y$ is the Horvitz–Thompson estimator (1.2.23).

The expected sample size is

$$E\{n\} = \sum_{i=1}^{N} \pi_i \tag{1.2.43}$$

and

$$V\{n\} = \sum_{i=1}^{N} \pi_i(1 - \pi_i). \tag{1.2.44}$$

A design-unbiased estimator for the variance of $\hat{T}_y$ is

$$\hat{V}\{\hat{T}_y \mid \mathcal{F}\} = \sum_{i \in A} (1 - \pi_i)\pi_i^{-2} y_i^2. \tag{1.2.45}$$

If $\pi_i \equiv \pi$,

$$E\left\{ n^{-1} \sum_{i \in A} y_i \mid (\mathcal{F},\ n),\ n > 0 \right\} = \bar{y}_N \tag{1.2.46}$$

and

$$V\left\{ n^{-1} \sum_{i \in A} y_i \mid (\mathcal{F},\ n),\ n > 0 \right\} = N^{-1}(N - n)n^{-1} S_{y,N}^2. \tag{1.2.47}$$

Proof. Results (1.2.42), (1.2.43), and (1.2.44) follow from the fact that the I_i of $\mathbf{d} = (I_1, I_2, \ldots, I_N)$ are independent Bernoulli random variables. If $\pi_i \equiv \pi$, the sample size is a binomial random variable because it is the sum of N *iid* Bernoulli random variables. The set of samples with size $n = n_0$ is the set of simple random nonreplacement samples of size n_0, because every sample of size n_0 has the same probability of selection. Results (1.2.46) and (1.2.47) then follow. ■

Theorem 1.2.5 gives another example of difficulties associated with unconsidered use of the Horvitz–Thompson estimator. If $\pi_i \equiv \pi$, the Horvitz–Thompson estimator of the total of y for a Poisson sample is

$$\hat{T}_y = \pi^{-1} \sum_{i \in A} y_i \tag{1.2.48}$$

with variance

$$V\{\hat{T}_y \mid \mathcal{F}\} = \pi^{-2}\sum_{i=1}^{N}\pi(1-\pi)y_i^2. \tag{1.2.49}$$

By (1.2.46), another estimator of the total of a Poisson sample with $\pi_i \equiv \pi$ is

$$\begin{aligned}\tilde{T}_y &= N\bar{y}_n \quad \text{if } n > 0 \\ &= 0 \quad \text{if } n = 0,\end{aligned} \tag{1.2.50}$$

where $\bar{y}_n = n^{-1}\Sigma_{i\in A}y_i$. The estimator $N\bar{y}_n$ is conditionally unbiased for T_y for each positive n, and if $n = 0$, $\tilde{T}_y = \hat{T}_y = 0$. The mean square error of $\tilde{T}_y$ is

$$N^2E\left\{(n^{-1}-N^{-1})S_{y,N}^2 \mid (\mathcal{F},n),\ n>0\right\}P\{n>0\} + T_y^2P\{n=0\}. \tag{1.2.51}$$

Now $E\{n\} = \mu_n = N\pi$ and the variance of the Horvitz–Thompson estimator can be written

$$N^2(\mu_n^{-1}-N^{-1})\left[N^{-1}(N-1)S_{y,N}^2 + \bar{y}_N^2\right].$$

While $E\{n^{-1}|\ n>0\} > \mu_n^{-1}$, it is difficult to think of a situation in which one would choose the Horvitz–Thompson estimator over estimator (1.2.50). Note also that given $\pi_i \equiv \pi$,

$$\hat{V}\{\tilde{T}_y \mid (\mathcal{F},\ n),\ n>1\} = N^2(n^{-1}-N^{-1})s_{y,n}^2, \tag{1.2.52}$$

where $s_{y,n}^2$ is as defined in (1.2.37), is a conditionally unbiased estimator of the conditional variance of $\tilde{T}_y$, conditional on $n > 1$.

1.2.3 Stratified sampling

Assume that the elements of a finite population are divided into H groups, indexed by $h = 1, 2, \ldots, H$, called *strata*. Assume that the hth stratum contains N_h elements and it is desired to estimate the finite population mean

$$\bar{y}_N = N^{-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}y_{hi} = \sum_{h=1}^{H}N^{-1}N_h\bar{y}_{Nh}, \tag{1.2.53}$$

where $\bar{y}_{Nh} = N_h^{-1}\Sigma_{i=1}^{N_h}y_{hi}$ and y_{hi} is the ith element in the hth stratum.

Assume that we are willing to treat the elements in each stratum as if they were a random sample from a population with mean μ_h and variance σ_h^2. That is, the unknown values in stratum h are considered to be a realization of N_h

iid random variables. Thus, if one were selecting a sample to estimate the mean of an individual stratum, it is reasonable, by Theorems 1.2.3 and 1.2.4, to select a simple random sample from that stratum. Then, the sample mean of the stratum sample is an unbiased estimator of the population stratum mean. That is,

$$E\{\bar{y}_h \mid \mathcal{F}\} = \bar{y}_{Nh},$$

where

$$\bar{y}_h = n_h^{-1} \sum_{i \in A_h} y_{hi}$$

and A_h is the set of indices for the sample in stratum h. It follows that

$$\bar{y}_{st} = \sum_{h=1}^{H} N^{-1} N_h \bar{y}_h \tag{1.2.54}$$

is unbiased for the population mean, where $\bar{y}_h$ is the mean of a simple nonreplacement sample of size n_h selected from stratum h and the H stratum samples are mutually independent.

The procedure of selecting independent samples from H mutually exclusive and exhaustive subdivisions of the population is called *stratified sampling*, a very common technique in survey sampling. The samples within a stratum need not be simple random samples, but we concentrate on that case.

Because the $\bar{y}_h$ are independent,

$$V\{\bar{y}_{st} - \bar{y}_N \mid \mathcal{F}\} = \sum_{h=1}^{H} (N^{-1}N_h)^2 N_h^{-1}(N_h - n_h) n_h^{-1} S_h^2, \tag{1.2.55}$$

where $\bar{y}_{st}$ is as defined in (1.2.54) and

$$S_h^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_{Nh})^2.$$

The estimated variance of the stratum mean $\bar{y}_h$ is

$$\hat{V}\{(\bar{y}_h - \bar{y}_{Nh}) \mid \mathcal{F}\} = N_h^{-1}(N_h - n_h) n_h^{-1} s_h^2,$$

where

$$s_h^2 = (n_h - 1)^{-1} \sum_{i \in A_h} (y_{hi} - \bar{y}_h)^2.$$

It follows that an unbiased estimator of the variance of $\bar{y}_{st}$ is

$$\hat{V}\{(\bar{y}_{st} - \bar{y}_N) \mid \mathcal{F}\} = \sum_{h=1}^{H} (N^{-1}N_h)^2 N_h^{-1}(N_h - n_h)n_h^{-1}s_h^2. \quad (1.2.56)$$

Under the model in which the y_{hi} are realizations of $iid(\mu_h, \sigma_h^2)$ random variables, the unconditional variance is the expected value of (1.2.55),

$$V\{\bar{y}_{st} - \bar{y}_N\} = E\{V[\bar{y}_{st} - \bar{y}_N \mid \mathcal{F}]\} = \sum_{h=1}^{H} N^{-2}N_h(N_h - n_h)n_h^{-1}\sigma_h^2. \quad (1.2.57)$$

Assume that the objective of the design and estimation operation is to estimate the population mean, $\bar{y}_N$, of the characteristic y. Assume that a total amount C is available for sample observation and that it costs c_h to observe an element in stratum h. Under this scenario, one would choose the n_h to minimize the variance (1.2.55), or the variance (1.2.57), subject to the condition that

$$\sum_{h=1}^{H} n_h c_h \leq C.$$

The minimization requires knowledge of the S_h^2 or of σ_h^2.

In practice, one seldom knows the σ_h^2 at the time that one is constructing a sampling design. Thus, it is reasonable for the designer to construct a model for the population, to postulate parameters for the model, and to use the model and parameters as the basis for determining a design. The model is called the *design model*, and the parameters of the model are called *design parameters* or *anticipated parameters*.

The expected value of the design variance of the planned estimator calculated under the designer's model using the postulated parameters is called the *anticipated variance*. Let $\hat{\theta}$ be an estimator of a finite population parameter θ_N. Then the anticipated variance of $\hat{\theta} - \theta_N$ is

$$AV\{\hat{\theta} - \theta_N\} = E\{E[(\hat{\theta} - \theta_N)^2 \mid \mathcal{F}]\} - [E\{E(\hat{\theta} - \theta_N \mid \mathcal{F})\}]^2.$$

For stratified sampling $E\{\bar{y}_{st} - \bar{y}_N \mid \mathcal{F}\} = 0$ and the anticipated variance of $\bar{y}_{st}$ is minimized by minimizing

$$AV\{\bar{y}_{st} - \bar{y}_N\} = \sum_{h=1}^{H} N^{-2}N_h^2(n_h^{-1} - N_h^{-1})\ddot{\sigma}_h^2, \quad (1.2.58)$$

subject to the cost restriction, where $\ddot{\sigma}_h^2$, $h = 1, 2, \ldots, H$, are the anticipated stratum variances. If one uses the method of Lagrange multipliers, one

obtains

$$n_h = \lambda^{-1/2} c_h^{-1/2} N^{-1} N_h \ddot{\sigma}_h, \tag{1.2.59}$$

where λ is the Lagrange multiplier and

$$\lambda^{1/2} = C^{-1} \sum_{h=1}^{H} N^{-1} N_h c_h^{1/2} \ddot{\sigma}_h.$$

In general, the n_h of (1.2.59) are not integers. Also, it is possible for n_h to exceed N_h. The formal feasible solution could be obtained by using integer programming. In practice, the n_h are rounded to integers with all n_h greater than or equal to 2 and less than or equal to N_h. The allocation with n_h proportional to $N_h \ddot{\sigma}_h$ is optimal for constant costs and is sometimes called *Neyman allocation*, after Neyman (1934).

Our discussion is summarized in the following theorem.

Theorem 1.2.6. Let $\mathcal{F}$ be a stratified finite population in which the elements in stratum h are realizations of $iid(\mu_h, \sigma_h^2)$ random variables. Let $\ddot{\sigma}_h^2$, $h = 1, 2, \ldots, H$, be the anticipated variances, let C be the total amount available for sample observation, and assume that it costs c_h to observe an element in stratum h. Then a sampling and estimation strategy for $\bar{y}_N$ that minimizes the anticipated variance in the class of linear unbiased estimators and probability designs is: Select independent simple random nonreplacement samples in each stratum, selecting n_h in stratum h, where n_h is defined in (1.2.59), subject to the integer and population size constraints, and use the estimator defined in (1.2.54).

If it is desired to obtain a particular variance for a minimum cost, one minimizes cost subject to the variance constraint

$$V_S = \sum_{h=1}^{H} N^{-2} N_h^2 (n_h^{-1} - N_h^{-1}) \ddot{\sigma}_h^2,$$

where V_S is the variance specified. In this case,

$$n_h = \left(V_S + \sum_{h=1}^{H} N^{-2} N_h \ddot{\sigma}_h^2 \right)^{-1} \left(\sum_{h=1}^{H} N^{-1} N_h c_h^{1/2} \ddot{\sigma}_h \right) N^{-1} N_h c_h^{-1/2} \ddot{\sigma}_h.$$

In both cases, n_h is proportional to $N^{-1} N_h c_h^{-1/2} \ddot{\sigma}_h$.

1.2.4 Systematic sampling

Systematic sampling is used widely because of the simplicity of the selection procedure. Assume that it is desired to select a sample of size n from a population of size N with probabilities π_i, $i = 1, 2, \ldots, N$, where $0 < \pi_i < 1$. To introduce the procedure, consider the population of 11 elements displayed in Table 1.1. Assume that it is desired to select a sample of four elements with probabilities of selection that are proportional to the measures of size. The sum of the sizes is 39. Thus, π_i is the size of the ith element divided by 39 and multiplied by 4. The third column contains the cumulated sizes, and the fourth column contains the cumulated sizes, normalized so that the sum is 4.

To select a systematic random sample of four elements, we select a random number in the interval (0, 1). For our example, assume that the random number is 0.4714. Then the elements in the cumulated normalized sum associated with the numbers 0.4714, 1.4714, 2.4714, and 3.4714 constitute the sample. Let the cumulated size for element t be C_t, where the elements are numbered $1, 2, \ldots, N$. Then an element is associated with the number ζ if

$$C_{t-1} < \zeta \leq C_t.$$

The probability that one of the numbers $(RN,\ RN+1,\ RN+2,\ RN+3)$, where RN is a random number in (0, 1), falls in the interval $(C_{t-1},\ C_t]$ is π_t.

Table 1.1 Selection of a Systematic Sample

Element Number	Measure of Size	Cumulated Size	Normalized Cumulated Size	Random Number and Increments
1	6	6	0.6154	0.4714
2	5	11	1.1282	
3	6	17	1.7436	1.4714
4	4	21	2.1538	
5	5	26	2.6667	2.4714
6	4	30	3.0769	
7	2	32	3.2821	
8	3	35	3.5897	3.4714
9	2	37	3.7949	
10	1	38	3.8974	
11	1	39	4.0000	

The selection is particularly simple if $N = nk$, where k is an integer and the elements are to be selected with equal probability. Then the selection consists of selecting a random integer between 1 and k inclusive, say r. The sample is composed of elements $r,\ r+k,\ r+2k,\ \ldots,\ r+(n-1)k$. For this situation, there are k possible samples and we have

$$\begin{aligned} \pi_i &= k^{-1} && \text{for all } i \\ \pi_{ij} &= k^{-1} && \text{if } j = i + km \text{ or } j = i - km \\ &= 0 && \text{otherwise,} \end{aligned} \tag{1.2.60}$$

where m is an integer. Because $\pi_{ij} = 0$ for some pairs, it is not possible to construct a design-unbiased estimator of the variance of a systematic sample.

If the elements are arranged in random order and if the elements are selected with equal probability, systematic sampling produces a simple random nonreplacement sample. Sometimes, for populations in natural order, the variance is estimated as if the sample were a random nonreplacement sample. Such variance calculation is appropriate if the natural order is equivalent to random order. More often, adjacent pairs are assigned to pseudo strata and the variance estimated as if the sample were a two-per-stratum stratified sample. See Section 5.3.

Systematic samples are sometimes defined with random sample sizes. For example, we might draw a sample from a population of size N by selecting a random integer between 1 and k inclusive, say r. Let the sample be elements $r,\ r+k,\ \ldots$, where the last element is $r+(q-1)k$ and $N-k < r+(q-1)k \leq N$. Let

$$N = kq + L,$$

where q and L are integers and $0 \leq L < k$; then L of the samples are of size $q+1$, and $k-L$ of the samples are of size q. Because every element has a probability k^{-1} of being selected, the Horvitz–Thompson estimator is unbiased for the population total. However, the estimator $\bar{y}_n$, where $\bar{y}_n$ is the sample mean, is slightly biased for the population mean.

To consider systematic sampling for the mean of a population arranged in natural order, assume that the superpopulation satisfies the stationary first-order autoregressive model

$$\begin{aligned} y_t &= \rho y_{t-1} + e_t, \\ e_t &\sim NI(0,\ \sigma^2), \end{aligned} \tag{1.2.61}$$

where $0 < \rho < 1$ and the symbol $\sim$ means "is distributed as." Under this model

$$V\{y_t\} = (1-\rho^2)^{-1}\sigma^2,$$

$$C\{y_t,\ y_{t+j}\} \quad = \quad (1-\rho^2)^{-1}\rho^{|j|}\sigma^2,$$

and the correlation between unit t and unit $t+j$, denoted by $\rho(j)$, is $\rho^{|j|}$. A systematic sample of size n selected from a population of size $N = nk$ generated by model (1.2.61) nearly minimizes the variance of the sample mean as an estimator of the finite population mean. Under the extended model with correlations that satisfy

$$\rho(i) - 2\rho(i+1) + \rho(i+2) \geq 0 \qquad \text{for } i = 0,\ 1,\ 2,\ \ldots,$$

Papageorgiou and Karakostas (1998) show that the optimal design for the population mean using the sample mean as the estimator is the systematic sample with the index of the first unit equal to the integer approximation of $(2n)^{-1}(N-n)$. Blight (1973) pointed out that the optimal linear estimator of the population mean under the model (1.2.61) is a weighted combination that gives more weight to the first and last observations than to the middle observations. Such selection and estimation procedures, although the best under the model, are not design unbiased.

Systematic sampling is efficient relative to simple random nonreplacement sampling for populations with a linear trend. Assume that the population satisfies the model

$$y_t \quad = \quad \beta_0 + \beta_1 t + e_t, \tag{1.2.62}$$

where e_t are $iid(0,\ \sigma^2)$ random variables. Then, for a population of size $N = kn$, the variance, under the model, of the random-start systematic sample mean as an estimator of the population mean is

$$\begin{aligned} V\{\bar{y}_{sys} - \bar{y}_N\} \quad &= \quad (12)^{-1}(k+1)(k^2-k)\beta_1^2 \\ &\quad + n^{-1}k^{-1}(k-1)\sigma^2 \\ &\doteq \quad (12)^{-1}k^3\beta_1^2 + n^{-1}\sigma^2 \end{aligned} \tag{1.2.63}$$

for large k. The variance of the sample mean for a simple random nonreplacement sample is approximately

$$V\{\bar{y}_{srs} - \bar{y}_N\} \quad \doteq \quad (12)^{-1}n^2k^3\beta_1^2 + n^{-1}\sigma^2\,. \tag{1.2.64}$$

If the ordered population is divided into n strata of size k and one element is selected in each stratum, the variance of the stratified mean as an estimator of the population mean is

$$\begin{aligned} V\{\bar{y}_{st} - \bar{y}_N\} \quad &= \quad n^{-1}k^{-1}(k-1)(12)^{-1}(k+1)(k^2-k)\beta_1^2 \\ &\quad + n^{-1}k^{-1}(k-1)\sigma^2 \\ &\doteq \quad n^{-1}(12)^{-1}k^3\beta_1^2 + n^{-1}\sigma^2. \end{aligned} \tag{1.2.65}$$

Thus, because the stratified sample mean averages over the local linear trends, it is more efficient than the systematic sample. It is not possible to construct a design-unbiased estimator of the variance for either the one-per-stratum or the systematic designs.

Systematic sampling can also be inefficient relative to simple random nonreplacement sampling. Assume that the y values satisfy

$$y_t = \sin 2\pi k^{-1} t.$$

Then the values in a systematic sample of interval k are identical. Hence, for this population, the variance of the mean is greater than the variance of a simple random nonreplacement sample. Furthermore, because the within-sample variation observed is zero, the estimated variance is zero when the variance is estimated as if the sample were a simple random sample.

See Bellhouse (1988) for a review of systematic sampling. Variance estimation for systematic samples is considered in Section 5.3.

1.2.5 Replacement sampling

Consider a sampling scheme in which repeated selections of a single element are made from a population of elements. Let the selection probabilities for each selection, or draw, for the N elements be p_{ri}, $i = 1, 2, \ldots, N$, where

$$\sum_{i=1}^{N} p_{ri} = 1.$$

Then a replacement sample of size n is that obtained by selecting an element from the N elements with probability p_{ri} at each of n draws. Such a procedure may produce a sample in which element i appears more than once. An estimator of the population total is

$$\hat{T}_{yR} = n^{-1} \sum_{i \in A} p_{ri}^{-1} y_i t_i = \sum_{d=1}^{n} n^{-1} p_d^{-1} y_d, \tag{1.2.66}$$

where t_i is the number of times that element i is selected in the sample, and (p_d, y_d) is the value of (p_{ri}, y_i) for the element selected on the dth draw. Although simple replacement sampling is seldom used in practice, its properties are useful in theoretical discussions.

We may omit the descriptor *nonreplacement* when discussing nonreplacement samples, but we always use the descriptor *replacement* when discussing replacement samples. A replacement sample can be considered to be a random sample selected from an infinite population with the value $p_{ri}^{-1} y_i =: z_i$

occurring with frequency p_{ri}, where the symbol $=:$ means "is defined to equal." Thus, the variance of the infinite population of z's is

$$\sigma_z^2 = E\{(z_i - \mu_z)^2\} = \sum_{i=1}^{N} p_{ri}(z_i - \mu_z)^2$$
$$= \sum_{i=1}^{N} p_{ri}(p_{ri}^{-1} y_i - T_y)^2, \qquad (1.2.67)$$

where

$$\mu_z = \sum_{i=1}^{N} p_{ri} z_i = \sum_{i=1}^{N} y_i = T_y.$$

The estimator $\hat{T}_{yR}$ is the mean of n iid random variables with mean μ_z and variance σ_z^2. Thus,

$$V\{\hat{T}_{yR}\} = V\left\{n^{-1}\sum_{d=1}^{n} z_d\right\} = n^{-1}\sigma_z^2, \qquad (1.2.68)$$

where z_d is the value of z obtained on the dth draw. Furthermore,

$$\hat{V}\{\hat{T}_{yr}\} = n^{-1}(n-1)^{-1}\sum_{d=1}^{n}(z_d - \bar{z}_n)^2, \qquad (1.2.69)$$

where

$$\bar{z}_n = n^{-1}\sum_{d=1}^{n} z_d$$

is unbiased for $V\{\hat{T}_{yR}\}$. The simplicity of the estimator (1.2.69) has led to its use as an approximation in nonreplacement unequal probability sampling when all of the np_{ri} are small.

The fact that some elements can be repeated in the estimator (1.2.66) is an unappealing property. That the estimator is not efficient is seen most easily when all p_{ri} are equal to N^{-1}. Then

$$\hat{T}_{yR} = Nn^{-1}\sum_{i\in A} y_i t_i. \qquad (1.2.70)$$

Because the draws are independent, the sample of unique elements is a simple random nonreplacement sample. Thus, an unbiased estimator of the mean is

$$\bar{y}_u = n_u^{-1}\sum_{i\in A} y_i, \qquad (1.2.71)$$

where n_u is the number of unique elements in the sample. The conditional variance of the mean associated with (1.2.66) conditional on $(t_1, t_2, \ldots, t_{n_u})$ is

$$V\{N^{-1}\hat{T}_{yR} \mid (t_1, t_2, \ldots, t_{n_u})\} = n^{-2}\sum_{i\in A} t_i^2\sigma_y^2, \quad (1.2.72)$$

while

$$V\{\bar{y}_u \mid (t_1, t_2, \ldots, t_{n_u})\} = n_u^{-1}\sigma_y^2. \quad (1.2.73)$$

Because $\Sigma t_i^2 \geq n_u$, with equality for $n_u = n$, the mean of unique units is conditionally superior to the mean associated with (1.2.66) for every $1 < n_u < n$.

1.2.6 Rejective sampling

Rejective sampling is a procedure in which a sample is selected by a particular rule but is accepted only if it meets certain criteria. The selection operation is repeated until an acceptable sample is obtained. The procedure is sometimes called *restrictive sampling*. In most situations, the rejection of certain samples changes the inclusion probabilities. Hájek (1964, 1981) studied two kinds of rejective sampling. In the first, a replacement sample is selected and the sample is kept only if it contains no duplicates. In the second, a Poisson sample is selected and is kept only if it contains exactly the number of elements desired.

To illustrate the effect of the restriction on probabilities, consider the selection of a Poisson sample from a population of size 4 with selection probabilities (0.2, 0.4, 0.6, 0.8) for i =1, 2, 3, 4. Let the sample be rejected unless exactly two elements are selected. The probabilities of the six possible samples of size 2 are (0.0064, 0.0144, 0.0384, 0.0384, 0.1024, 0.2304) for the samples [(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)], respectively. It follows that the rejective procedure gives inclusion probabilities (0.1375, 0.3420, 0.6580, 0.8625) for $i = 1, 2, 3, 4$. See Section 1.4 for references on the use of rejective sampling with unequal probabilities.

To illustrate how the inclusion probabilities are changed by other rejection rules, consider the selection of a sample of size 3 from a sample of 6, where the elements are numbered from 1 to 6. Assume that the procedure is to select a simple random sample of size 3 but to reject the sample if it contains three adjacent elements. Thus, samples (1, 2, 3), (2, 3, 4), (3, 4, 5), and (4, 5, 6) are rejected. If the sample is rejected, a new simple random sample is selected until an acceptable sample is obtained. There are 20 possible simple random samples and 16 acceptable samples. Therefore, the probabilities of inclusion

in an acceptable sample are (9/16, 8/16, 7/16, 7/16, 8/16, 9/16) for elements (1, 2, 3, 4, 5, 6), respectively.

As a second example, let x be the ordered identification for a population of size 8. Assume that we select samples using simple random sampling but reject any sample with a mean of x less than 2.5 or greater than 6.5. Thus, the four samples (1, 2, 3), (1, 2, 4), (5, 7, 8), and (6, 7, 8) are rejected. The resulting probabilities of inclusion are (19/52, 19/52, 20/52, 20/52, 20/52, 20/52, 19/52, 19/52) for elements (1, 2, 3, 4, 5, 6, 7, 8), respectively. These two examples illustrate the general principles that rejecting adjacent items increases the relative probability of boundary elements, and rejecting samples with large $\mid \bar{x}_n - \bar{x}_N \mid$ decreases the relative probability of extreme observations. In these simple examples, one can construct the Horvitz–Thompson estimator using the correct inclusion probabilities.

Many practitioners employ modest types of rejective sampling when the unit identification carries information. For example, let an ordered population be divided into m strata of size k, with two elements selected in each stratum. Practitioners would be tempted to reject a sample composed of the two largest elements in each stratum. The probability of such a sample is $[0.5k(k-1)]^{-m}$ for m strata of size k. If only this sample and the similar sample of the two smallest elements are rejected, the inclusion probabilities will be little affected for large k and m. On the other hand, if a large fraction of possible samples are rejected, the inclusion probabilities can be changed by important amounts.

1.2.7 Cluster samples

In much of the discussion to this point we have considered a conceptual list of units, where the units can be given an identification and the identifications can be used in sample selection. In Example 1.2.1 we introduced the possibility that the units on the frame are not the units of final interest. In that example, households are of interest and are the units observed, but the units sampled are blocks, where there will be several households in a block. Samples of this type are called *cluster samples*. It is also possible for the units of analysis to differ from the sampling units and from the observation units. Assume that data are collected for all persons in a household using a single respondent for the household and that the analyst is interested in the fraction of people who had flu shots. Then the analysis unit is a person, the observation unit is the household, and the sampling unit is the block.

In estimation formulas such as (1.2.23) and (1.2.33), the variable y_i is the total for the ith sampling unit. In Example 1.2.1, y_i is the total for a block. It is very easy for analysts to treat analysis units or observation units incorrectly as sampling units. One must always remember the nature of the units on the sampling frame.

From a statistical point of view, no new concepts are involved in the construction of estimators for cluster samples. If we let M_i be the number of elements in the ith cluster and let y_{ij} be the value for the jth element in the ith cluster, then

$$y_i = \sum_{j=1}^{M_i} y_{ij}$$

and the estimator of a total is

$$\hat{T}_y = \sum_{i \in A} \pi_i^{-1} y_i, \tag{1.2.74}$$

where π_i is the probability of selection for the ith cluster and y_i is the total of the characteristic for all persons in the ith cluster. Similarly, variance estimators such as (1.2.32) and (1.2.33) are directly applicable.

1.2.8 Two-stage sampling

In many situations it is efficient first to select a sample of clusters and then select a subsample of the units in each cluster. In this case, the cluster is called a *primary sampling unit* (PSU), and the sample of primary sampling units is called the *first-stage sample*. The units selected in the subsample are called *secondary sampling units* (SSUs), and the sample of secondary sampling units is called the *second-stage sample*.

We adopt the convention described by Särndal, Swensson, and Wretman (1992, p. 134). If the sample is selected in two steps (stages), if units selected at the second step are selected independently in each first-step unit, and if the rules for selection within a first-step unit depend only on that unit and not on other first step units in the sample, the sample is called a *two-stage sample*.

The Horvitz–Thompson estimator of the total for a two-stage sample is

$$\hat{T}_{2s} = \sum_{i \in A_1} \sum_{j \in B_i} \pi_{(ij)}^{-1} y_{ij}, \tag{1.2.75}$$

where $\pi_{(ij)} = \pi_i \pi_{(ij)|i}$ is the probability that second-stage unit ij is selected in the sample, π_i is the probability that first-stage unit i is selected, $\pi_{(ij)|i}$ is the probability that second-stage unit ij is selected given that first-stage unit i is selected, A_1 is the set of indices for first-stage units in the sample, and B_i is the set of second-stage units in first-stage unit i that are in the sample.

The estimator (1.2.75) is unbiased for the total by the properties of the Horvitz–Thompson estimator. The joint probabilities are

$$\begin{aligned} \pi_{(ij)(km)} &= \pi_i \pi_{(ij)(im)|i} && \text{if } i = k \text{ and } j \neq m \\ &= \pi_{ik} \pi_{(ij)|i} \pi_{(km)|k} && \text{if } i \neq k \text{ and } j \neq m, \end{aligned} \tag{1.2.76}$$

where π_{ik} is the probability that first-stage units i and k are selected, and $\pi_{(ij)(im)|i}$ is the probability that elements ij and im are selected given that PSU i is selected. Given these probabilities, the variances and estimated variances of the Horvitz–Thompson estimator are defined. We present some more convenient expressions for the variance and estimated variance.

Consider a sample of n_1 PSUs selected from a finite population which is, itself, a sample of N PSUs selected from an infinite population of PSUs. Let the ith PSU be selected with probability π_i and let the ith PSU contain M_i secondary sampling units. Let a nonreplacement probability sample of m_i units be selected from the M_i. Then an alternative expression for the estimator of (1.2.75) is

$$\hat{T}_{2s} = \sum_{i \in A_1} \pi_i^{-1} \hat{y}_i, \tag{1.2.77}$$

where

$$\hat{y}_i = \sum_{j \in B_i} \pi_{(ij)|i}^{-1} y_{ij}$$

and B_i is as defined in (1.2.75). The design variance of $\hat{T}_{2s}$ is

$$\begin{aligned} V\{\hat{T}_{2s} \mid \mathcal{F}\} &= V\{E[\hat{T}_{2s} \mid (A_1, \mathcal{F})] \mid \mathcal{F}\} + E\{V[\hat{T}_{2s} \mid (A_1, \mathcal{F})] \mid \mathcal{F}\} \\ &= V_1\{\hat{T}_{1s} \mid \mathcal{F}\} + E\{V[\hat{T}_{2s} \mid (A_1, \mathcal{F})] \mid \mathcal{F}\}, \end{aligned} \tag{1.2.78}$$

where $\hat{T}_{1s}$ is the estimated total with all $m_i = M_i$, $V_1\{\hat{T}_{1s} \mid \mathcal{F}\}$ is the variance of the estimated total with $m_i = M_i$ for all i, and $V[\hat{T}_{2s}|(A_1, \mathcal{F})]$ is the conditional design variance, conditional on the first-stage units selected. Generally, a design consistent estimator of $V[\hat{T}_{2s} \mid (A_1, \mathcal{F})]$ is available and can be used to estimate $E\{V[\hat{T}_{2s} \mid (A_1, \mathcal{F})]\}$. Estimation of $V_1\{\hat{T}_{1s} \mid \mathcal{F}\}$ is more difficult. Consider a quadratic function of the y_i, such as the Horvitz–Thompson estimator, for $V_1\{\hat{T}_{1s} \mid \mathcal{F}\}$,

$$\tilde{V}_1\{\hat{T}_{1s} \mid \mathcal{F}\} = \sum_{i \in A_1} \sum_{j \in A_1} \alpha_{ij} y_i y_j,$$

where α_{ij} are fixed coefficients. If y_i is replaced with $\hat{y}_i$, we have

$$\begin{aligned} E\left\{\sum_{i \in A_1} \sum_{j \in A_1} \alpha_{ij} \hat{y}_i \hat{y}_j \mid (A_1, \mathcal{F})\right\} &= \sum_{i \in A_1} \sum_{j \in A_1} \alpha_{ij} y_i y_j \\ &\quad + \sum_{i \in A_1} \alpha_{ii} V\{\hat{y}_i \mid (A_1, \mathcal{F})\}, \end{aligned} \tag{1.2.79}$$

because

$$E\left\{\hat{y}_i^2 \mid (A_1, \mathcal{F})\right\} = y_i^2 + V\left\{\hat{y}_i - y_i \mid (A_1, \mathcal{F})\right\}, \tag{1.2.80}$$

and given that samples are selected independently within each PSU,

$$E\left\{\hat{y}_i \hat{y}_j \mid (A_1, \mathcal{F})\right\} = y_i y_j \qquad \text{for } i \neq j. \tag{1.2.81}$$

Thus, given a quadratic estimator of variance for the first stage, an estimator of $V\{\hat{T}_{2s} \mid \mathcal{F}\}$ is

$$\hat{V}\{\hat{T}_{2s} \mid \mathcal{F}\} = \hat{V}_1\{\hat{T}_{1s} \mid \mathcal{F}\} + \sum_{i \in A_1} (\pi_i^{-2} - \alpha_{ii}) \hat{V}\{\hat{y}_i \mid (A_1, \mathcal{F})\}, \tag{1.2.82}$$

where $\hat{V}_1\{\hat{T}_{1s} \mid \mathcal{F}\}$ is the estimated design variance for the first-stage sample computed with $\hat{y}_i$ replacing y_i.

The coefficient for y_i^2 in the design variance of a design linear estimator is $\pi_i^{-1}(1 - \pi_i)$. It follows that the α_{ii} in a design-unbiased quadratic estimator of the variance of a design linear estimator is $\pi_i^{-2}(1 - \pi_i)$. Therefore, the bias in $\hat{V}_1\{\hat{T}_{2s} \mid \mathcal{F}\}$ as an estimator of $V\{\hat{T}_{2s} \mid \mathcal{F}\}$ is the sum of the $\pi_i V\{y_i \mid (A_1, \mathcal{F})\}$, and the bias is small if the sampling rates are small.

For a simple random sample of PSUs, the variance of the estimator of the total for a complete first stage is

$$V_1\{\hat{T}_{1s} \mid \mathcal{F}\} = N^2(1 - f_1) n_1^{-1} S_{1y}^2, \tag{1.2.83}$$

where $f_1 = N^{-1} n_1$,

$$\begin{aligned} S_{1y}^2 &= (N-1)^{-1} \sum_{i \in U} (y_i - \bar{y}_N)^2, \\ y_i &= \sum_{j=1}^{M_i} y_{ij}, \end{aligned}$$

and $\bar{y}_N = N^{-1} \Sigma_{i \in U} y_i$. Given that the samples within the first-stage units are simple random nonreplacement samples,

$$V\{\hat{T}_{2s} \mid (A_1, \mathcal{F})\} = \sum_{i \in A_1} \pi_i^{-2} M_i^2 \left(1 - M_i^{-1} m_i\right) m_i^{-1} S_{2yi}^2, \tag{1.2.84}$$

where

$$S_{2yi}^2 = (M_i - 1)^{-1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{Mi})^2$$

and

$$\bar{y}_{Mi} = M_i^{-1} \sum_{j=1}^{M_i} y_{ij}.$$

For simple random sampling at the second stage,

$$\hat{V}\{\hat{y}_i \mid (A_1, \mathcal{F})\} = M_i^2 \left(1 - M_i^{-1} m_i\right) m_i^{-1} s_{2yi}^2, \tag{1.2.85}$$

where

$$s_{2yi}^2 = (m_i - 1)^{-1} \sum_{j \in B_i} (y_{ij} - \bar{y}_i)^2$$

and

$$\bar{y}_{i\cdot} = m_i^{-1} \sum_{j \in B_i} y_{ij}$$

is design consistent for $E\{V[\hat{T}_{2s} \mid (A_1, \mathcal{F})] \mid \mathcal{F}\}$. The expected value of the estimator of the variance (1.2.83) constructed by replacing y_i with $\hat{y}_i$ is

$$\begin{aligned} E\{\hat{V}_{1,srs}(\hat{T}_{1s} \mid \mathcal{F}) \mid \mathcal{F}\} &= E[C_f (n_1 - 1)^{-1} \sum_{i \in A_1} (\hat{y}_i - \bar{y}_{n1})^2 \mid \mathcal{F}] \\ &= C_f S_{1y}^2 + C_f n_1^{-1} \sum_{i \in A_1} M_i (M_i - m_i) m_i^{-1} S_{2yi}^2, \end{aligned} \tag{1.2.86}$$

where $C_f = N^2 (1 - f_1) n_1^{-1}$, $f_1 = N^{-1} n_1$, and

$$\bar{y}_{n1} = n_1^{-1} \sum_{i \in A_1} \hat{y}_i.$$

Therefore, an unbiased estimator of the variance of $\hat{T}_{2s}$ is, for simple random nonreplacement sampling at both stages,

$$\begin{aligned} \hat{V}\{\hat{T}_{2s} \mid \mathcal{F}\} &= \hat{V}_{1,srs}\{\hat{T}_{2s} \mid \mathcal{F}\} \\ &\quad + f_1 N^2 n_1^{-2} \sum_{i \in A_1} M_i^2 \left(1 - M_i^{-1} m_i\right) m_i^{-1} s_{2yi}^2, \end{aligned} \tag{1.2.87}$$

where $\hat{V}_{1,srs}\{\hat{T}_{1s} \mid \mathcal{F}\}$ is defined in (1.2.86) and s_{2yi}^2 is defined in (1.2.85). The first-stage estimated variance in (1.2.86) is a quadratic in y_i with $\alpha_{ii} = (1 - f_1) N^2 n_1^{-2}$. Furthermore, $\pi_i^{-2} = N^2 n_1^{-2}$ and $\hat{V}_{1,srs}\{\hat{T}_{2s} \mid \mathcal{F}\}$ is the

dominant term in $\hat{V}\{\hat{T}_{2s} \mid \mathcal{F}\}$ when the finite population correction is close to 1.

To construct estimator (1.2.87), we require that $m_i \geq 2$ for all i and require the assumption of independent simple random nonreplacement samples within first-stage units. Estimator (1.2.82) only requires that the second-stage design be such that a reasonable estimator of the second-stage variance is available for every PSU.

If the finite population correction can be ignored, the estimator $\hat{V}_1\{\hat{T}_{2s} \mid \mathcal{F}\}$ is consistent for the variance under any selection scheme for secondary units, such that $\hat{T}_{2s}$ is an unbiased estimator and the selection within a PSU is independent of the selection in other PSUs. This follows from (1.2.80) and (1.2.81). Thus, for example, one could stratify each of the PSUs and select stratified samples of secondary units within each PSU.

Example 1.2.2. We use data from the U.S. National Resources Inventory (NRI) in a number of examples. The NRI is conducted by the U.S. Natural Resources Conservation Service in cooperation with the Iowa State University Center for Survey Statistics and Methodology. The survey is a panel survey of land use conducted in 1982, 1987, 1992, 1997, and yearly since 2000. Data are collected on soil characteristics, land use, land cover, wind erosion, water erosion, and conservation practices. The sample is a stratified area sample of the United States, where the primary sampling units are areas of land called *segments*. Data are collected for the entire segment on such items as urban lands, roads, and water. Detailed data on soil properties and land use are collected at a random sample of points within the segment. The sample for 1997 contained about 300,000 segments with about 800,000 points. The yearly samples are typically about 70,000 segments. See Nusser and Goebel (1997) for a more complete description of the survey.

We use a very small subsample of the Missouri NRI sample for the year 1997 to illustrate calculations for a two-stage sample. The true first-stage sampling rates are on the order of 2%, but for the purposes of illustration, we use the much higher rates of Table 1.2. In Missouri, segments are defined by the Public Land Survey System. Therefore, most segments are close to 160 acres in size, but there is some variation in size due to variation in sections defined by the Public Land Survey System and due to truncation associated with county boundaries. The segment size in acres is given in the fourth column of the table. The points are classified using a system called *broaduse*, where example broaduses are urban land, cultivated cropland, pastureland, and forestland. Some of the broaduses are further subdivided into categories called *coveruses*, where corn, cotton, and soybeans are some of the coveruses within the cropland broaduse.

Table 1.2 Missouri NRI Data

Stratum	PSU	Weight	Segment Size	Total No. Pts	No. Pts. Forest	s^2_{2yi}	$\hat{y}_i$
1	1	3.00	195	3	2	0.1111	130
	2	3.00	165	3	3	0	165
	3	3.00	162	3	2	0.1111	54
	4	3.00	168	3	0	0	0
	5	3.00	168	3	2	0.1111	112
	6	3.00	100	2	1	0.2500	50
	7	3.00	180	3	0	0	0
2	1	5.00	162	3	1	0.1111	54
	2	5.00	174	3	1	0.1111	58
	3	5.00	168	3	2	0.1111	112
	4	5.00	174	3	0	0	0

In this example, we estimate the acres of forestland and define

$$\begin{aligned} y_{ij} &= 1 \quad \text{if point } j \text{ in PSU } i \text{ is forest} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The total number of points in the segment is given in the fifth column and the number that are forest is given in the sixth column. In a typical data set there would be a row for each point, and the sum for the segment would be calculated as part of the estimation program. Treating each point as if it represents 1 acre, we have

$$\hat{y}_{hi} = M_{hi} m_{hi}^{-1} \sum_{j=1}^{m_{hi}} y_{hij},$$

where M_{hi} represents the acres (SSUs) in segment i of stratum h and m_{hi} the number of sample points (SSUs) in the segment. Thus, the estimated total acres of forest for PSU 1 in stratum 1 is 130, and the estimated variance for that estimated segment total is

$$\hat{V}\{\bar{y}_{1,1} \mid (A_1, \mathcal{F})\} = 195(195 - 3)3^{-1}(0.1111) = 1386.67,$$

where $s^2_{2y,1,1} = 0.1111$ is as defined in (1.2.85).

The estimated acres of forest for this small region is

$$\hat{T}_{2s} = \sum_{h=1}^{2} N_h n_{1h}^{-1} \sum_{i \in A_{1h}} \hat{y}_{hi} = 3(514) + 5(222) = 2652,$$

where n_{1h} is the number of sample segments (PSUs) in stratum h.

Equation (1.2.87) extends immediately to stratified sampling and we have

$$\begin{aligned}\hat{V}\{\hat{T}_{2s} \mid \mathcal{F}\} &= \sum_{h=1}^{2} N_h^2(1-f_{1h})n_{1h}^{-1}\hat{s}_{1h}^2 \\ &\quad + \sum_{h=1}^{2} N_h^2 n_{1h}^{-2} f_{1h} \sum_{i=1}^{n_{1h}} M_{hi}(M_{hi}-m_{hi})m_{hi}^{-1}s_{2y,hi}^2 \\ &= 340.940 + 29.190 = 370.130,\end{aligned}$$

where $f_{1h} = N_h^{-1}n_{1h}$,

$$\hat{s}_{1h}^2 = (n_{1h}-1)^{-1} \sum_{i \in A_{1h}} (\hat{y}_{hi} - \bar{y}_{h,n1})^2,$$

and $\bar{y}_{h,n1}$ is the stratum analog of $\bar{y}_{n1}$ of (1.2.86). The values of the first-stage estimated variances are $(\hat{s}_{1,1}^2, \hat{s}_{1,2}^2) = (4130.3, 2093.3)$. There is a sizable correlation between points within a segment for a broaduse such as forest, and the between-PSU portion dominates the variance. ■■

1.3 LIMIT PROPERTIES

1.3.1 Sequences of estimators

We define sequences that will permit us to establish large-sample properties of sample designs and estimators. Our sequences will be sequences of finite populations and associated probability samples. A set of indices is used to identify the elements of each finite population in the sequence. To reduce the number of symbols required, we usually assume that the Nth finite population contains N elements. Thus, the set of indices for the Nth finite population is

$$U_N = \{1, 2, \ldots, N\}, \tag{1.3.1}$$

where $N = 1, 2, \ldots$. Associated with the ith element of the Nth population is a column vector of characteristics, denoted by $\mathbf{y}_{iN}$. Let

$$\mathcal{F}_N = (\mathbf{y}_{1N}, \mathbf{y}_{2N}, \ldots, \mathbf{y}_{NN})$$

be the set of vectors for the Nth finite population. The set $\mathcal{F}_N$ is often called simply the Nth *finite population* or the Nth *finite universe*.

Two types of sequences $\{\mathcal{F}_N\}$ may be specified. In one, the set $\mathcal{F}_N$ is a set of fixed vectors from a fixed sequence. In the other, the vectors

$\mathbf{y}_{iN}$, $i = 1, 2, \ldots, N$, are random variables. For example, the $\{\mathbf{y}_{iN}\}$, $i = 1, 2, \ldots, N$, might be the first N elements of the sequence $\{y_i\}$ of iid random variables with distribution function $F(y)$ such that

$$E\{y_i\} = \mu \tag{1.3.2}$$

and

$$E\{(y_i - \mu)^2\} = \sigma^2. \tag{1.3.3}$$

If necessary to avoid confusion, we will add subscripts so that, for example, μ_y denotes the mean of y and σ_y^2 or σ_{yy} denotes the variance of y.

As defined previously, the finite population mean and variance for scalar y are

$$\bar{y}_N = N^{-1} \sum_{i=1}^{N} y_{iN} \tag{1.3.4}$$

and

$$S_{y,N}^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_{iN} - \bar{y}_N)^2. \tag{1.3.5}$$

The corresponding quantities for vectors are

$$\bar{\mathbf{y}}_N = N^{-1} \sum_{i=1}^{N} \mathbf{y}_{iN} \tag{1.3.6}$$

and

$$\mathbf{S}_{yy,N} = (N-1)^{-1} \sum_{i=1}^{N} (\mathbf{y}_{iN} - \bar{\mathbf{y}}_N)\,(\mathbf{y}_{iN} - \bar{\mathbf{y}}_N)'. \tag{1.3.7}$$

Recall that a sample is defined by a subset of the population indices and let A_N denote the set of indices appearing in the sample selected from the Nth finite population. The number of distinct indices appearing in the sample is called the *sample size* and is denoted by n_N. We assume that samples are selected according to the probability rule $p_N(A)$ introduced in Section 1.2.

Example 1.3.1. As an example of a sequence of populations, consider the sequence of sets of $N = 10j$ elements, where $j = 1, 2, \ldots$. Let iid Bernoulli random variables be associated with the indexes $1, 2, \ldots, N$. From each set of $10j$ values realized, a simple random nonreplacement sample

of size $n_N = j$ is selected. In this case it is possible to give the exact form of the relevant distributions. Assume that the Bernoulli variable is such that

$$\begin{aligned} x_i &= 1 \quad \text{with probability } p \\ &= 0 \quad \text{with probability } (1-p). \end{aligned}$$

Then the distribution of

$$X_N = \sum_{i=1}^{N} x_i$$

is that of a binomial random variable with parameters $(N,\ p)$ and

$$P\{X_N = a\} = \binom{N}{a} p^a (1-p)^{N-a}.$$

Because the elements are independent, the unconditional distribution of the sample sum, X_n, is binomial with parameters $(n,\ p)$,

$$P\{X_n = a\} = \binom{n}{a} p^a (1-p)^{n-a}.$$

Now a particular finite population, $\mathcal{F}_N$, has X_N elements equal to 1. The conditional distribution of X_n given $\mathcal{F}_N$ is the hypergeometric distribution and

$$P\{X_n = a \mid \mathcal{F}_N\} = \binom{X_N}{a}\binom{N - X_N}{j-a}\left[\binom{N}{j}\right]^{-1}$$

■■

A fully specified sequence will contain a description of the structure of the finite populations and of the sampling probability rules. For example, it might be assumed that the finite population is composed of N *iid* random variables with properties (1.3.2) and (1.3.3), and that the samples are simple nonreplacement samples of size n_N selected from the N population elements. In that situation, a simple random sample of size n_N selected from the finite universe is a set of *iid* random variables with common distribution function $F_y(y)$. A proof, due to F. Jay Breidt, is given in Theorem 1.3.1.

Theorem 1.3.1. Suppose that $y_1,\ y_2,\ \ldots,\ y_N$ are *iid* with distribution function $F(y)$ and corresponding characteristic function $\varphi(t) = E\{e^{ity}\}$. Let $\mathbf{d} = (I_1,\ I_2,\ \ldots,\ I_N)'$ be a random vector with each component supported on $\{0,\ 1\}$. Assume that $\mathbf{d}$ is independent of $(y_1,\ y_2,\ \ldots,\ y_N)'$. Let

$U = \{1, 2, \ldots, N\}$ and define $A = \{k \in U : I_k = 1\}$. If A is nonempty, the random variables $(y_k, k \in A) \mid \mathbf{d}$ are iid with characteristic function $\varphi(t)$.

Proof. Let $(t_1, t_2, \ldots, t_N)'$ be an element of N-dimensional Euclidean space. Then, given $\mathbf{d}$, the joint characteristic function of $(y_k, k \in A)$ is

$$\begin{aligned} E\left\{exp\left(\mathrm{i}\sum_{k\in A} t_k y_k\right) \middle| \mathbf{d}\right\} &= E\left\{exp\left(\mathrm{i}\sum_{k\in U} t_k I_k y_k\right) \middle| \mathbf{d}\right\} \\ &= E\left\{\prod_{k\in U} exp(\mathrm{i}t_k I_k y_k) \mid \mathbf{d}\right\} \\ &= \prod_{k\in U} E\{exp(it_k I_k y_k) \mid \mathbf{d}\} \\ &= \prod_{k\in U} \varphi(t_k I_k) \\ &= \prod_{k\in A} \varphi(t_k), \end{aligned} \tag{1.3.8}$$

since $\varphi(0) = 1$. The result follows because (1.3.8) is the characteristic function of $n = \Sigma_{k\in U} I_k$ iid random variables with distribution function $F(y)$. ■

The crucial assumption of the theorem is that the probability rule defining membership in the sample, the probability function for $\mathbf{d}$, is independent of $(y_1, y_2, \ldots, y_N)$. It then follows that given $\mathbf{d}$ with component support on $\{0, 1\}$, the sets $\{y_k, k \in A\}$ and $\{y_k, k \notin A\}$ are sets of n and $N - n$ iid random variables with distribution function $F_y(y)$. Furthermore, the two sets are independent. The conditional distribution of the two sets is the same for all $\mathbf{d}$ with the same sample size, where the sample size is

$$n = \sum_{k=1}^{N} I_k.$$

Thus, for fixed-sample-size nonreplacement designs and iid random variables, the unconditional distribution over all samples is the same as the conditional distribution for a particular sample set of indices.

Example 1.3.2. As a second example of a sequence of populations and samples, let $\mathcal{F}_N = (y_1, y_2, \ldots, y_N)$ be the first N elements in a sequence of independent random variables selected from a normal distribution with mean

μ_y and variance σ_y^2. Let n_N be the largest integer less than or equal to fN, where f is a fixed number in (0, 1). Assume that a simple random sample of size n_N is selected from $\mathcal{F}_N$, let A_N be the set of indices of the sample selected, and let A_N^c be the set of indexes of the $N - n_N$ elements not in A_N. The n_N sample elements are $NI(\mu_y, \sigma_y^2)$ random variables, independent of the $N - n_N$ nonsample $NI(\mu_y, \sigma_y^2)$ random variables. It follows that

$$\bar{y}_n \sim N(\mu_y, n_N^{-1}\sigma_y^2),$$

$$\bar{y}_{N-n} \sim N\{\mu_y, (N - n_N)^{-1}\sigma_y^2\},$$

$$\bar{y}_n - \bar{y}_N \sim N\{0, n_N^{-1}(1 - f_N)\sigma_y^2\},$$

and

$$\left[(1 - f_N)n_N^{-1}s_{y,n}^2\right]^{-1/2}(\bar{y}_n - \bar{y}_N) \sim t_{n-1},$$

where $f_N = N^{-1}n_N$,

$$s_{y,n}^2 = (n_N - 1)^{-1}\sum_{i \in A_N}(y_i - \bar{y}_n)^2,$$

$$\bar{y}_n = n_N^{-1}\sum_{i \in A_N} y_i,$$

$$\bar{y}_{N-n} = (N - n_N)^{-1}\sum_{i \in A_N^c} y_i,$$

and t_{n-1} is Student's t-distribution with $n_N - 1$ degrees of freedom. ■■

Given a model for the stochastic mechanism generating the finite population, we can consider expectations conditional on properties of the random variables. Most often we are interested in a set of samples with some of the same characteristics as those observed in the current sample.

Example 1.3.3. Let $\mathcal{F}_N = [(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)]$ be the first N elements in a sequence of independent random variables from a bivariate normal distribution,

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim NI\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}\right].$$

Let a sequence of simple random samples be selected as described in Example 1.3.2. Let $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$, let $\mathbf{z}_i = (x_i, y_i)$, and let

$$\begin{pmatrix} s_{x,n}^2 & s_{xy,n} \\ s_{xy,n} & s_{y,n}^2 \end{pmatrix} = (n - 1)^{-1}\sum_{i \in A_N}(\mathbf{z}_i - \bar{\mathbf{z}}_n)'(\mathbf{z}_i - \bar{\mathbf{z}}_n),$$

where $\bar{\mathbf{z}}_n$ is the simple sample mean. The least squares regression coefficient for the regression of y on x is

$$\hat{\beta}_n = s_{x,n}^{-2} s_{xy,n}.$$

By Theorem 1.3.1, the sample is a realization of iid normal vectors. It follows that under the model, we have the conditional mean and variance,

$$E\{\hat{\beta}_n \mid \mathbf{x}_n\} = \beta$$

and

$$V\{\hat{\beta}_n \mid \mathbf{x}_n\} = \left[\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right]^{-1} \sigma_e^2,$$

where $\beta = \sigma_x^{-2}\sigma_{xy}$ and $\sigma_e^2 = \sigma_y^2 - \beta\sigma_{xy}$. ■■

In describing rates of convergence for real-valued sequences and for sequences of random variables, a notation for order is useful. We use the conventions given by Fuller (1996, Chapter 5). See Appendix 1A.

Example 1.3.4. In previous examples we have considered sequences of finite populations generated by a random mechanism. To study sampling properties for a sequence of finite populations generated from a fixed sequence, let $\{y_i\}$ be a sequence of real numbers and assume that

$$\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}(y_i,\ y_i^2) = (\theta_1,\ \theta_2),$$

where $(\theta_1,\ \theta_2)$ are finite and $\theta_2 - \theta_1^2 > 0$. Define a sequence of finite populations $\{\mathcal{F}_N\}$, where the Nth finite population is composed of the first N values of the sequence $\{y_i\}$. Let a simple random sample of size $n_N = [fN]$ be selected from the Nth finite population, where $0 < f < 1$ and $[fN]$ is the largest integer less than or equal to fN. By the results of Section 1.2,

$$V\{\bar{y}_n - \bar{y}_N \mid \mathcal{F}_N\} = (1 - f_N)n_N^{-1}S_{y,N}^2,$$

where $f_N = N^{-1}n_N$. By assumption,

$$\lim_{N\to\infty} S_{y,N}^2 = \theta_2 - \theta_1^2$$

is a finite positive number. It follows that we can write

$$V\{\bar{y}_n - \bar{y}_N \mid \mathcal{F}_N\} = O(n_N^{-1})$$

and

$$\bar{y}_n - \bar{y}_N \mid \mathcal{F}_N \;=\; O_p(n_N^{-1/2}),$$

where $\{(\bar{y}_n - \bar{y}_N) \mid \mathcal{F}_N\}$ denotes the sequence of $\bar{y}_n - \bar{y}_N$ calculated from the sequence of samples selected from the sequence $\{\mathcal{F}_N\}$. Because the sampling fraction is fixed, n_N and N are of the same order. For a sequence of finite populations that is generated from a sequence of fixed numbers such as this example, the notational reference to $\mathcal{F}_N$ is not required because the random variation comes only from the design. In complex situations the notation serves to identify properties derived from the sampling design. Even in situations where not required, we often employ the notation. ■■

Once the sequence of populations, sample designs, and estimators is specified, the properties of the sequence of estimators can be obtained. The unconditional properties of the estimator, the properties conditional on the particular finite population, and the properties conditional on some attributes of the particular sample are all of interest. Because of the central importance of the sampling design, it is common in the survey sampling literature to use the term *design consistent* for a procedure that is consistent conditional on the particular sequence of finite populations. The sequence of populations can be composed of fixed numbers, as in Example 1.3.4, or can be a sequence of random variables, as in Example 1.3.2. For a sequence of random variables, the property is assumed to hold almost surely (a.s.); that is, the property holds for all sequences except for a set of measure zero.

Definition 1.3.1. Given a sequence of finite populations $\{\mathcal{F}_N\}$ and an associated sequence of sample designs, the estimator $\hat{\theta}$ is *design consistent* for the finite population parameter θ_N if for every $\epsilon > 0$,

$$\lim_{N\to\infty} P\{|\,\hat{\theta} - \theta_N\,| > \epsilon \mid \mathcal{F}_N\} \;=\; 0 \quad \text{a.s.}, \tag{1.3.9}$$

where the notation indicates that for the sequence of finite populations, the probability is that determined by the sample design.

Observe that $\bar{y}_n$ of Example 1.3.4 is design consistent for $\bar{y}_N$ because $V\{\bar{y}_n - \bar{y}_N \mid \mathcal{F}_N\}$ is $O(n_N^{-1})$.

Example 1.3.5. For the sequence of populations and samples of Example 1.3.2,

$$V\{\bar{y}_n - \bar{y}_N \mid \mathcal{F}_N\} = (1 - f_N)n_N^{-1}S_{y,N}^2,$$

where $f_N = N^{-1}n_N$ and

$$S^2_{y,N} = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2.$$

The sequence of populations is created from a sequence of $NI(\mu, \sigma_y^2)$ random variables. Therefore,

$$\lim_{N\to\infty} S^2_{y,N} = \sigma_y^2 \quad \text{a.s.}$$

It follows that

$$V\{\bar{y}_n - \bar{y}_N \mid \mathcal{F}_N\} = O_p(n_N^{-1}) \quad \text{a.s.},$$

$$(\bar{y}_n - \bar{y}_N) \mid \mathcal{F}_N = O_p(n_N^{-1/2}) \quad \text{a.s.}$$

and hence $\bar{y}_n$ is design consistent for $\bar{y}_N$. ■■

1.3.2 Central limit theorems

Central limit theorems are critical to our ability to make probability statements on the basis of sample statistics. Our first results are for a stratified finite population, where the strata are composed of realizations of *iid* random variables. Under mild conditions, the properly standardized stratified mean converges to a normal random variable. In the theorem statement, $N(0, \sigma^2)$ denotes the normal distribution with mean zero and variance σ^2, and the symbol $\xrightarrow{\mathcal{L}}$ is used to denote convergence in distribution (convergence in law).

Theorem 1.3.2. Let $\{\mathcal{F}_N\}$, where $\mathcal{F}_N = \{y_{hiN}\}$, $h = 1, 2, \ldots, H_N$; $i = 1, 2, \ldots, N_{hN}$, be a sequence of finite populations composed of H_N strata, where the y_{hiN} in stratum h are a sample of $iid(\mu_h, \sigma_h^2)$ random variables with bounded $2+\delta$, $\delta > 0$, moments. Let the sample for the Nth population be a simple random stratified sample with $n_{hN} \geq 1$ for all h, where $\{n_{hN}\}$ is a fixed sequence. Let

$$\begin{aligned}
A_N &= \{hi \in U_N : I_{hiN} = 1\}, \\
n_N &= \sum_{h=1}^{H_N} n_{hN} = \sum_{h=1}^{H_N}\sum_{i=1}^{N_{hN}} I_{hiN}, \\
\bar{y}_N &= N^{-1}\sum_{h=1}^{H_N}\sum_{i=1}^{N_{hN}} y_{hiN} = N^{-1}\sum_{hi\in U_N} y_{hiN}, \\
\hat{\theta}_n &= \sum_{h=1}^{H_N} N^{-1}N_{hN}\bar{y}_{hn},
\end{aligned}$$

and

$$\bar{y}_{hn} = n_{hN}^{-1} \sum_{i=1}^{n_{hN}} y_{hiN},$$

where U_N is the set of indices hi for population N and I_{hiN} is the indicator for sample membership. Assume that

$$\lim_{N\to\infty} \sup_{1\le h\le H_N} \frac{n_{hN}^{-2}\left(N_{hN}-n_{hN}\right)^2+1}{\sum_{g=1}^{H_N} N_{gN}\left(N_{gN}-n_{gN}\right) n_{gN}^{-1}\sigma_g^2} = 0. \tag{1.3.10}$$

Then

$$[V\{\hat{\theta}_n - \bar{y}_N\}]^{-1/2}(\hat{\theta}_n - \bar{y}_N) \xrightarrow{\mathcal{L}} N(0,1), \tag{1.3.11}$$

where

$$V\{\hat{\theta}_n - \bar{y}_N\} = \sum_{h=1}^{H_N} N^{-2} N_{hN}(N_{hN}-n_{hN}) n_{hN}^{-1}\sigma_h^2.$$

Furthermore, if the y_{hiN} have bounded fourth moments, if $n_{hN} \ge 2$ for all h, and

$$\sum_{h=1}^{H_N} \lambda_{hN}^2 n_{hN}^{-1} = o\left(\left(\sum_{h=1}^{H_N} \lambda_{hN}\right)^2\right), \tag{1.3.12}$$

then

$$[\hat{V}\{\hat{\theta}_n - \bar{y}_N\}]^{-1/2}(\hat{\theta}_n - \bar{y}_N) \xrightarrow{\mathcal{L}} N(0,1), \tag{1.3.13}$$

where $\lambda_{hN} = N^{-2} N_{hN}\left(N_{hN}-n_{hN}\right) n_{hN}^{-1}$,

$$\hat{V}\{\hat{\theta}_n - \bar{y}_N\} = \sum_{h=1}^{H_N} N^{-2} N_{hN}(N_{hN}-n_{hN}) n_{hN}^{-1} s_h^2,$$

and

$$s_h^2 = (n_{hN}-1)^{-1} \sum_{j=1}^{N_{hN}} (y_{hjN} - \bar{y}_{hn})^2.$$

Proof. For each N, the design is a fixed-size design and by Theorem 1.3.1 the sample in each stratum is a set of iid random variables. Therefore, the

stratified estimator is a weighted average of independent random variables and we write

$$\begin{aligned}\hat{\theta}_n - \bar{y}_N &= N^{-1} \sum_{hi \in U_N} \left(N_{hN} n_{hN}^{-1} I_{hiN} - 1\right) y_{hiN} \\ &=: N^{-1} \sum_{hi \in U_N} c_{hN} y_{hiN},\end{aligned}$$

where

$$\begin{aligned}c_{hN} &= N^{-1}\left(N_{hN} - n_{hN}\right) n_{nN}^{-1} \qquad \text{if } hi \in A_N \\ &= -N^{-1} \qquad \text{if } hi \notin A_N.\end{aligned}$$

Because the random variables are identically distributed and the n_{hN} are fixed, we can treat the c_{hN} as fixed.

The Lindeberg criterion is

$$\begin{aligned}&\lim_{N\to\infty} V_N^{-1} \sum_{hi \in U_N} c_{hN}^2 \int_{R_{hN}} (y - \mu_h)^2 dF_{hy}(y) \\ &\leq \lim_{N\to\infty} V_N^{-1} \sum_{hi \in U_N} c_{hN}^2 \int_{R_{0N}} (y - \mu_h)^2 dF_{hy}(y) \\ &\leq \lim_{N\to\infty} V_N^{-1} \sum_{hi \in U_N} c_{hN}^2 \epsilon^{\delta} B_N^{\delta} \int_{R_{0N}} |\, y - \mu_h \,|^{2+\delta}\, dF_{hy}(y) \\ &\leq \lim_{N\to\infty} V_N^{-1} \sum_{hi \in U_N} c_{hN}^2 \epsilon^{\delta} B_N^{\delta} E\{|\, y_{hi} - \mu_h \,|^{2+\delta}\}, \qquad (1.3.14)\end{aligned}$$

where $V_N = V\{\hat{\theta}_n - \bar{y}_N\}$,

$$R_{hN} = \{y : |\, y - \mu_h \,| \geq \epsilon V_N^{1/2} \,|\, c_{hN} \,|^{-1}\},$$

$$R_{0N} = \{y : |\, y - \mu_h \,| \geq \epsilon B_N^{-1}\},$$

and

$$B_N = V_N^{-1/2} \sup_{1 \leq h \leq H_N} |\, c_{hN} \,|\,.$$

By assumption (1.3.10) B_N is converging to zero and (1.3.14) converges to zero because the $2 + \delta$ moments are bounded. Thus, the first result is proven.

If the y_{hiN} have fourth moments bounded by, say, M_4, the variance of the estimated variance is

$$\begin{aligned} V\left\{\sum_{h=1}^{H_N}\lambda_{hn}s_h^2\right\} &= \sum_{h=1}^{H_N}\lambda_{hN}^2(n_{hN}-1)^{-2}V\left\{\sum_{i\in A_h}y_{hiN}^2 - n_{hN}\bar{y}_{hn}^2\right\} \\ &\leq \sum_{h=1}^{H_N}\lambda_{hN}^2(n_{hN}-1)^{-2}(n_{hN}M_4+M_5) \\ &= o([V\{\hat{\theta}_n-\bar{y}_N\}]^2) \end{aligned}$$

by assumption (1.3.12), where M_5 is the bound on $V\{n_{hN}\bar{y}_h^2\} - 2C\{n_{hN}\bar{y}_h^2, \sum y_{hiN}^2\}$. See Exercise 49. Therefore,

$$[V\{\hat{\theta}_n-\bar{y}_N\}]^{-1}\hat{V}\{\hat{\theta}_n-\bar{y}_N\} \xrightarrow{P} 1$$

and result (1.3.13) is proven. ■

By Theorem 1.3.2, the stratified mean is approximately normally distributed for a large number of small strata or for a small number of large strata.

It is important to note that in Theorem 1.3.2, as in Theorem 1.3.1, results are obtained by averaging over all possible finite populations under the assumption that the design vector $\mathbf{d}$ is independent of $(y_1, y_2, \ldots, y_N)$. The independence assumption is reasonable for stratified samples because selection in each stratum is simple random sampling. Simple random sampling is a special case of stratified sampling and hence the sample mean of a simple random sample is normally distributed in the limit.

Corollary 1.3.2.1. Let $\{\mathcal{F}_N\}$, where $\mathcal{F}_N = (y_{1N}, y_{2N}, \ldots, y_{NN})$, be a sequence of finite populations in which the $y_{iN}, i = 1, 2, \ldots, N$, are realizations of independent (μ, σ^2) random variables with bounded $2+\delta, \delta > 0$, moments. Let A_N be a simple random nonreplacement sample of size n_N selected from the Nth population. Assume that

$$\lim_{N\to\infty} n_N = \infty$$

and

$$\lim_{N\to\infty} N - n_N = \infty.$$

Let $\bar{y}_n$, $\bar{y}_N$, $S_{y,N}^2$, and $s_{y,n}^2$ be as defined in (1.2.35), (1.2.22), (1.2.36), and (1.2.37), respectively. Then

$$\left[N^{-1}(N-n_N)n_N^{-1}S_{y,N}^2\right]^{-1/2}(\bar{y}_n-\bar{y}_N) \xrightarrow{\mathcal{L}} N(0,1) \tag{1.3.15}$$

and

$$\left[N^{-1}(N-n_N)n_N^{-1}s_{y,n}^2\right]^{-1/2}(\bar{y}_n-\bar{y}_N)\xrightarrow{\mathcal{L}}N(0,1). \tag{1.3.16}$$

Proof. Because both n_N and $N-n_N$ increase without bound as N increases, (1.3.10) is satisfied for $H=1$, and result (1.3.15) follows. By the assumption that $E\{|\ y\ |^{2+\delta}\}$ is bounded,

$$\lim_{N\to\infty} s_{y,n}^2=\sigma^2 \quad \text{a.s.}$$

See Hall and Heyde (1980, p. 36). Result (1.3.16) then follows. ■

Result (1.3.13) permits one to use the estimated variance to construct confidence intervals that are appropriate in large samples. That is,

$$\lim_{N\to\infty} P\{\bar{y}_n-t_\alpha[\hat{V}\{\bar{y}_n\}]^{0.5}\le\bar{y}_N\le\bar{y}_n+t_\alpha[\hat{V}\{\bar{y}_n\}]^{0.5}\}=\alpha,$$

where the probability that a standard normal random variable exceeds t_α is α and

$$\hat{V}\{\bar{y}_n\}=(1-f_N)n_N^{-1}s_{y,n}^2.$$

In Theorem 1.3.2, the basis for the limiting result is a sequence of all possible samples from all possible finite populations. One can also obtain limiting normality for Poisson sampling from a fixed sequence of finite populations. The result is due to Hájek (1960).

Consider the Poisson sampling design introduced in Section 1.2.2. For such a design, define the vector random variable

$$\mathbf{x}_i = \mathbf{g}_i I_i, \qquad i=1,\ 2,\ \ldots,\ N, \tag{1.3.17}$$

where I_i is the indicator variable with $I_i=1$ if element i is selected and $I_i=0$ otherwise,

$$\mathbf{g}_i=(1,\ \mathbf{y}_i',\ \alpha_N\pi_i^{-1},\ \alpha_N\pi_i^{-1}\mathbf{y}_i')', \qquad i=1,\ 2,\ \ldots,\ N, \tag{1.3.18}$$

π_i is the probability that element i is included in the sample, $\mathbf{y}_i$ is a column vector associated with element i, $\alpha_N=N^{-1}n_{BN}$, and $n_{BN}=E\{n_N\mid N\}$, where $n_N=\Sigma_{i\in A}I_i$. The ratio α_N is required only for normalization purposes in limit operations, and is required only if $N^{-1}n_N$ or $1-N^{-1}n_N$ goes to zero as N increases. For a fixed $\mathbf{g}_i$, the mean of $\mathbf{x}_i$ is $\mathbf{g}_i\pi_i$ and

$$V\{\mathbf{x}_i\mid\mathbf{g}_i\}=\pi_i(1-\pi_i)\,\mathbf{g}_i\,\mathbf{g}_i'.$$

The Horvitz–Thompson estimator, $N^{-1}\hat{\mathbf{T}}_y$, of the population mean of $\mathbf{y}$ is the vector associated with $\alpha_N \pi_i^{-1}\mathbf{y}_i$ in the estimated mean vector,

$$\hat{\boldsymbol{\mu}}_x = n_{BN}^{-1} \sum_{i=1}^{N} \mathbf{x}_i. \tag{1.3.19}$$

Theorem 1.3.3. Let $\mathbf{y}_1, \mathbf{y}_2, \ldots,$ be a sequence of real vectors and let $\pi_1, \pi_2, \ldots,$ be a sequence of probabilities, with $0 < \pi_i < 1$. Let a Poisson sample be selected from $\mathcal{F}_N = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$, and let $\mathbf{g}_i$ be defined by (1.3.18). Assume that

$$\lim_{N\to\infty} n_{BN}^{-1} \sum_{i=1}^{N} \mathbf{g}_i \pi_i = \boldsymbol{\mu}_x, \tag{1.3.20}$$

$$\lim_{N\to\infty} n_{BN}^{-1} \sum_{i=1}^{N} \pi_i(1-\pi_i)\mathbf{g}_i \mathbf{g}_i' = \boldsymbol{\Sigma}_{xx}, \tag{1.3.21}$$

the submatrix of $\boldsymbol{\Sigma}_{xx}$ associated with $(1, \mathbf{y}_i')$ is positive definite, and the submatrix of $\boldsymbol{\Sigma}_{xx}$ associated with $(\alpha_N \pi_i^{-1}, \alpha_N \pi_i^{-1}\mathbf{y}_i')$ is positive definite. Also assume that

$$\lim_{N\to\infty} \sup_{1\le k\le N} \left(\sum_{i=1}^{N} (\boldsymbol{\gamma}' \mathbf{g}_i)^2 \pi_i(1-\pi_i) \right)^{-1} (\boldsymbol{\gamma}'\mathbf{g}_k)^2 = 0 \tag{1.3.22}$$

for every fixed row vector $\boldsymbol{\gamma}'$ such that $\boldsymbol{\gamma}' \boldsymbol{\Sigma}_{xx} \boldsymbol{\gamma} > 0$. Let $\mathbf{x}_i$, $i = 1, 2, \ldots,$ be the independent random variables defined by (1.3.17). Then

$$n_{BN}^{1/2}(\hat{\boldsymbol{\mu}}_x - \boldsymbol{\mu}_{xN}) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \boldsymbol{\Sigma}_{xx}), \tag{1.3.23}$$

where

$$\boldsymbol{\mu}_{xN} = n_{BN}^{-1} \sum_{i=1}^{N} \mathbf{g}_i \pi_i$$

and $\hat{\boldsymbol{\mu}}_x$ is defined in (1.3.19).

If, in addition,

$$\lim_{N\to\infty} n_{BN}^{-1} \sum_{i=1}^{N} \pi_i \mid \mathbf{g}_i \mid^4 = M_g \tag{1.3.24}$$

for some finite M_g, then

$$[\hat{V}\{\hat{\mathbf{T}}_y \mid \mathcal{F}_N\}]^{-1/2}(\hat{\mathbf{T}}_y - \mathbf{T}_y) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}), \tag{1.3.25}$$

where $\hat{\mathbf{T}}_y$ is the Horvitz–Thompson estimator, $\mid \mathbf{g}_i \mid = (\mathbf{g}_i'\mathbf{g}_i)^{1/2}$, and

$$\hat{V}\{\hat{\mathbf{T}}_y \mid \mathcal{F}_N\} = \sum_{i \in A_N} (1 - \pi_i)\pi_i^{-2}\mathbf{y}_i\mathbf{y}_i'.$$

Proof. Let

$$Z_i = \boldsymbol{\gamma}' \mathbf{g}_i(I_i - \pi_i), \tag{1.3.26}$$

where $\boldsymbol{\gamma}$ is an arbitrary real-valued column vector, $\boldsymbol{\gamma} \neq \mathbf{0}$. Then $\{Z_i\}$ is a sequence of independent random variables with zero means and $V\{Z_i\} = \pi_i(1 - \pi_i)\ (\boldsymbol{\gamma}'\ \mathbf{g}_i)^2 =: v_{ii}$. Letting

$$V_N = \sum_{i=1}^{N} \pi_i\ (1 - \pi_i)\ (\boldsymbol{\gamma}'\mathbf{g}_i)^2 = \sum_{i=1}^{N} v_{ii}, \tag{1.3.27}$$

the arguments of the proof of Theorem 1.3.2 can be used to show that

$$V_N^{-1/2} \sum_{i=1}^{N} Z_i \xrightarrow{\mathcal{L}} N(0,\ 1). \tag{1.3.28}$$

Note that all moments exist for the random variables I_i. Multivariate normality follows because $\boldsymbol{\gamma}$ is arbitrary. Now

$$\lim_{N \to \infty} n_{BN}^{-1} \sum_{i=1}^{N} (\boldsymbol{\gamma}'\mathbf{g}_i)^2 \pi_i(1 - \pi_i) = \boldsymbol{\gamma}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\gamma}$$

and we have result (1.3.23).

By assumption (1.3.24), the variance of $\hat{V}\{N^{-1}\boldsymbol{\gamma}_y'\hat{\mathbf{T}}_y \mid \mathcal{F}_N\}$ is

$$\begin{aligned} V\left\{N^{-2} \sum_{i \in A_N} (1 - \pi_i)\pi_i^{-2}(\boldsymbol{\gamma}_y'\mathbf{y}_i)^2\right\} &= N^{-4} \sum_{i=1}^{N} \pi_i(1 - \pi_i)^3\pi_i^{-4}(\boldsymbol{\gamma}_y'\mathbf{y}_i)^4 \\ &= O(N^{-3}) \end{aligned}$$

for any fixed vector $\boldsymbol{\gamma}_y$. Because $V\{\hat{\mathbf{T}}_y\}$ is positive definite,

$$[\hat{V}\{\hat{\mathbf{T}}_y \mid \mathcal{F}_N\}]^{-1/2}[V\{\hat{\mathbf{T}}_y \mid \mathcal{F}_N\}]^{1/2} = \mathbf{I} + O_p(N^{-1/2}) \quad (1.3.29)$$

and result (1.3.25) follows. ■

By Theorem 1.3.3, the limiting distribution of the pivotal statistic for the mean of a Poisson sample is $N(0, 1)$ for any sequence of finite populations satisfying conditions (1.3.20), (1.3.21), and (1.3.22). Condition (1.3.22) can be replaced with conditions on the moments of y and on the probabilities.

Corollary 1.3.3.1. Let the sequence of populations and vectors satisfy (1.3.20) and (1.3.21) of Theorem 1.3.3. Replace assumption (1.3.22) with

$$\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \mid \mathbf{y}_i \mid^{2+\delta} < \infty \quad (1.3.30)$$

for some $\delta > 0$, and assume that

$$K_L < \pi_i < K_U \quad (1.3.31)$$

for all i where K_L and K_U are fixed positive numbers and n_{BN} was defined for (1.3.18). Then the limiting normality of (1.3.23) holds.

Proof. The result follows from the fact that (1.3.30) and (1.3.31) are sufficient for the Lindeberg condition. ■

Hájek (1960) showed that the result for Poisson sampling can be extended to simple random nonreplacement sampling .

Theorem 1.3.4. Let the assumptions of Theorem 1.3.3 hold for a sequence of scalars, $\{y_j\}$, with the exception of the assumption of Poisson sampling. Instead, assume that samples of fixed size $n = \pi N$ are selected by simple random nonreplacement sampling. Then

$$\hat{V}_n^{-1/2}(\bar{y}_n - \bar{y}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(0, 1), \quad (1.3.32)$$

where $\hat{V}_n = N^{-1}n^{-1}(N-n)s_{y,n}^2$, and $s_{y,n}^2$ is defined for (1.2.37).

Proof. The probability that any set of r elements, $1 \leq r \leq n_o$, is included in a Poisson sample of size n_o is

$$\frac{\binom{N-r}{n_o-r}\pi^{n_o}(1-\pi)^{N-n_o}}{\binom{N}{n_o}\pi^{n_o}(1-\pi)^{N-n_o}},$$

which is also the probability for the set of r selected as a simple random sample. Hence, the conditional distribution of the Poisson sample given that n_o elements are selected is that of a simple random nonreplacement sample of size n_o.

Let n_B be the expected sample size of a Poisson sample selected with probability π, where n_B is an integer, and let a realized sample of size n_o be given. We create a simple random sample of size n_B starting with the sample of size n_o. If $n_o > n_B$, a simple random sample of $n_o - n_B$ elements is removed from the original sample. If $n_B > n_o$, a simple random sample of $n_B - n_o$ elements is selected from the $N - n_o$ nonsample elements and added to the original n_o elements. If $n_o > n_B$, the n_B elements form a simple random sample from the n_o, and if $n_o < n_B$, n_o is a simple random sample from n_B.

Consider, for $n_o > n_B$, the difference

$$\bar{y}_o - \bar{y}_B = \frac{\sum_{i\in A_o} y_i}{n_o} - \frac{\sum_{i\in A_o} y_i - \sum_{i\in A_k} y_i}{n_B},$$

where $\bar{y}_o$ is the mean of the original Poisson sample, $\bar{y}_B = \bar{y}_{SRS} = \bar{y}_n$ is the mean of the created simple random sample, A_o is the set of indices in the original Poisson sample, and A_k represents the indices of the $k = n_o - n_B$ elements removed from the original Poisson sample. Because the n_B elements are selected from the n_o elements, $E\{\bar{y}_B \mid (n_o, A_o)\} = \bar{y}_o$ and

$$V\{\bar{y}_o - \bar{y}_B \mid (n_o, A_o)\} = (n_B^{-1} - n_o^{-1})s_{yo}^2,$$

where

$$\begin{aligned} s_{yo}^2 &= (n_o - 1)^{-1} \sum_{i\in A_o} (y_i - \bar{y}_o)^2, \\ \bar{y}_B &= n_B^{-1} \sum_{i\in A_B} y_i, \\ \bar{y}_o &= n_o^{-1} \sum_{i\in A_o} y_i, \end{aligned}$$

and A_B is the set of indices for the n_B elements. Furthermore,

$$\begin{aligned} V\{\bar{y}_o - \bar{y}_B \mid n_o\} &= E\{E[(\bar{y}_o - \bar{y}_B)^2 \mid (n_o, A_o)] \mid n_o\} \\ &= (n_B^{-1} - n_o^{-1})S_{y,N}^2, \end{aligned}$$

where $S_{y,N}^2$ is the finite population variance.

If $0 < n_o \leq n_B$, $E\{\bar{y}_o \mid (n_o, A_B)\} = \bar{y}_B$,

$$V\{\bar{y}_o - \bar{y}_B \mid (n_o, A_B)\} = (n_o^{-1} - n_B^{-1})s_{y,B}^2$$

and

$$V\{\bar{y}_o - \bar{y}_B \mid n_o\} = (n_o^{-1} - n_B^{-1})S_{y,N}^2,$$

where

$$s_{y,B}^2 = (n_B - 1)^{-1} \sum_{i \in A_B} (y_i - \bar{y}_B)^2.$$

Then, for $n_o > 0$,

$$V\{\bar{y}_o - \bar{y}_B \mid n_o\} = \mid n_o^{-1} - n_B^{-1} \mid S_{y,N}^2.$$

We note that n_o satisfies

$$E\{(n_B^{-1} n_o - 1)^{2r}\} = O(n_B^{-r})$$

for positive integer r because $N^{-1} n_o$ is the mean of N Poisson random variables. Now $n_o^{-1} n_B$ is bounded by n_B for $n_o > 0$ and by $K_1^{-1} n_B$ for $n_o > K_1$. It follows from Theorems 5.4.4 and 5.4.3 of Fuller (1996) that

$$E\{(n_o^{-1} - n_B^{-1})^2 \mid n_o > 0\} = O(n_B^{-3}).$$

See Exercise 1.34.

To evaluate $E\{(\bar{y}_o - \bar{y}_B)^2\}$, we define $\bar{y}_o - \bar{y}_B = \bar{y}_B$ when $n_o = 0$, and write

$$E\{(\bar{y}_o - \bar{y}_B)^2\} = E\{(\bar{y}_o - \bar{y}_B)^2 \mid n_o > 0\} P\{n_o > 0\} + \bar{y}_B^2 P\{n_o = 0\}.$$

Because $P\{n_o = 0\}$ goes to zero exponentially as $n_B \to \infty$,

$$E\left\{(\bar{y}_B - \bar{y}_o)^2\right\} = O(n_B^{-3/2})$$

and the limiting distribution of $n_B^{1/2}(\bar{y}_B - \bar{y}_N)$ is the same as that of $n_B^{1/2}(\bar{y}_o - \bar{y}_N)$. By Theorem 1.3.3, the limiting distribution of $n_B^{1/2}(\bar{y}_o - \bar{y}_N)$ is normal. By assumption (1.3.24), $s_{y,n}^2 - S_{y,N}^2$ converges to zero in probability, and result (1.3.32) is proven. ■

Theorem 1.3.4 is for simple random samples, but the result extends immediately to a sequence of stratified samples with a fixed number of strata.

Corollary 1.3.4.1. Let $\{\mathcal{F}_N\}$ be a sequence of populations, where the Nth population is composed of H strata with $\mathcal{F}_{h,N} = \{y_{h1}, y_{h2}, \ldots, y_{h,N_{hN}}\}$, $h = 1, 2, \ldots, H$. Assume that $\{y_{hi}\}$, $h = 1, 2, \ldots, H$, are sequences of real numbers satisfying

$$\lim_{N \to \infty} N_{hN}^{-1} \sum_{i \in U_{hN}} [y_{hi}, (y_{hi} - \bar{y}_{hN})^2, y_{hi}^4] = (M_{1h}, S_h^2, M_{4h}),$$

where the M_{4h} are finite and the S_h^2 are positive. Then

$$[\hat{V}\{(\bar{y}_{st} - \bar{y}_N) \mid \mathcal{F}_N\}]^{-1/2}(\bar{y}_{st} - \bar{y}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,\ 1),$$

where $\bar{y}_{st}$ is the stratified mean and $\hat{V}\{(\bar{y}_{st} - \bar{y}_N) \mid \mathcal{F}_N\}$ is the usual stratified estimator defined in (1.2.56).

Proof. Omitted. ∎

In proving Theorems 1.3.3 and 1.3.4, we assumed the elements of the finite population to be fixed and obtained results based on the sequence of fixed populations. In Theorem 1.3.2, the sequence of finite populations was created as samples from an infinite population, and the results were for averages over all possible samples from all possible finite populations. It is also useful to have conditional properties for a sequence of finite populations created as samples from an infinite population. Using the strong law of large numbers, it is shown in Theorem 1.3.5 that the central limit theorem holds conditionally, except for a set of probability zero. The sequence $\{\pi_i\}$ in the theorem can be fixed or random.

Theorem 1.3.5. Consider a sequence of populations, $\{\mathcal{F}_N\}$, where the Nth population is the set $(y_1,\ y_2,\ \ldots,\ y_N)$ and $\{y_i\}$ is a sequence of independent $(\mu,\ \sigma_i^2)$ random variables with bounded $4 + \delta,\ \delta > 0$, moments. Let a Poisson sample be selected from the Nth finite population with probabilities π_i, where the $Nn_{BN}^{-1}\pi_i$ are bounded as described in (1.3.31). If the π_i are random, $(\pi_i,\ y_i)$ is independent of $(\pi_j,\ y_j)$ for $i \neq j$. Assume that

$$\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\sigma_i^2 = V_{11} \tag{1.3.33}$$

and

$$\lim_{N\to\infty} n_{BN}N^{-2}\sum_{i=1}^{N}\pi_i^{-1}(1,\ y_i)'(1,\ y_i) = \mathbf{\Sigma}_{22} \text{ a.s.,} \tag{1.3.34}$$

where $E\{n_N \mid \mathcal{F}_N\} = n_{BN}$. Assume that V_{11} is positive and that $\mathbf{\Sigma}_{22}$ is positive definite. Then

$$[\hat{V}\{\hat{T}_y \mid \mathcal{F}_N\}]^{-1/2}(\hat{T}_y - T_{yN}) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,\ 1) \text{ a.s.,} \tag{1.3.35}$$

where T_{yN} is the population total for the Nth population,

$$\hat{T}_y = \sum_{i\in A_N}\pi_i^{-1}y_i,$$

and

$$\hat{V}\{\hat{T}_y \mid \mathcal{F}_N\} = \sum_{i \in A_N} (1 - \pi_i)\pi_i^{-2} y_i^2.$$

Proof. By the $2 + \delta$ moments of the superpopulation,

$$\lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} (y_i,\ y_i^2) = (\mu,\ \mu^2 + V_{11}) \quad \text{a.s.}$$

and

$$\lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} \mid y_i \mid^{2+0.5\delta} < \infty \quad \text{a.s.}$$

Therefore, conditions (1.3.20), (1.3.21), and (1.3.22) are satisfied almost surely and

$$[V\{\hat{T}_y \mid \mathcal{F}_N\}]^{-1/2} (\hat{T}_y - T_{yN}) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,\ 1) \quad \text{a.s.}, \tag{1.3.36}$$

where

$$V\{\hat{T}_y \mid \mathcal{F}_N\} = \sum_{i=1}^{N} (1 - \pi_i)\pi_i^{-1} y_i^2,$$

by Corollary 1.3.3.1.

Now $n_{BN} N^{-1} (1 - \pi_i)\pi_i^{-1}$ is bounded by, say, K_c, and

$$V\left\{ \sum_{i \in A} \left[n_{BN} N^{-1} (1 - \pi_i)\pi_i^{-1} \right] \pi_i^{-1} y_i^2 \mid \mathcal{F}_N \right\} \leq \sum_{i=1}^{N} K_c^2 (1 - \pi_i)\pi_i^{-1} y_i^4.$$

Therefore, by the $4 + \delta$ moments of y_i,

$$\lim_{N \to \infty} n_{BN} N^{-1} \sum_{i=1}^{N} K_c^2 (1 - \pi_i)\pi_i^{-1} y_i^4$$

is a well-defined finite number, almost surely. It follows that

$$n_{BN}N^{-2}[\hat{V}\{\hat{T}_y \mid \mathcal{F}_N\} - V\{\hat{T}_y \mid \mathcal{F}_N\}] \mid \mathcal{F}_N = O_p(n_{BN}^{-1/2}) \text{ a.s.} \tag{1.3.37}$$

and result (1.3.35) follows. ■

Theorem 1.3.5 is for Poisson sampling but the result holds for a sequence of stratified samples with a fixed number of strata, by Corollary 1.3.4.1.

Corollary 1.3.5.1. Let $\{\mathcal{F}_N\}$ be a sequence of populations, where the Nth population is composed of H strata with $\mathcal{F}_{h,N} = \{y_{h1}, y_{h2}, \ldots, y_{h,N_{hN}}\}$, $h = 1, 2, \ldots, H$. Assume that the $\{y_{hi}\}$, $h = 1, 2, \ldots, H$, are sequences of independent (μ_h, σ_h^2) random variables with bounded $4 + \delta$, $\delta > 0$ moments. Let a sequence of stratified samples be selected, where $n_{h,N} \geq n_{h,N-1}$, $\lim_{N\to\infty} n_{h,N} = \infty$,

$$\lim_{N\to\infty} N_{h,N}^{-1} n_{h,N} = f_{h,\infty},$$

and

$$\lim_{N\to\infty} N^{-1} N_{h,N} = W_h$$

for $h = 1, 2, \ldots, H$. Assume that $0 \leq f_{h,\infty} < 1$ and $W_h > 0$ for $h = 1, 2, \ldots, H$. Then

$$[\hat{V}\{\hat{T}_y \mid \mathcal{F}_N\}]^{-1/2}(\hat{T}_y - T_{yN}) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(0, 1) \text{ a.s.},$$

where

$$\hat{V}\{\hat{T}_y \mid \mathcal{F}_N\} = \sum_{h=1}^{H} N_h^2(n_h^{-1} - N_h^{-1})s_h^2.$$

Proof. The conditions of Corollary 1.3.3.1 hold almost surely and the result follows by Corollary 1.3.4.1. ■

To extend the results of Theorem 1.3.5 to estimation of parameters of the superpopulation, we require the following theorem, adapted from Schenker and Welsh (1988).

Theorem 1.3.6. Let $\{\mathcal{F}_N\}$ be a sequence of finite populations, let θ_N be a function of the elements of $\mathcal{F}_N$, and let a sequence of samples be selected from $\{\mathcal{F}_N\}$ by a design such that

$$\theta_N - \theta_N^{\circ} \xrightarrow{\mathcal{L}} N(0, V_{11}), \tag{1.3.38}$$

and

$$(\hat{\theta} - \theta_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,\ V_{22}) \quad \text{a.s.}, \tag{1.3.39}$$

for a fixed sequence $\{\theta_N^\circ\}$ and an estimator, $\hat{\theta}$, where $V_{11} + V_{22} > 0$. Then

$$(V_{11} + V_{22})^{-1/2}(\hat{\theta} - \theta_N^\circ) \xrightarrow{\mathcal{L}} N(0,\ 1). \tag{1.3.40}$$

Proof. Let $\Phi_{Vj}(\cdot)$ denote the normal cumulative distribution function with mean zero and variance V_{jj}, $j = 1,\ 2$, and let $\Phi_{V3}(\cdot) = \Phi_{V1}(\cdot) * \Phi_{V2}(\cdot)$ denote the normal cumulative distribution function with mean zero and variance $V_{11} + V_{22} = V_{33}$, where $\Phi_{V1}(\cdot) * \Phi_{V2}(\cdot)$ is the convolution of $\Phi_{V1}(\cdot)$ and $\Phi_{V2}(\cdot)$. Consider

$$\mid P\{(\hat{\theta} - \theta_N^\circ) \le t\} - \Phi_{V3}(t) \mid \ = \ \mid E[P\{(\hat{\theta} - \theta_N^\circ) \le t\} \mid \mathcal{F}_N] - \Phi_{V3}(t) \mid .$$

Letting $s = t - (\theta_N - \theta_N^\circ)$ yields

$$\begin{aligned}
&\mid P\{(\hat{\theta} - \theta_N^\circ) \le t\} - \Phi_{V3}(t) \mid \\
&\qquad \le \ \mid E[P\{(\hat{\theta} - \theta_N) \le s \mid \mathcal{F}_N\}] - E\{\Phi_{V2}(s) \mid \mathcal{F}_N\} \mid \\
&\qquad\quad + \mid E\{\Phi_{V2}(s) \mid \mathcal{F}_N\} - \Phi_{V3}(t) \mid \\
&\qquad \le \ E[\ \mid P\{(\hat{\theta} - \theta_N) \le s \mid \mathcal{F}_N\} - \Phi_{V2}(s) \mid ,\ \mid \mathcal{F}_N] \\
&\qquad\quad + \mid E\{\Phi_{V2}(s) \mid \mathcal{F}_N\} - \Phi_{V3}(t) \mid .
\end{aligned} \tag{1.3.41}$$

Because $\mid P\{(\hat{\theta} - \theta_N) \le s \mid \mathcal{F}_N\} - \Phi_{V2}(s) \mid$ is bounded for all $s \in \mathcal{R}$,

$$\begin{aligned}
&\lim_{N\to\infty} \mid P\{(\hat{\theta} - \theta_N) \le s \mid \mathcal{F}_N\} - \Phi_{V2}(s) \mid \mathcal{F}_N \mid \\
&\qquad \le E\{\lim_{N\to\infty} (\sup_{s\in\mathcal{R}} \mid P\{(\hat{\theta} - \theta_N) \le s \mid \mathcal{F}_N\} - \Phi_{V2}(s) \mid ,\ \mid \mathcal{F}_N)\}
\end{aligned} \tag{1.3.42}$$

by the dominated convergence theorem. By Lemma 3.2 of R. R. Rao (1962), assumption (1.3.39) implies that

$$\lim_{N\to\infty} \{\sup_{s\in\mathcal{R}} \mid P\{(\hat{\theta} - \theta_N) \le s \mid \mathcal{F}_N\} - \Phi_{V2}(s) \mid ,\ \mid \mathcal{F}_N\} \ = \ 0 \quad \text{a.s.} \tag{1.3.43}$$

and the expectation in (1.3.42) is zero.

Now

$$\lim_{N\to\infty} E\{\Phi_{V2}(s) \mid \mathcal{F}_N\} = E\left\{\lim_{N\to\infty} \Phi_{V2}[t-(\theta_N-\theta_N^\circ)]\ ,\ \mid \mathcal{F}_N\right\}$$

$$= \Phi_{V2}(t) * \Phi_{V1}(t) = \Phi_{V3}(t) \qquad (1.3.44)$$

by (1.3.38) and the dominated convergence theorem. It follows from (1.3.43) and (1.3.44) that (1.3.41) converges to zero as $N \to \infty$. ■

The V_{11} of (1.3.38) or the V_{22} of (1.3.39) can be zero, but the sum $V_{11}+V_{22}$ is never zero. For example, let $\bar{y}_n$ be the mean of a simple random sample from a finite population that is a set of $iid(\mu, \sigma^2)$ random variables, let $\hat{\theta}_n - \theta_N = N^{1/2}(\bar{y}_n - \bar{y}_N)$, and let $\theta_N - \theta_N^\circ = N^{1/2}(\bar{y}_N - \mu)$. Then if $N - n \to 0$ as N increases, $V_{22} = 0$. Conversely, if $N^{-1}n \to 0$, the limiting variance of $n^{1/2}(\bar{y}_n - \bar{y}_N) = \sigma^2$ and the limiting variance of $n^{1/2}(\bar{y}_N - \mu) = 0$. If $\lim N^{-1}n = f$, for $0 < f < 1$, both V_{11} and V_{22} are positive. These theoretical results have a commonsense interpretation. If the sample is a very small fraction of the finite population, the fact that there is the intermediate step of generating a large finite population is of little importance. Conversely, if we have a very large sampling rate, say a census, we still have variability in the estimator of the superpopulation parameter. See Deming and Stephan (1941) on the use of a census in this context.

Using Theorems 1.3.6 and 1.3.5, one can prove that the limiting distribution of the standardized $\bar{y}_{HT} - \mu$ is normal.

Corollary 1.3.6.1. Let the assumptions of Theorem 1.3.5 hold and assume that

$$\lim_{N\to\infty} N^{-1}n_{BN} = f_\infty,$$

where $0 \le f_\infty \le 1$. Then

$$\left[\hat{V}\{\bar{y}_{HT} - \mu\}\right]^{-1/2}(\bar{y}_{HT} - \mu) \xrightarrow{\mathcal{L}} N(0,\ 1), \qquad (1.3.45)$$

where $\bar{y}_{HT}$ is as defined in (1.2.24),

$$\hat{V}\{\bar{y}_{HT} - \mu\} = \hat{V}\{\bar{y}_{HT} \mid \mathcal{F}_N\} + N^{-1}\hat{S}_y^2,$$

$$\hat{S}_y^2 = N^{-1}\sum_{i\in A_N} \pi_i^{-1}(y_i - \bar{y}_{HT})^2,$$

$\hat{V}\{\bar{y}_{HT} \mid \mathcal{F}_N\} = N^{-2}\hat{V}\{\hat{T}_y \mid \mathcal{F}_N\}$, and $\hat{V}\{\hat{T}_y \mid \mathcal{F}_N\}$ is as defined in (1.3.25).

Proof. By the moment properties of the superpopulation,

$$\theta_N = N^{1/2}(\bar{y}_N - \mu) \xrightarrow{\mathcal{L}} N(0,\ V_{11})$$

or, equivalently, for $f_\infty > 0$,

$$n_{BN}^{1/2}(\bar{y}_N - \mu) \xrightarrow{\mathcal{L}} N(0,\ f_\infty V_{11}).$$

By (1.3.36) of the proof of Theorem 1.3.5,

$$n_{BN}^{1/2}(\bar{y}_{HT} - \bar{y}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,\ V_{22}) \quad \text{a.s.},$$

where

$$\lim_{N\to\infty} n_{BN} N^{-2} \sum_{i=1}^{N} (1-\pi_i)\pi_i^{-1} y_i^2 \ = \ V_{22} \quad \text{a.s.}$$

Therefore, by Theorem 1.3.6,

$$n_B^{1/2}(\bar{y}_{HT} - \mu) \xrightarrow{\mathcal{L}} N(0,\ V_{22} + f_\infty V_{11}). \tag{1.3.46}$$

The estimated variance satisfies

$$\hat{V}\{\bar{y}_{HT} \mid \mathcal{F}_N\} - V\{\bar{y}_{HT} \mid \mathcal{F}_N\} \ = \ O_p(n_{BN}^{-1.5})$$

by (1.3.37). Also,

$$V\left\{N^{-1}\sum_{i\in A_N} \pi_i^{-1} y_i^2\right\} \ = \ O(n_{BN}^{-1})$$

and $\bar{y}_{HT}^2 = \mu^2 + O_p(n_{BN}^{-1/2})$ by the fourth moments of y_i. Therefore,

$$\hat{V}\{\bar{y}_{HT} - \mu\} \ = \ V\{\bar{y}_{HT} - \mu\} + O_p(n_{BN}^{-1.5}) \tag{1.3.47}$$

and result (1.3.45) follows from (1.3.46) and (1.3.47). ■

1.3.3 Functions of means

Theorems 1.3.2, 1.3.3, and 1.3.4 give the limiting distributions for means, but functions of means also occur frequently in the analysis of survey samples. It is a standard result that the limiting distribution of a continuous differentiable function of a sample mean is normal, provided that the standardized mean converges in distribution to a normal distribution.

Theorem 1.3.7. Let $\bar{\mathbf{x}}_n$ be a vector random variable with $E\{\bar{\mathbf{x}}_n\} = \boldsymbol{\mu}_x$ such that

$$n^{1/2}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_x) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \boldsymbol{\Sigma}_{xx})$$

as $n \to \infty$. Let $g(\bar{\mathbf{x}}_n)$ be a function of $\bar{\mathbf{x}}$ that is continuous at $\boldsymbol{\mu}_x$ with a continuous derivative at $\boldsymbol{\mu}_x$. Then

$$n^{1/2}\left[g(\bar{\mathbf{x}}_n) - g(\boldsymbol{\mu}_x)\right] \xrightarrow{\mathcal{L}} N\left[\mathbf{0},\ \mathbf{h}_x(\boldsymbol{\mu}_x)\boldsymbol{\Sigma}_{xx}\mathbf{h}_x'(\boldsymbol{\mu}_x)\right]$$

as $n \to \infty$, where $\mathbf{h}_x(\boldsymbol{\mu}_x)$ is the row vector of derivatives of $g(\bar{\mathbf{x}})$ with respect to $\bar{\mathbf{x}}$ evaluated at $\bar{\mathbf{x}} = \boldsymbol{\mu}_x$.

Proof. By a Taylor expansion

$$g(\bar{\mathbf{x}}_n) \quad = \quad g(\boldsymbol{\mu}_x) + (\bar{\mathbf{x}}_n - \boldsymbol{\mu}_x)\mathbf{h}_x'(\boldsymbol{\mu}_x^*),$$

where $\boldsymbol{\mu}_x^*$ is on the line segment joining $\bar{\mathbf{x}}_n$ and $\boldsymbol{\mu}_x$. Now $\bar{\mathbf{x}}_n - \boldsymbol{\mu}_x = O_p(n^{-1/2})$ and $\mathbf{h}_x(\mathbf{x})$ is continuous at $\mathbf{x} = \boldsymbol{\mu}_x$. Therefore, given $\delta > 0$ and $\varepsilon_x > 0$, there is some n_0 and a closed set B containing $\boldsymbol{\mu}_x$ as an interior point such that $\mathbf{h}_x(\mathbf{x})$ is uniformly continuous on B and $P\{\bar{\mathbf{x}}_n \in B$ and $\mid \bar{\mathbf{x}}_n - \boldsymbol{\mu}_x \mid < \varepsilon_x\} > 1 - \delta$ for $n > n_0$. Therefore, given $\delta > 0$ and an $\varepsilon_h > 0$, there is an $\varepsilon_x > 0$ and an n_0 such that

$$P\left\{\mid g(\bar{\mathbf{x}}_n) - g(\boldsymbol{\mu}_x) - (\bar{\mathbf{x}}_n - \boldsymbol{\mu}_x)\mathbf{h}_x'(\boldsymbol{\mu}_x) \mid > \varepsilon_h\right\} < \delta$$

for all $n > n_0$. The limiting normality follows because $n^{1/2}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_x)$ converges to a normal vector and $\mathbf{h}_x'(\boldsymbol{\mu}_x)$ is a fixed vector. ■

Ratios of random variables, particularly ratios of two sample means, play a central role in survey sampling. Because of their importance, we give a separate theorem for the large-sample properties of ratios.

Theorem 1.3.8. Let a sequence of finite populations be created as samples from a superpopulation with finite fourth moments. Let $\mathbf{x}_j = (x_{1j},\ x_{2j})$ and assume that $\mu_{x1} \neq 0$, where $\boldsymbol{\mu}_x = (\mu_{x1},\ \mu_{x2})$ is the superpopulation mean. Assume that the sequence of designs is such that

$$n_{BN}^{1/2}N^{-1}(\hat{\mathbf{T}}_x - N\boldsymbol{\mu}_x) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \boldsymbol{\Sigma}_{xx}) \tag{1.3.48}$$

and

$$n_{BN}^{1/2}N^{-1}(\hat{\mathbf{T}}_x - \mathbf{T}_{xN}) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{M}_{xx}), \tag{1.3.49}$$

where $\boldsymbol{\Sigma}_{xx}$ and $\mathbf{M}_{xx}$ are positive definite, $n_{BN} = E\{n_N \mid \mathcal{F}_N\}$,

$$\hat{\mathbf{T}}_x \quad = \quad \sum_{i \in A_N} \pi_i^{-1}\mathbf{x}_i,$$

and $\mathbf{T}_{xN} = N\bar{\mathbf{x}}_N$. Then

$$n_{BN}^{1/2}(\hat{R} - R_S) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{h}_S \boldsymbol{\Sigma}_{xx} \mathbf{h}_S') \tag{1.3.50}$$

and

$$n_{BN}^{1/2}(\hat{R} - R_N) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{h}_N \mathbf{M}_{xx} \mathbf{h}_N'), \tag{1.3.51}$$

where $\hat{R} = \hat{T}_{x1}^{-1}\hat{T}_{x2}$, $R_S = \mu_{x1}^{-1}\mu_{x2}$, $R_N = \bar{x}_{1N}^{-1}\bar{x}_{2N}$,

$$\mathbf{h}_S = \left(\frac{\partial R}{\partial x_1}, \frac{\partial R}{\partial x_2}\right) \mid \mathbf{x} = \boldsymbol{\mu}_x = \mu_{x1}^{-1}(1,\ -R_s)$$

and

$$\mathbf{h}_N = \left(\frac{\partial R}{\partial x_1}, \frac{\partial R}{\partial x_2}\right) \mid \mathbf{x} = \bar{\mathbf{x}}_N = \bar{x}_{1N}^{-1}(1,\ -R_N).$$

Let the designs be such that

$$[V\{\hat{\mathbf{T}}_x \mid \mathcal{F}_N\}]^{-1}\hat{V}_{HT}\{\hat{\mathbf{T}}_x\} - \mathbf{I} = O_p\left(n_{BN}^{-1/2}\right) \tag{1.3.52}$$

for any x-variable with finite fourth moments, where $\hat{V}_{HT}\{\hat{\mathbf{T}}_x\}$ is the Horvitz–Thompson variance estimator. Then

$$[\hat{V}_{HT}\{\hat{T}(\hat{d})\}]^{-1/2}(\hat{R} - R_N) \xrightarrow{\mathcal{L}} N(0,\ 1), \tag{1.3.53}$$

where $\hat{V}_{HT}\{\hat{T}(\hat{d})\}$ is the Horvitz–Thompson variance estimator calculated for

$$\hat{T}(\hat{d}) = \sum_{i \in A_N} \pi_i^{-1}\hat{d}_i,$$

and $\hat{d}_i = \hat{T}_{x1}^{-1}(x_{2i} - \hat{R}x_{1i})$.

Proof. Results (1.3.50) and (1.3.51) follow from (1.3.48), (1.3.49), and Theorem 1.3.7. The Taylor expansion for the estimator of the finite population ratio is

$$\begin{aligned}\hat{R} &= R_N + T_{x1,N}^{-1}(\hat{T}_{x2} - T_{x2,N}) - T_{x1,N}^{-2}T_{x2,N}(\hat{T}_{x1} - T_{x1,N}) + O_p(n_{BN}^{-1}) \\ &= R_N + T_{x1,N}^{-1}\left(\sum_{i \in A_N} \pi_i^{-1} e_i\right) + O_p(n_{BN}^{-1}),\end{aligned} \tag{1.3.54}$$

where $e_i = x_{2i} - R_N x_{1i}$. The remainder is $O_p(n_{BN}^{-1})$ because the second derivatives are continuous at $(T_{x1,N},\ T_{x2,N})$. Now

$$\hat{V}_{HT}\{\hat{T}(\hat{d})\} = \sum_{i,j \in A_N}\sum \pi_{ij}^{-1}(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\hat{d}_i\pi_j^{-1}\hat{d}_j,$$

where

$$\hat{d}_i = d_i - \hat{T}_{x1}^{-1} T_{x1,N}^{-1} (\hat{T}_{x1} - T_{x1,N}) e_i + \hat{T}_{x1}^{-1} (\hat{R} - R_N) x_{1i},$$

and $d_i = T_{x1,N}^{-1} e_i$. Because $\boldsymbol{\Sigma}_{xx}$ is positive definite, $V\{\hat{T}(d)\}$ is the same order as $V\{\hat{T}_{x1}\}$. It follows from (1.3.52) that

$$[V\{\hat{T}(d)\}]^{-1} \hat{V}_{HT}\{\hat{T}(\hat{d})\} = 1 + O_p(n_{BN}^{-1/2}), \tag{1.3.55}$$

because, for example,

$$[V\{\hat{T}(d)\}]^{-1} \hat{T}_{x1}^{-2} (\hat{R} - R_N)^2 \sum_{i,j \in A_N} \sum g_{ij} x_{1i} x_{2j} = O_p(n_B^{-1}),$$

$$[V\{\hat{T}(d)\}]^{-1} \hat{T}_{x1}^{-1} T_{x1,N}^{-1} (\hat{T}_{x1} - T_{x1,N}) \sum_{i,j \in A_N} \sum g_{ij} d_i e_j = O_p(n_B^{-1/2}),$$

where $g_{ij} = \pi_{ij}^{-1}(\pi_{ij} - \pi_i \pi_j)\pi_i^{-1}\pi_j^{-1}$. Result (1.3.53) follows from (1.3.55) and (1.3.54). ■

Theorem 1.3.8 is for unconditional properties derived under the probability structure defined by sampling from a finite population that is a sample from a superpopulation. It is possible to prove a theorem analogous to Theorem 1.3.5 in which the result holds almost surely for the sequence of finite populations.

Observe that the error in the ratio estimator is approximated by a design linear estimator in e_i in expression (1.3.54). This approximation is what leads to the limiting normality in (1.3.53). Although the estimator is not exactly normally distributed and $\hat{V}_{HT}\{\hat{T}(\hat{d})\}$ is not exactly a multiple of a chi-square random variable, the limiting distribution of the "t-statistic" is $N(0, 1)$. This type of result will be used repeatedly for nonlinear functions of Horvitz–Thompson estimators.

Expression (1.3.53) provides an efficient way to compute the estimated variance of the ratio using

$$\hat{d}_i = \hat{T}_{x1}^{-1}(x_{2i} - \hat{R} x_{1i}).$$

The $\hat{d}_i$ is sometimes called the *estimated Taylor deviate*. In the notation of Theorem 1.3.7, the Taylor deviate is

$$d_i = \mathbf{h}_x(\boldsymbol{\mu}_x)(\mathbf{x}_i - \bar{\mathbf{x}}_\pi)'$$

and the estimated variance of $g(\hat{\mathbf{T}}_x)$ is the estimated variance of $\bar{d}_{HT}$ calculated with $\hat{d}_i$.

We can use Theorem 1.3.8 to define an estimator for the population mean of y by letting $(x_{1i},\ x_{2i}) = (1,\ y_i)$. Then we obtain

$$\bar{y}_\pi = \left(\sum_{i\in A}\pi_i^{-1}\right)^{-1}\sum_{i\in A}\pi_i^{-1}y_i \tag{1.3.56}$$

as an estimator of $\bar{y}_N$. We gave the unbiased estimator

$$\bar{y}_{HT} = N^{-1}\hat{T}_y = N^{-1}\sum_{j\in A}\pi_j^{-1}y_j \tag{1.3.57}$$

in (1.2.24). The estimators (1.3.56) and (1.3.57) are identical for many designs, including stratified sampling, but can differ considerably for designs such as Poisson sampling with unequal probabilities. In general, and under mild regularity conditions, $N^{-1}\hat{T}_y$ is design unbiased and design consistent, whereas $\bar{y}_\pi$ is only design consistent. However, $\bar{y}_\pi$ is location and scale invariant, whereas $N^{-1}\hat{T}_y$ is only scale invariant. See (1.2.29). The estimator (1.3.56) is sometimes called the *Hájek estimator.* See Hájek (1971).

The estimators (1.3.56) and (1.3.57) can be compared under models for the population. One superpopulation model is

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + e_i, \\ e_i &\sim ind(0,\ x_i^\alpha\sigma^2), \end{aligned} \tag{1.3.58}$$

where α is positive, the x_i are positive, e_j is independent of x_i for all i and j, and $\sim$ *ind* denotes distributed independently. Let $(x_1,\ x_2,\ \ldots,\ x_N)$ be a finite population of positive x values, let the finite population of y_i values be generated by model (1.3.58), and let a sample be selected with probabilities $\pi_i = n(\sum_{j\in U}x_j)^{-1}x_i$.

Then the conditional expected value of the finite population mean is

$$E\{\bar{y}_N \mid \bar{x}_N\} = \beta_0 + \beta_1\bar{x}_N. \tag{1.3.59}$$

For fixed-size designs, the conditional expectations of the estimators are

$$E\{N^{-1}\hat{T}_y \mid \mathbf{x}_A\} = \beta_0 N^{-1}\hat{N}_{HT} + \beta_1\bar{x}_N, \tag{1.3.60}$$

$$E\{\bar{y}_\pi \mid \mathbf{x}_A\} = \beta_0 + \hat{N}_{HT}^{-1}N\beta_1\bar{x}_N, \tag{1.3.61}$$

where $\hat{N}_{HT} = \sum_{i\in A}\pi_i^{-1}$ and $\mathbf{x}_A$ is the set of x values in the sample. If $\beta_0 = 0$, $N^{-1}\hat{T}_y$ is conditionally unbiased, and if $\beta_1 = 0$, $\bar{y}_\pi$ is conditionally unbiased, conditional on $\mathbf{x}_A$.

The conditional variances are

$$V\{N^{-1}\hat{T}_y \mid \mathbf{x}_A\} = N^{-2}\sum_{i\in A}\pi_i^{-2+\alpha}\sigma^2 \tag{1.3.62}$$

and

$$V\{\bar{y}_\pi \mid \mathbf{x}_A\} = \left(\sum_{i\in A}\pi_i^{-1}\right)^{-2}\sum_{i\in A}\pi_i^{-2+\alpha}\sigma^2. \tag{1.3.63}$$

Thus, the conditional variance of $\bar{y}_\pi$ can be larger or smaller than that of $N^{-1}\hat{T}_y$.

The design variance of $\bar{y}_{HT}$ is

$$V\{\bar{y}_{HT} \mid \mathcal{F}\} = V\left\{N^{-1}\sum_{i\in A}\pi_i^{-1}y_i \mid \mathcal{F}\right\} \tag{1.3.64}$$

and the design variance of the approximate distribution of $\bar{y}_\pi$ is

$$V\{\bar{y}_\pi \mid \mathcal{F}\} = V\left\{N^{-1}\sum_{i\in A}\pi_i^{-1}(y_i - \bar{y}_N) \mid \mathcal{F}\right\}. \tag{1.3.65}$$

Thus, as suggested by (1.3.64) and (1.3.65), $\bar{y}_{HT}$ will have smaller design variance than $\bar{y}_\pi$ if the ratio of y_i to π_i is nearly constant and $\bar{y}_\pi$ will have smaller design variance than $\bar{y}_{HT}$ if $y_i - \bar{y}_N$ is nearly a constant multiple of π_i. Also see Exercise 6.

Because $\bar{y}_\pi$ is location invariant, we generally begin estimation for more complex situations with $\bar{y}_\pi$. A regression estimator that is conditionally model unbiased, conditional on $\mathbf{x}_A$, is discussed in Chapter 2.

In the analysis of survey samples, subpopulations are often called *domains of study* or, simply, *domains*. Thus, in reporting unemployment rates, the rate might be reported for a domain composed of females aged 35 to 44. To study the properties of the estimated mean for a domain, let

$$\begin{aligned} y_{Di} &= y_i && \text{if element } i \text{ is in domain } D \\ &= 0 && \text{otherwise,} \\ z_{Di} &= 1 && \text{if element } i \text{ is in domain } D \\ &= 0 && \text{otherwise.} \end{aligned}$$

Then the estimator of the domain mean is the ratio estimator

$$\hat{\theta}_D = \bar{z}_{D\pi}^{-1}\bar{y}_{D\pi} = \hat{T}_{zD}^{-1}\hat{T}_{yD}, \tag{1.3.66}$$

where

$$(\hat{T}_{zD},\ \hat{T}_{zD}) \quad = \quad \sum_{i \in A} \pi_i^{-1}(z_{Di},\ y_{Di}).$$

The Horvitz–Thompson variance estimator of Theorem 1.3.8 is

$$\hat{V}_{LS}\{\hat{\theta}_D\} \quad = \quad \bar{z}_{D\pi}^{-2} \sum_{i,j \in A} \sum \pi_{ij}^{-1}(\pi_{ij} - \pi_i \pi_j)\pi_i^{-1} \hat{e}_i \pi_j^{-1} \hat{e}_j, \qquad (1.3.67)$$

where $\hat{e}_i = y_{Di} - \hat{\theta}_D z_{Di}$. Observe that $\hat{e}_i$ is zero if element i is not in the domain.

The properties of the estimated domain mean illustrate the care required in the use of large-sample results. Assume that we have a simple random sample from a finite population that is, in turn, a random sample from a normal distribution. Assume that the finite population correction can be ignored. Then the domain mean is the simple mean of the elements in the domain,

$$\hat{\theta}_D \quad = \quad \bar{y}_D \quad = \quad n_D^{-1} \sum_{i \in A_D} y_i, \qquad (1.3.68)$$

where $n_D = \Sigma_{i \in A} z_{Di}$ is the number of elements in domain D and A_D is the set of indices of elements in domain D. The variance estimator (1.3.67) becomes

$$\hat{V}_{LS}\{\hat{\theta}_D\} \quad = \quad n_D^{-2} n^2 \left[n(n-1)\right]^{-1} \sum_{i \in A_D} (y_i - \bar{y}_D)^2. \qquad (1.3.69)$$

Because the original sample is a simple random sample, the n_D elements selected from domain D are a simple random sample from that domain. Therefore,

$$t_{srs,D} \quad = \quad [\hat{V}_{srs}\{\bar{y}_D\}]^{-1/2}(\bar{y}_D - \mu_D) \qquad (1.3.70)$$

is, conditional on $n_D, n_D > 1$, distributed as Student's t with $n_D - 1$ degrees of freedom, where

$$\hat{V}_{srs}\{\bar{y}_D\} \quad = \quad [n_D(n_D - 1)]^{-1} \sum_{i \in A_D} (y_i - \bar{y}_D)^2 \qquad (1.3.71)$$

and μ_D is the population domain mean. If we use (1.3.69) to construct the estimated variance, we have

$$\hat{V}_{LS}\{\hat{\theta}_D\} \quad = \quad n_D^{-1}(n_D - 1)\hat{V}_{srs}\{\hat{\theta}_D\}. \qquad (1.3.72)$$

Thus, the estimator based on the large-sample approximation underestimates the variance and

$$[\hat{V}_{LS}\{\hat{\theta}_D\}]^{-1/2}(\hat{\theta}_D - \mu_D) \sim \left[n_D(n_D-1)^{-1}\right]^{1/2} t_{srs,D}, \qquad (1.3.73)$$

where t_{n_D-1} is distributed as Student's t with $n_D - 1$ degrees of freedom. For small n_D, $N(0, 1)$ will be a very poor approximation for the distribution of (1.3.73).

The assumptions of Theorem 1.3.8 require the distributions of $\bar{z}_{D\pi}$ and $\bar{y}_{D\pi}$ to have small variances. This condition does not hold for the components of the domain mean if n_D is small, no matter how large the original sample.

1.3.4 Approximations for complex estimators

An estimator is often defined as the solution to a system of equations, where the solution may be implicit. In Theorem 1.3.9 we show that Taylor methods can be used to obtain an approximation to the distribution of such an estimator. Results are given for $\hat{\boldsymbol{\theta}}$ as an estimator of the finite population parameter and for $\hat{\boldsymbol{\theta}}$ as an estimator of the parameter of the superpopulation that generated the finite population.

The theorem contains a number of technical assumptions. They can be summarized as assumptions of existence of moments for the superpopulation, assumptions pertaining to the design, and assumptions about the functions defining the estimator. The design must be such that a central limit theorem holds for the Horvitz–Thompson estimator, and the function must be continuous with at least a continuous second derivative with respect to the parameter.

It is assumed that the estimator is consistent. See (1.3.80) and (1.3.81). The consistency assumption is required because some functions have more than one root. If the function $\mathbf{g}(\mathbf{x}, \boldsymbol{\theta})$ is the vector of derivatives of an objective function, it may be possible to use the properties of the objective function to prove (1.3.80) and (1.3.81). See, for example, Gallant (1987). The usual w_i of the theorem is π_i^{-1}, but alternative weights, some considered in Chapter 6, are possible. Equation (1.3.75), which defines the estimator, is sometimes called an *estimating equation*. See Godambe (1991).

Theorem 1.3.9. Let $\mathcal{F}_N = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be the Nth finite population in the sequence $\{\mathcal{F}_N\}$, where $\{\mathbf{x}_i\}$ is a sequence of iid random variables with finite fourth moments. Assume that the sequence of designs is such that for any $\mathbf{x}_j$ with positive variance,

$$V\{n_{BN}^{1/2}N^{-1}(\hat{\mathbf{T}}_x - \mathbf{T}_{x,N}) \mid \mathcal{F}_N\} = \mathbf{V}_{T,xx,N}, \qquad (1.3.74)$$

where $\mathbf{V}_{T,xx,N}$ is positive semidefinite almost surely, $\mathbf{T}_{x,N}$ is the population total of $\mathbf{x}$, $\hat{\mathbf{T}}_{\mathbf{x}}$ is the Horvitz–Thompson estimator of the total, and $n_{BN} = E\{n_N \mid \mathcal{F}_N\}$. Let an estimator $\hat{\boldsymbol{\theta}}$, be defined by

$$\sum_{i \in A} w_i \mathbf{g}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{1.3.75}$$

and let $\boldsymbol{\theta}_N$ satisfy

$$\sum_{i \in U} \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}_N) = \mathbf{0}, \tag{1.3.76}$$

where we have omitted the subscript N on U_N and A_N. Assume that $\mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta}$ in a closed set $\mathcal{B}$ containing $\boldsymbol{\theta}^\circ$ as an interior point and all $\mathbf{x}_i$, where $\boldsymbol{\theta}^\circ$ satisfies

$$E\left\{\sum_{i \in U} \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}^\circ)\right\} = \mathbf{0}. \tag{1.3.77}$$

Assume that $\mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}) = \partial \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$ is continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta}$ in $\mathcal{B}$ and all $\mathbf{x}_i$. Assume for all $\boldsymbol{\theta}$ in $\mathcal{B}$ that

$$N^{-1} \sum_{i \in A} w_i \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}) = N^{-1} \sum_{i \in U} \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}) + O_p(n_{BN}^{-1/2}) \tag{1.3.78}$$

and

$$\lim_{N \to \infty} N^{-1} \sum_{i \in U} \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta}) \quad \text{a.s.},$$

where $\mathbf{H}(\boldsymbol{\theta})$ is nonsingular. Assume that

$$\mid \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}) \mid < K(\mathbf{x}_i) \tag{1.3.79}$$

for some $K(\mathbf{x})$ with finite fourth moment for all $\mathbf{x}_i$ and all $\boldsymbol{\theta}$ in $\mathcal{B}$. Assume that

$$p \lim_{N \to \infty} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ) = \mathbf{0} \tag{1.3.80}$$

and

$$p \lim_{N \to \infty} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) \mid \mathcal{F}_N = \mathbf{0} \quad \text{a.s.} \tag{1.3.81}$$

Then

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ = \left(\sum_{i \in A} w_i \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}^\circ)\right)^{-1} \sum_{i \in A} w_i \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}^\circ) + o_p(n_{BN}^{-1/2}) \tag{1.3.82}$$

and

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N = \left(\sum_{i\in A} w_i \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}_N)\right)^{-1} \sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}_N) + o_p(n_{BN}^{-1/2}). \tag{1.3.83}$$

Let $\mathbf{V}_{t,xx,N}$ denote the conditional variance of $N^{-1}(\hat{\mathbf{T}}_x - \mathbf{T}_{x,N})$ conditional on $\mathcal{F}_N$, and assume that

$$\mathbf{V}_{t,xx,N}^{-1/2} N^{-1}(\hat{\mathbf{T}}_x - \mathbf{T}_{x,N}) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \quad \text{a.s.} \tag{1.3.84}$$

and

$$\lim_{N\to\infty} (n_{BN}\mathbf{V}_{t,xx,N} - \boldsymbol{\Sigma}_{t,xx}) = \mathbf{0} \quad \text{a.s.}, \tag{1.3.85}$$

where $\boldsymbol{\Sigma}_{t,xx}$ is positive definite for any $\mathbf{x}_j$ with positive definite covariance matrix. Then

$$[V_\infty\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\}]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \tag{1.3.86}$$

and

$$[V_\infty\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \mathcal{F}_N\}]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \quad \text{a.s.}, \tag{1.3.87}$$

where

$$\begin{aligned}
V_\infty\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \mathcal{F}_N\} &= \mathbf{H}^{-1}(\boldsymbol{\theta}_N)\mathbf{V}_{t,gg,N}\mathbf{H}^{-1}(\boldsymbol{\theta}_N), \\
\mathbf{H}(\boldsymbol{\theta}_N) &= N^{-1}\sum_{i\in U} \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}_N), \\
\mathbf{V}_{t,gg,N} &= V\left\{N^{-1}\sum_{i\in A} w_i g(\mathbf{x}_i, \boldsymbol{\theta}_N) \mid \mathcal{F}_N\right\}, \\
V_\infty\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\} &= \mathbf{H}^{-1}(\boldsymbol{\theta}^\circ)(N^{-1}\boldsymbol{\Sigma}_{gg} + \mathbf{V}_{t,gg,N}), \mathbf{H}^{-1}(\boldsymbol{\theta}^\circ),
\end{aligned}$$

and $\boldsymbol{\Sigma}_{gg} = E\{\mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}^\circ)\mathbf{g}'(\mathbf{x}_i, \boldsymbol{\theta}^\circ)\}$.

Proof. For $\hat{\boldsymbol{\theta}} \in \mathcal{B}$, by a Taylor expansion,

$$\begin{aligned}
N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) &= N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}^\circ) \\
&\quad + N^{-1}\sum_{i\in A} w_i \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ)
\end{aligned}$$

$$= N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i,\ \boldsymbol{\theta}^\circ) + N^{-1}\sum_{i\in A} w_i \mathbf{H}(\mathbf{x}_i,\ \boldsymbol{\theta}^\circ)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ) + o_p(|\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\ |), \tag{1.3.88}$$

where $\boldsymbol{\theta}^*$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\circ$. The continuity of $\mathbf{H}(\mathbf{x}_i,\ \boldsymbol{\theta})$, (1.3.80), and (1.3.78) were used to obtain the second equality. Given $\epsilon > 0$, by (1.3.80), there is an n_o such that for $n > n_o$, $P\{\hat{\boldsymbol{\theta}} \in \mathcal{B}\} > 1 - \epsilon$. Therefore, result (1.3.88) holds in general. Now, by (1.3.79) and (1.3.74),

$$N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i,\ \boldsymbol{\theta}^\circ) - N^{-1}\sum_{i\in U} \mathbf{g}(\mathbf{x}_i,\ \boldsymbol{\theta}^\circ) \ = \ O_p(n_{BN}^{-1/2})$$

and by (1.3.79),

$$V\left\{N^{-1}\sum_{i\in U} \mathbf{g}(\mathbf{x}_i,\ \boldsymbol{\theta}^\circ)\right\} \ = \ O(N^{-1}).$$

Therefore,

$$N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i,\ \boldsymbol{\theta}^\circ) \ = \ O_p(n_{BN}^{-1/2}) \tag{1.3.89}$$

and result (1.3.82) is proven. Also see Exercise 31.

Result (1.3.83) follows by analogous arguments.

By (1.3.83) and (1.3.78),

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \mathcal{F}_N = \left(\sum_{i\in U} \mathbf{H}(\mathbf{x}_i,\ \boldsymbol{\theta}_N)\right)^{-1} \sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i,\ \boldsymbol{\theta}_N) + o_p(n_{BN}^{-1/2}) \quad \text{a.s.} \tag{1.3.90}$$

and result (1.3.87) follows from (1.3.79) and the independence assumption. By (1.3.79) and the Lindeberg Central Limit Theorem,

$$N^{-1/2}\sum_{i\in U} \mathbf{g}(\mathbf{x}_i,\ \boldsymbol{\theta}^\circ) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \boldsymbol{\Sigma}_{gg})$$

and, by (1.3.78),

$$N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ) \xrightarrow{\mathcal{L}} N\left[\mathbf{0},\ \mathbf{H}^{-1}(\boldsymbol{\theta}^\circ)\boldsymbol{\Sigma}_{gg}\mathbf{H}^{-1}(\boldsymbol{\theta}^\circ)\right]. \tag{1.3.91}$$

Result (1.3.86) follows from (1.3.90) and (1.3.91) by Theorem 1.3.6. ■

To apply Theorem 1.3.9, we require estimators of the variances. Estimators are obtained by substituting estimators for unknown parameters.

Corollary 1.3.9.1. Let the assumptions of Theorem 1.3.9 hold. Then

$$[\hat{V}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \mathcal{F}_N\}]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \quad \text{a.s.}$$

and

$$[\hat{V}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\}]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}),$$

where

$$\hat{V}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \mathcal{F}_N\} = \hat{\mathbf{T}}_H^{-1}\hat{V}_{HT}\left\{\sum_{i\in A}\mathbf{g}(\mathbf{x}_i, \hat{\boldsymbol{\theta}})\right\}\hat{\mathbf{T}}_H^{-1},$$

$$\hat{V}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\} = \hat{V}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \mathcal{F}_N\} + \hat{\mathbf{T}}_H^{-1}N\hat{\boldsymbol{\Sigma}}_{gg}\hat{\mathbf{T}}_H^{-1},$$

$$\hat{\mathbf{T}}_H = \sum_{i\in A} w_i \mathbf{H}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}),$$

and

$$N\hat{\boldsymbol{\Sigma}}_{gg} = \sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i, \hat{\boldsymbol{\theta}})\mathbf{g}'(\mathbf{x}_i, \hat{\boldsymbol{\theta}}).$$

Proof. By the assumptions that $\mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta})$ and $\mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta})$ are continuous in $\boldsymbol{\theta}$,

$$N^{-1}\sum_{i\in A} w_i \mathbf{H}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) - N^{-1}\sum_{i\in A} w_i \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}^\circ) = o_p(1),$$

$$N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) - N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}^\circ) = O_p(n_{BN}^{-1/2}),$$

and

$$\hat{\boldsymbol{\Sigma}}_{gg} - N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}^\circ)\mathbf{g}'(\mathbf{x}_i, \boldsymbol{\theta}^\circ) = O_p(n_{BN}^{-1/2}).$$

Therefore, by (1.3.78),

$$N^{-1}\sum_{i\in A} w_i \mathbf{H}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) - N^{-1}\sum_{i\in U} \mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}^\circ) = o_p(1).$$

Furthermore, by (1.3.79),

$$N^{-1}\sum_{i\in A} w_i \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}^\circ) - N^{-1}\sum_{i\in U} \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}^\circ) = O_p(n_{BN}^{-1/2})$$

and

$$\hat{\boldsymbol{\Sigma}}_{gg} - \boldsymbol{\Sigma}_{gg} = O_p(n_{BN}^{-1/2}).$$

The conclusions then follow because the estimators of the variances are consistent estimators. ■

1.3.5 Quantiles

Means, totals, and functions of means are the most common statistics, but estimators of the quantiles of the distribution function are also important. Let y be the variable of interest and define the finite population distribution function by

$$F_{y,N}(a) = N^{-1}\sum_{i=1}^{N} d_{ai}, \tag{1.3.92}$$

where

$$\begin{aligned} d_{ai} &= 1 \quad \text{if } y_i \leq a \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Given a sample, an estimator of the distribution function at point a is the sample mean of the indicator function

$$\hat{F}_y(a) = \bar{d}_{a\pi} = \left(\sum_{i\in A}\pi_i^{-1}\right)^{-1}\sum_{i\in A}\pi_i^{-1}d_{ai}. \tag{1.3.93}$$

The finite population quantile is defined as

$$\xi_{b,N} =: Q_{y,N}(b) = \inf\{a : F_{y,N}(a) \geq b\} \tag{1.3.94}$$

and the sample quantile by

$$\hat{\xi}_b =: \hat{Q}_y(b) = \inf\{a : \hat{F}_y(a) \geq b\}. \tag{1.3.95}$$

Estimated quantiles are not simple functions of means, and therefore the results of Section 1.3.3 are not applicable. However, the relationship between the distribution function and the quantile function can be exploited to obtain useful results.

Let $\hat{s}_{ca}^2$ be the estimated variance of $\hat{F}_y(a)$ and assume that the sample is large enough so that $\hat{F}_y(a)$ can be treated as being normally distributed. Then the hypothesis that $F_y(a) = b$ will be accepted at the α level if $\hat{F}(a)$ falls in the interval

$$(b - t_\alpha \hat{s}_{ca},\ b + t_\alpha \hat{s}_{ca}), \tag{1.3.96}$$

where t_α is the α percentage point of the normal distribution. If $\hat{F}_y(a)$ is in the interval defined in (1.3.96), then

$$\hat{Q}_y(b - t_\alpha \hat{s}_{ca}) \leq Q_y(b) \leq \hat{Q}_y(b + t_\alpha \hat{s}_{ca}), \tag{1.3.97}$$

where $F(a) = b$. Therefore, $[\hat{Q}_y(b - t_\alpha \hat{s}_{ca}),\ \hat{Q}(b + t_\alpha \hat{s}_{ca})]$ is a $1 - \alpha$ confidence interval for $Q_y(b)$. Intervals of this type are sometimes called *test inversion intervals*. The interval (1.3.97) is also called the *Woodruff interval* in the survey sampling literature. See Woodruff (1952).

Using a plot of the distribution function, one can see that shifting the function up by an amount δ will shift the quantile left by an amount approximately equal to δ divided by the slope of the distribution function. This local approximation can be used to approximate the distribution of a quantile. For simple random samples from a distribution with a density, the limiting distribution of a quantile associated with a positive part of the density is normal because the error in the quantile can be written

$$\hat{\xi}_b - \xi_b = [f_y(a)]^{-1}(b - \bar{d}_{a,\pi}) + o_p(n^{-1/2}), \tag{1.3.98}$$

where $\xi_b = a$ and $f_y(a)$ is the density of y evaluated at a. Equation (1.3.98) is called the *Bahadur representation*. See Bahadur (1966), Ghosh (1971), and David (1981, Section 9.2). Francisco and Fuller (1991) extended representation (1.3.98) to a more general class of samples and used the representation to show that sample quantiles for complex samples are normally distributed in the limit.

Theorem 1.3.10. Let a sequence of finite populations be created as samples from a superpopulation with cumulative distribution function $F_y(\cdot)$ and finite fourth moments. Let $\xi_b^\circ = a^\circ$ be the bth quantile. Assume that the cumulative distribution function $F_y(a)$ is continuous with a continuous positive derivative on a closed interval B containing a° as an interior point. Assume that the sequence of designs is such that

$$n_{BN}^{1/2} N^{-1}(\hat{T}_x - N\mu_x) \xrightarrow{\mathcal{L}} N(0,\ \sigma_{xx}),$$

$$n_{BN}^{1/2} N^{-1}(\hat{T}_x - N\bar{x}_N) \xrightarrow{\mathcal{L}} N(0,\ M_{xx}),$$

$$[V\{\hat{T}_x \mid \mathcal{F}_N\}]^{-1} \hat{V}_{HT}\{\hat{T}_x \mid \mathcal{F}_N\} - 1 = O_p(n_{BN}^{-1/2}),$$

$$[V\{\bar{x}_\pi\}]^{-1} \hat{V}\{\bar{x}_\pi\} - 1 = O_p(n_{BN}^{-1/2}),$$

for any x with positive variance and fourth moment, where $n_{BN} = E\{n_N\}$, $\hat{V}_{HT}\{\hat{T}_x \mid \mathcal{F}_N\}$ is the Horvitz–Thompson estimator of the variance of $\hat{T}_x - T_x$ given $\mathcal{F}_N$, $\hat{V}\{\bar{x}_\pi\}$ is an estimator of the unconditional variance of $\bar{x}_\pi - \mu_x$, and μ_x is the superpopulation mean. Assume that $n_N V\{\hat{F}_y(a)\}$ and $n_N V\{\hat{F}_y(a) - F_{y,N}(a) \mid \mathcal{F}_N\}$ are positive and continuous in a for $a \in B$. Assume that

$$V\{\hat{F}_y(a + \delta) - \hat{F}_y(a)\} \leq C n_N^{-1} \mid \delta \mid$$

for some $0 < C < \infty$, for all N, and for all a and $a + \delta$ in B. Then

$$\hat{s}_{ca}^{-1}\hat{f}_y(\hat{\xi}_b)(\hat{\xi}_b - a^\circ) \xrightarrow{\mathcal{L}} N(0,\ 1), \tag{1.3.99}$$

where $\hat{s}_{ca}^2 = \hat{V}\{\hat{F}_y(a)\}$,

$$\hat{f}_y(\hat{\xi}_b) = (2t_\alpha \hat{s}_{ca})[\hat{Q}(b + t_\alpha \hat{s}_{ca}) - \hat{Q}(b - t_\alpha \hat{s}_{ca})]^{-1},$$

$\hat{a} = \hat{\xi}_b$, t_α is defined by $\Phi(t_\alpha) = 1 - 0.5\alpha$, and $\Phi(\cdot)$ is the distribution function of a standard normal random variable.

Also,

$$\hat{s}_{HT,ca}^{-1}\hat{f}_y(\hat{\xi}_b)(\hat{\xi}_b - \xi_{b,N}) \xrightarrow{\mathcal{L}} N(0, 1),$$

where $\hat{s}_{HT,ca}^2 = \hat{V}\{\hat{F}_y(a) - F_{y,N}(a) \mid \mathcal{F}_N\}$.

Proof. Omitted. See Francisco and Fuller (1991) and Shao (1994). ∎

In Theorem 1.3.10, the ratio of the difference between two values of the sample distribution function to the distance between the points defining the values is used to estimate the density. The use of $t_\alpha = 2$ in (1.3.99) to estimate $f_y(a)$ seems to work well in practice. The estimator of $f_y(a)$ in (1.3.99) can be viewed as a regression in the order statistics for order statistics "close" to $\hat{\xi}_b$.

Let $y_{(r)}$ be the largest order statistic less than $\hat{Q}(\hat{\xi}_b - t_\alpha \hat{s}_{ca})$, let $y_{(m)}$ be the smallest order statistic greater than $\hat{Q}(\hat{\xi}_b + t_\alpha \hat{s}_{ca})$, and let a "smoothed" estimator of the distribution function of $y_{(i)}$ be

$$z_i \ = \ 0.5[\hat{F}(y_{(i)}) + \hat{F}(y_{(i-1)})]. \tag{1.3.100}$$

Let $\hat{\theta}_0$ and $\hat{\theta}_1$ be the regression coefficients obtained in a weighted regression of z_i on $(1,\ y_{(i)})$ for $i = r,\ r+1,\ \ldots,\ m$. Then $\hat{\theta}_1$ is an estimator of $f_y(\xi_b)$ and

$$\tilde{\xi}_b \ = \ \hat{\theta}_1^{-1}(b - \hat{\theta}_0) \tag{1.3.101}$$

is a smoothed estimator of ξ_b. If only $y_{(m)}$ and $y_{(r)}$ are used in the regression

$$\hat{\theta}_1 \ = \ (y_{(m)} - y_{(r)})^{-1}(z_m - z_r),$$

$\hat{\theta}_0 = z_k - \hat{\theta}_1 y_{(k)}$, and

$$\hat{\xi}_b \ = \ \bar{y}_{(r)} + (z_m - z_r)^{-1}(y_{(m)} - y_{(r)})(b - z_r).$$

There are many smoothed estimators of quantiles. See Silverman (1986) and Scott (1992).

1.4 METHODS OF UNEQUAL PROBABILITY SAMPLE SELECTION

The literature contains numerous procedures for the selection of nonreplacement unequal probability samples. The number of procedures is indicative of the difficulty of constructing a completely general procedure that is not extremely cumbersome computationally. We consider only sampling schemes where selection is not a function of the y values. For selection procedures that are functions of y, see Thompson and Seber (1996).

Selection procedures have been classified by Carroll and Hartley (1964) as draw-by-draw methods, mass draw procedures wherein samples are rejected if duplication occurs, and systematic procedures. The draw-by-draw and mass draw methods require computation of "working probabilities" if the probability of selection is to be maintained at the values specified. The working probabilities are typically given as the solutions to a system of N, Nn, or $N(n-1)$ equations. Some procedures have been demonstrated to be superior to replacement sampling for $n = 2$ (Fellegi, 1963), while others are justified on the basis of joint probabilities that guarantee nonnegative estimators of variance (Hanurav, 1967; Vijayan, 1968). Alternative procedures have been discussed by Jessen (1969) and Rao (1978), and procedures have been reviewed extensively by Brewer and Hanif (1983) and Tillé (2006).

Perhaps the most common method of selecting an unequal probability sample is the systematic procedure described in Section 1.2.4. The systematic procedure is easy to implement but has the disadvantage that no design-unbiased estimator of the variance is available.

The following two draw-by-draw methods of selecting a sample of size 2 yield the same joint inclusion probabilities. The first was suggested by Brewer (1963a) and the second by Durbin (1967). Let p_i be a set of positive numbers (probabilities) with the properties that $\Sigma_{i=1}^{N} p_i = 1$ and $p_i < 0.5$ for all i. The selection probability is then $\pi_i = 2p_i$.

Procedure 1

1. Select a unit with probability q_{1j}, where

$$q_{1j} = \left(\sum_{i=1}^{N} (1-2p_i)^{-1} p_i (1-p_i) \right)^{-1} (1-2p_j)^{-1} p_j (1-p_j). \tag{1.4.1}$$

2. Select a second unit with probability

$$q_{2j} = (1-p_{i(1)})^{-1} p_j, \tag{1.4.2}$$

where $p_{i(1)}$ is the value of p for the unit selected at the first draw.

Procedure 2

1. Select a unit with probability p_i.
2. Select a second unit with probability $p_i^{-1}\pi_{ij}$, $j \neq i$, where i is the unit selected on the first draw and

$$\pi_{ij} = \frac{\pi_i \pi_j}{2(1+A)} \left(\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j} \right), \tag{1.4.3}$$

where

$$A = \frac{1}{2} \sum_{i=1}^{N} \frac{\pi_i}{1-\pi_i} = \sum_{i=1}^{N} \frac{p_i}{1-2p_i}. \tag{1.4.4}$$

For both procedures, the joint probability is given by (1.4.3) and the total probability of selecting unit i is $\pi_i = 2p_i$. Under selection procedure 2, the probability that the unit is selected on the first draw is p_i and the probability that it is selected on the second draw is p_i. Fuller (1971) gave the following motivation for the joint probabilities (1.4.3).

Assume that the population of N values of $p_i^{-1}y_i$ is a random sample from a normal population with variance σ^2. Then considering the population of all such populations, the variance of the Yates–Grundy–Sen estimated variance for samples of size 2 is

$$8\sigma^4 \sum_{i<j}^{N} \pi_{ij}^{-1} (\pi_i \pi_j - \pi_{ij})^2. \tag{1.4.5}$$

Therefore, under this model, minimization of the summation with respect to π_{ij} will result in a minimum variance for the estimated variance. Since minimization of this expression leads to a system of nonlinear equations for the π_{ij}, consider the approximation obtained by replacing the π_{ij} in the denominator by $\pi_i \pi_j$.

For $n = 2$ those π_{ij} that minimize

$$\sum_{i<j}^{N} \frac{(\pi_i \pi_j - \pi_{ij})^2}{\pi_i \pi_j} \tag{1.4.6}$$

subject to the restrictions that

$$\sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_{ij} = \pi_i \qquad \text{for all} \quad i \tag{1.4.7}$$

are the π_{ij} of (1.4.3). The π_{ij} of (1.4.3) are positive, and furthermore, for $0 < \pi_i < 1.0$,

$$\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j} = \frac{\pi_i}{1-\pi_i} + \frac{\pi_j}{1-\pi_j} + 2 < 2 + 2A$$

and hence $\pi_{ij} < \pi_i \pi_j$. Therefore, the joint probabilities (1.4.3) permit the construction of an unbiased nonnegative estimator of variance. The following theorem demonstrates that the sampling scheme is always more efficient than replacement sampling.

Theorem 1.4.1. The variance of the Horvitz–Thompson estimator for the sampling scheme with joint probabilities (1.4.3) is never greater than that of estimator (1.2.70) for replacement sampling, equality holding only if all $(z_i - z_j)^2 = 0$, where $z_i = y_i \pi_i^{-1}$.

Proof. Using the variance expression (1.2.28), the variance of the Horvitz–Thompson estimated total is

$$V_N(\hat{Y}) = \sum_{i<j} \left[\pi_i \pi_j - \frac{\pi_i \pi_j}{2(1+A)} \left(\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j} \right) \right] (z_i - z_j)^2$$

and the variance of the replacement sampling estimator (1.2.70) is

$$V_R(\hat{Y}) = \frac{1}{2} \sum_{i<j} \pi_i \pi_j (z_i - z_j)^2 = \sum_{j=1}^{N} \pi_j (z_j - 0.5Y)^2 .$$

Then

$$V_R - V_N = \frac{1}{1+A} \left[-(1+A) V_R(\hat{Y}) + \frac{1}{2} \sum_i \frac{\pi_i}{1-\pi_i} \sum_j \pi_j (z_i - z_j)^2 \right]$$

$$= \frac{1}{1+A} \left[\sum_i \frac{\pi_i^2}{1-\pi_i} (z_i - 0.5Y)^2 \right] \geq 0,$$

equality holding only if all $z_i \equiv 0.5Y$. ■

Two procedures that maintain inclusion probabilities equal to np_i for $n > 2$ are that of Brewer (1963a) and that proposed by Rao (1965) and Sampford (1967). In the Brewer (1963a) procedure, the first selection for a sample of size n is made with probability

$$q_{jN} = \left[\sum_{i=1}^{N} (1 - np_i) p_i (1 - p_i) \right]^{-1} (1 - np_j)^{-1} p_j (1 - p_j),$$

the next with probabilities proportional to

$$[1-(n-1)p_j]^{-1}p_j(1-p_j),$$

and so on.

Another way to select unequal probability samples is to select a replacement sample and reject the sample if the sample contains duplicates. Hájek (1964) studied this procedure. To illustrate, consider the selection of a sample of size 2 from a population of size N with draw probabilities p_i. The total probability of selecting the sample is

$$1 = \left(\sum_{i=1}^{N} p_i\right)^2$$

and the probabilities of selecting one of the units twice is

$$P\{\text{repeated selection}\} = \sum_{i=1}^{N} p_i^2.$$

Thus, the joint probability of element i and element j, $i \neq j$, appearing in a sample where samples with repeated elements are rejected is

$$\pi_{ij|NR} = P\{(i,\, j) \in A \mid NR\} = \left(1-\sum_{k=1}^{N} p_k^2\right)^{-1} p_i p_j,$$

where NR denotes no repeated elements in the sample. The probability that element i appears in the sample is

$$\pi_{i|NR} = \sum_{\substack{j=1 \\ j\neq i}}^{N} \pi_{ij} = \left(1-\sum_{k=1}^{N} p_k^2\right)^{-1} (1-p_i)\, p_i.$$

If a nonreplacement sample with selection probabilities close to the specified π_i is desired, working probabilities must be specified. Hájek (1964) suggested approximate p_i, and Carroll and Hartley (1964) gave an iterative procedure, described by Brewer and Hanif (1983), for determining working probabilities. Chen, Dempster, and Liu (1994) give a computational algorithm that can be used for sample selection. For a complete discussion, see Tillé (2006, Chapter 5). Also see Section 3.4.

1.5 REFERENCES

Sections 1.1, 1.2. Brewer (1963b), Cochran (1946, 1977), Goldberger (1962), Graybill (1976), Hansen and Hurwitz (1943), Horvitz and

Thompson (1952), Narain (1951), Royall (1970), Sen (1953), Stuart and Ord (1991), Yates (1948), Yates and Grundy (1953).

Section 1.3. Bickel and Freedman (1984), Binder (1983), Blight (1973), Cochran (1946, 1977), Francisco (1987), Francisco and Fuller (1991), Fuller (1975, 1987b, 1996), Hájek (1960), Hannan (1962), Isaki and Fuller (1982), Krewski and Rao (1981), Madow (1948), Madow and Madow (1944), Papageorgiou and Karakostas (1998), Rao and Wu (1987), R. R. Rao (1962), Rubin-Bleuer and Kratina (2005), Sen (1988), Shao (1994), Thompson (1997), Woodruff (1952, 1971), Xiang (1994).

Section 1.4. Brewer (1963a), Brewer and Hanif (1983), Carroll and Hartley (1964), Durbin (1967), Fellegi (1963), Fuller (1971), Hájek (1964), Hanurav (1967), Hedayat and Sinha (1991), Jessen (1969), Rao (1965, 1978), Rao, Hartley, and Cochran (1962), Sampford (1967), Vijayan (1968), Yates and Grundy (1953).

1.6 EXERCISES

1. (Section 1.2.1) Let $\mathbf{d} = (I_1, I_2, \ldots, I_N)$, as defined in (1.2.4). Show that a matrix expression for the variance of the design linear estimator $\hat{\theta}$ of (1.2.17) is

$$V\{\hat{\theta} \mid \mathcal{F}\} = \mathbf{y}_N \mathbf{W}_N \boldsymbol{\Sigma}_{dd} \mathbf{W}_N \mathbf{y}'_N,$$

where $\mathbf{y}_N = (y_1, y_2, \ldots, y_N)$, $\mathbf{W}_N = \text{diag}(w_1, w_2, \ldots, w_N)$ is a diagonal matrix whose diagonal elements starting in the upper left corner are $w_1, w_2, \ldots, w_N$, and $\boldsymbol{\Sigma}_{dd}$ is the covariance matrix of $\mathbf{d}$.

2. (Section 1.2.4) Derive the joint probabilities of selection for systematic samples of size 3 selected from the population of Table 1.1 with the measures of size of Table 1.1.

3. (Section 1.2.4) Assume that a population satisfies

$$y_t = \sin 2\pi k^{-1} t$$

for $t = 1, 2, \ldots, N$. Give the variance of the sample mean of a systematic sample of size 10 as an estimator of the population mean for a population with $k = 6$ and $N = 60$. Compare this to the variance of the mean of a simple random nonreplacement sample and to the variance of the mean of a stratified sample, where the population is divided into

two equal-sized strata with the smallest 30 indices in the first stratum. How do the results change if the sample size is 12?

4. (Section 1.2.3) Let a stratified population be of the form described in Section 1.2.3 with H strata of sizes $N_1, N_2, \ldots, N_H$. Find the optimal allocation to strata to estimate the linear function

$$\theta = \sum_{h=1}^{H} \alpha_h \bar{y}_{Nh},$$

where α_h, $h = 1, 2, \ldots, H$, are fixed constants. Assume equal costs for observations in the strata.

5. (Section 1.2, 1.3) Consider the following sampling scheme. A simple random sample of n households is selected from N households. The ith household contains M_i family members. In each household selected, one family member is selected at random and interviewed. Give the probability that person ij (the jth person in the ith household) is interviewed. Define the Horvitz–Thompson estimator of the total of y. Give the joint probability that any two people appear in the sample. Is it possible to construct an unbiased estimator of the variance of the Horvitz–Thompson estimator?

Consider the estimator of the variance,

$$\hat{V}\{\bar{y}_\pi \mid \mathcal{F}_N\} = n(n-1)^{-1} \left(\sum_{t=1}^{n} M_t \right)^{-2} \sum_{t=1}^{n} M_t^2 (y_{tj} - \bar{y}_\pi)^2,$$

where

$$\bar{y}_\pi = \left(\sum_{t=1}^{n} M_t \right)^{-1} \sum_{t=1}^{n} M_t y_{tj}$$

and y_{ij} is the value observed for person ij. Assume that the household size satisfies $1 \leq M_t \leq K$ for some K and assume that the finite population is a random sample from a superpopulation, where (y_{tj}, M_t) has a distribution with finite fourth moments and (y_{tj}, M_t) is independent of (y_{ij}, M_i) for $t \neq i$. Show that

$$n[\hat{V}\{\bar{y}_\pi \mid \mathcal{F}_N\} - V\{\bar{y}_\pi \mid \mathcal{F}_N\}] = o_{p(1)} \quad \text{a.s.}$$

as $N \to \infty$, $n \to \infty$, and $N^{-1}n \to 0$, where

$$V\{\bar{y}_\pi \mid \mathcal{F}_N\} = E\left\{ \left[\bar{y}_\pi - \left(\sum_{i=1}^{N} M_i \right)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \right]^2 \middle| \mathcal{F}_N \right\} \quad \text{a.s.}$$

6. (Section 1.3.3) Consider a population of x values with $x_i > 0$ for all i. Let model (1.3.58) hold with $E\{\bar{y}_N \mid \bar{x}_N\} = \bar{x}_N$, $\sigma^2 = 1$, and $\alpha = 2$. Let samples be selected from a finite population generated by (1.3.58) with π_i proportional to x_i. Using the approximate design variances, for what values of β_0 and β_1 is $V\{\bar{y}_\pi \mid \mathcal{F}\} < V\{\bar{y}_{HT} \mid \mathcal{F}\}$? You may consider the set of finite populations with common $(x_1, x_2, \ldots, x_N)$.

7. (Section 1.2.4) In Section 1.2.4 it is stated that the sample mean for a systematic sample with equal probabilities for samples of unequal sizes is biased for the population mean. Derive the bias and construct an unbiased estimator of the population mean. Assign probabilities to the two types of samples so that the sample mean is unbiased for the population mean.

8. (Section 1.2) Consider a population of size 9 that has been divided into two strata of size 4 and 5, respectively. Assume that a stratified sample of size 5 is to be selected, with two in stratum 1 and three in stratum 2. Let $\mathbf{d}$ be the nine-dimensional vector of indicator variables defined in (1.2.4). Give the mean and covariance matrix of $\mathbf{d}$.

9. (Section 1.2.5) Assume that a replacement sample of size 3 is selected from a population of size N with draw probabilities $(p_1, p_2, \ldots, p_N)$.

   (a) What is the probability that element i appears in the sample three times?

   (b) What is the probability that element i is observed given that the sample contains only one distinct unit?

   (c) What is the probability that element i is selected twice?

   (d) What is the probability that element i is selected twice given that some element was selected twice?

   (e) What is the probability that element i appears in the sample at least once?

   (f) What is the probability that element i appears in the sample given that the sample contains three distinct units?

10. (Section 1.2) Consider a design for a population of size N, where the design has $N+1$ possible samples. N of the samples are of size 1, where each sample contains one of the possible N elements. One sample is of size N, containing all elements in the population. Each of the $N+1$ possible samples is given an equal probability of selection.

    (a) What is the probability that element j is included in the sample?

(b) What is the expected sample size?

(c) What is the variance of the sample size?

(d) If the finite population is a realization of $NI(0,\ \sigma^2)$ random variables, what is the variance (over all populations and samples) of $\hat{T}_y - T_y$, where $\hat{T}_y$ is the Horvitz–Thompson estimator of the finite population total?

(e) Compare the variance of the estimated total of part (d) with the variance of $N(\bar{y}_n - \bar{y}_N)$ for a simple random sample of size n.

(f) Consider the estimator that conditions on sample size

$$\tilde{T}_y \ = \ Nn^{-1}\sum_{i\in A} y_i,$$

where n is the realized sample size. Show that this estimator is design unbiased for T_y.

(g) Give the variance of $\hat{T}_y - T_y$ of part (f) under the conditions of part (d).

11. (Section 1.2) Assume an R-person list, where the ith person appears on the list r_i times. The total size of the list is N. Assume that a simple random nonreplacement sample of n lines is selected from the list. For each line selected, a person's characteristic, denoted by y_i, the total number of lines for person i, denoted by r_i, and the number of times that person i occurs in the sample, denoted by t_i, are determined. Assume that r_i is known only for the sample.

(a) Give an estimator for the number of people on the list.

(b) Give an estimator for the total of y.

12. (Section 1.2) The possible samples of size 3 selected from a population of size 5 are enumerated in Table 1.3. The table also contains probabilities of selection for a particular design.

(a) Compute the probabilities of selection, π_i, for $i = 1,\ 2,\ \ldots,\ 10$.

(b) Compute the joint probabilities of selection, π_{ij}, for all possible pairs.

(c) Compute the joint probability of selection for each pair of pairs.

(d) Assume that a population of finite populations of size 5 is such that each finite population is a sample of 5 $NI(\mu,\ \sigma^2)$ random variables. What are the mean and variance of the Horvitz–Thompson

Table 1.3 Design for Samples of Size 3 from a Population of Size 5

Sample	Sample Elements	Prob. of Sample	Sample	Sample Elements	Prob. of Sample
1	1,2,3	0.06	6	1,4,5	0.10
2	1,2,4	0.07	7	2,3,4	0.11
3	1,2,5	0.08	8	2,3,5	0.12
4	1,3,4	0.09	9	2,4,5	0.13
5	1,3,5	0.10	10	3,4,5	0.14

estimator of the total of the finite population when the sample is selected according to the design of the table?

(e) Under the assumptions of part (d), find the mean and variance of the variance estimator (1.2.33).

(f) Under the assumptions of part (d), find the mean and variance of $\hat{\theta}_k - \bar{y}_N$, $k = 1, 2$, where

$$\hat{\theta}_1 = N^{-1} \sum_{i \in A} \pi_i^{-1} y_i$$

and

$$\hat{\theta}_2 = \left(\sum_{i \in A} \pi_i^{-1} \right)^{-1} \left(\sum_{i \in A} \pi_i^{-1} y_i \right).$$

13. (Section 1.2) [Sirken (2001)] Let a population be composed of N units with integer measures of size m_i, $i = 1, 2, \ldots, N$. Let $M_0 = 0$ and let $M_j = \Sigma_{i=1}^{j} m_i$, $j = 1, 2, \ldots, N$. Consider two sampling procedures:

(a) A replacement simple random sample of n integers is selected from the set $\{1, 2, \ldots, M_N\}$. If the selected integer, denoted by k, satisfies

$$M_{i-1} < k \leq M_i,$$

element i is in the sample.

(b) A nonreplacement simple random sample of n integers is selected from the set $\{1, 2, \ldots, M_N\}$. The rule for identifying selected units is the same as for procedure (a).

For each procedure:

i. Determine the probability that element i is selected for the sample exactly once.

ii. Determine the probability that element i is selected for the sample at least once.

iii. Determine the joint probability that elements i and k appear together in the sample.

14. (Section 1.2.2) Let a finite population of size N be a random sample from an infinite population satisfying the model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + e_i, \\ e_i &\sim NI(0,\ x_i\sigma^2), \end{aligned}$$

where x_i is distributed as a multiple of a chi-square random variable with d, $d \geq 3$, degrees of freedom, and e_i is independent of x_j for all i and j. Let a Poisson sample of expected size n_B be selected with the selection probability for element i proportional to x_i. What is the expected value of

$$n_B^{-1} \sum_{i \in U} x_i \left(\sum_{i \in A} x_i^{-1},\ \sum_{i \in A} y_i,\ \sum_{i \in A} x_i^{-2} y_i,\ \sum_{i \in A} x_i^{-1} y_i \right)?$$

15. (Section 1.2.2) Assume that a sample of n elements is selected using Poisson sampling with probabilities π_i, $i = 1, 2, \ldots, N$. Find the design variance of the linear function

$$\hat{\theta} = \sum_{i \in A} g_i y_i,$$

where the g_i, $i = 1, 2, \ldots, N$, are fixed coefficients. Determine an estimator of the variance of $\hat{\boldsymbol{\theta}}$.

16. (Section 1.3) Let a sequence of populations of size N be selected from a distribution with mean μ and variance σ^2. Give an example of a sequence $(N,\ n_N)$ such that

$$N^{-1} \left(\sum_{i \in A} y_i + (N - n_N) n_N^{-1} \sum_{i \in A} y_i - \sum_{i \in U} y_i \right) = O_p(n_N^{-\beta}),$$

where $\beta > 0.5$.

17. (Section 1.3.3) Let $y_i \sim NI(0,\ 1)$ and define x_i by

$$\begin{aligned} x_i &= y_i && \text{with probability } 0.5 \\ &= -y_i && \text{with probability } 0.5, \end{aligned}$$

where the event defining x_i is independent of y_i. Let $(\bar{x},\ \bar{y}) = n^{-1}\sum_{i=1}^{n}(x_i,\ y_i)$.

(a) Prove that

$$n^{1/2}(\bar{x},\ \bar{y}) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{I}).$$

(b) Does the conditional distribution of $\bar{x}$ given $\bar{y}$ converge to a normal distribution almost surely?

(c) Let

$$\begin{aligned} \bar{z} &= \bar{y} && \text{with probability } 0.5 \\ &= -\bar{y} && \text{with probability } 0.5, \end{aligned}$$

where the event defining $\bar{z}$ is independent of $\bar{y}$. Show that $\bar{z}$ is a normal random variable. Is the conditional distribution of $\bar{z}$ given $\bar{y}$ normal?

18. (Section 1.3) Assume that a finite population of size N is a realization of N binomial trials with probability of success equal to p. Let the finite population proportion be p_N. Assume that a sample of size n is selected with replacement from the finite population. Show that the variance of the sample proportion, $\hat{p}$, as an estimator of the infinite population proportion is

$$V\{\hat{p} - p\} = N^{-1}n^{-2}\left[(n-1)N + n^2\right]p(1-p),$$

where $\hat{p}$ is the replacement estimator of the mean obtained by dividing the estimated total (1.2.66) by N.

19. (Section 1.2.2) Assume that a Poisson sample is selected with known probabilities π_i, where the π_i differ. Let n_B be the expected sample size. Find the design mean and variance of the estimator

$$\hat{T}_y = Nn_B^{-1}\sum_{i\in A} y_i.$$

Is it possible to construct a design-unbiased estimator of the design variance of $\hat{T}_y$?

20. (Section 1.2) Assume that a simple random sample of size n is selected from a population of size N and then a simple random sample of size m

is selected from the remaining $N - n$. Show that the $n + m$ elements constitute a simple random sample from the population of size N. What is $C\{\bar{y}_n, \bar{y}_m \mid \mathcal{F}\}$, where $\bar{y}_n$ is the mean of the first n elements and $\bar{y}_m$ is the mean of the second m elements?

21. (Section 1.3) Show that the assumptions

    (a) $K_L < \pi_i < K_U$

    (b) $\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \mid y_i \mid^{2+\delta} = M_{2\delta} > 0$

    for positive constants δ, K_L and K_U are sufficient for

$$\lim_{N\to\infty} \sup_{l\le k\le N} \left[\sum_{i=1}^{N} y_i^2 \pi_i (1 - \pi_i)\right]^{-1} y_k^2 = 0.$$

22. (Section 1.3) Consider a sequence of finite populations composed of H_N strata. Assume that random samples of size 2 are selected from each stratum. Do the μ_h need to be bounded for the results of Theorem 1.3.2 to hold?

23. (Section 1.2.2) Assume that the data in Table 1.4 are a Poisson sample selected with the probabilities given in the table.

    (a) Estimate the fraction of managers who are over 50. Estimate the variance of your estimator.

    (b) Estimate the fraction of employees who have a manager over 50. Estimate the variance of your estimator.

    (c) Estimate the population covariance between age of manager and number of employees for the population of managers.

24. (Section 1.3.3) Consider the estimator $\bar{\pi}_N \hat{R}_{\pi y}$, where

$$\hat{R}_{\pi y} = \left(\sum_{i\in A} \pi_i^{-1} \pi_i\right)^{-1} \sum_{i\in A} \pi_i^{-1} y_i.$$

    The denominator of the ratio is n, but the summation expression emphasizes the fact that n is an estimator of $N\bar{\pi}_N$. Thus, under the assumptions of Theorem 1.3.8, $\bar{\pi}_N \hat{R}_{\pi y} = \bar{y}_N + O_p(n_{BN}^{-1/2})$. Find $E\{\bar{\pi}_N \hat{R}_{\pi y} \mid \boldsymbol{\pi}_n\}$ under model (1.3.58), where $\boldsymbol{\pi}_n$ is the set of π_i in A. Assume that $\alpha = 2$ and $\beta_0 = 0$ in model (1.3.58). Find the best linear unbiased estimator

Table 1.4 Poisson Sample of Managers

Probability	Age of Manager	Number of Employees	Probability	Age of Manager	Number of Employees
0.016	35	10	0.070	50	31
0.016	36	4	0.100	51	47
0.036	41	15	0.100	56	55
0.040	45	25	0.120	62	41
0.024	45	8	0.100	64	50

of β_1, conditional on π_n. What is the best linear unbiased predictor of T_y?

25. (Section 1.3.2) In the proof of Theorem 1.3.4 we demonstrated that for a sequence of Bernoulli samples there is a corresponding sequence of simple random samples such that the difference between the two means is $O_p(n_B^{-3/2})$. Given a sequence of simple random samples, construct a corresponding sequence of Bernoulli samples such that the difference between the two means is $O_p(n_B^{-3/2})$.

26. (Section 1.3.1) Prove the following.

    Result. Let $\{X_n\}$ be a sequence of random variables such that

$$\begin{aligned} E\{X_n\} &= 0, \\ V_n\{X_n\} &= O_p(n^{-\alpha}), \end{aligned}$$

    where $V_n\{X_n\}$ is the sequence of variances of X_n and $\alpha > 0$. Then

$$X_n = O_p(n^{-0.5\alpha}).$$

27. (Section 1.2.5) Let a population of size N be given and denoted by $\mathcal{F}_N$. Let a second finite population of size nN be created by replicating each of the original observations n times. Denote the second population by $\mathcal{F}_{nN}$. Is $V\{\bar{x}_{rn} - \bar{x}_N \mid \mathcal{F}_N\}$ for an equal probability replacement sample of size n selected from $\mathcal{F}_N$ the same as the variance of $V\{\bar{x}_n - \bar{x}_N \mid \mathcal{F}_{nN}\}$ for a simple random nonreplacement sample of size n selected from $\mathcal{F}_{nN}$? The statistic $\bar{x}_{rN}$ for the replacement sample is the mean of the y for the n draws, not the mean of distinct units.

28. (Section 1.2.3) Show that the estimator (1.2.56) is the Horvitz–Thompson variance estimator.

29. (Section 1.2.1) Show that

$$\begin{aligned} S^2 &= (N-1)^{-1}\sum_{i\in U}(y_i - \bar{y}_N)^2 \\ &= 0.5N^{-1}(N-1)^{-1}\sum_{i\in U}\sum_{j\in U}(y_i - y_j)^2. \end{aligned}$$

Hence, show that expression (1.2.28) for simple random sampling is

$$V\{N\bar{y}_n \mid \mathcal{F}\} = N^2(n^{-1} - N^{-1})S^2.$$

30. (Section 1.2.6) In the example in the text, the selection of a sample of size 3 from a population of size 6 led to selection probabilities of (9/16, 8/16, 7/16, 7/16, 8/16, 9/16). What is the joint selection probability of units 1 and 2? Of units 3 and 4? Of units 1 and 6?

31. (Section 1.3.1) Prove:

Lemma 1.6.1. If $\hat{\theta}_n = B_n + o_p(|\ \hat{\theta}_n\ |)$, then $\hat{\theta}_n = O_p(\ |B_n|\ )$.

32. (Section 1.2.1) Let A_1 be the indexes of a simple random nonreplacement sample of size n_1 selected from a finite population of size N. Let A_2 be the indexes of a simple random sample of size n_2 selected from the remaining $N - n_1$ elements. Let $\bar{y}_1$ be the mean for sample A_1 and $\bar{y}_2$ be the mean for sample A_2. What is $C\{\bar{y}_1, \bar{y}_2 \mid \mathcal{F}\}$?

33. (Section 1.4) Let a population of size N have assigned probabilities $(p_1, p_2, \ldots, p_N)$ and consider the following successive selection scheme. At step 1 a unit is selected from the N units with probability p_i. At step 2 a unit is selected from the remaining $N-1$ units with probability $p_j(1-p_i)^{-1}$. What is the probability that unit j is in a sample of size 2? What is the probability that units j and k will be in a sample of size 2? Rosen (1972) has studied this selection scheme for n selections.

34. (Section 1.3.2) In Theorem 1.3.4 it is asserted that

$$E\{(n_o^{-1} - n_B^{-1})^2 \mid n_o > 0\} = O(n_B^{-3}).$$

Show that the conditions of Theorem 5.4.4 of Fuller (1996) are satisfied for $n_o^{-2}n_B^2$ and hence that the conditions of Theorem 5.4.3 of Fuller (1996) are satisfied for $(n_o^{-1}n_B - 1)^2$. *Hint*: Let x of Theorem 5.4.4 be $n_B^{-1}n_o$.

35. (Section 1.3) Prove:

Theorem 1.6.1. Let $\{y_i\}$ be a sequence of fixed numbers and let $\mathcal{F}_N = \{y_1, y_2, \ldots, y_N\}$. Assume that

$$\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} y_i = \mu$$

and

$$\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} | y_i |^{1+\epsilon} = K_\epsilon$$

for some $\epsilon > 0$, where μ and K_ϵ are finite. Let $\{\pi_i\}$ be a sequence of probabilities with $0 < c_s < \pi_i < c_g < 1$. Let a sequence of Poisson samples be defined with selection probabilities π_i, where $A_{N-1} \subseteq A_N$. Then

$$\lim_{N\to\infty} N^{-1} \sum_{i\in A_N} \pi_i^{-1} y_i = \mu \text{ a.s.} \tag{1.6.1}$$

and

$$\lim_{N\to\infty} \left(\sum_{i\in A_N} \pi_i^{-1} \right)^{-1} \sum_{i\in A_N} \pi_i^{-1} y_i = \mu \text{ a.s.} \tag{1.6.2}$$

36. (Section 1.3) Let two finite populations of size N_1 and N_2 be realizations of iid random variables from a distribution $F(y)$. Let a simple random sample of size n_1 be selected from N_1 and a simple random sample of size n_2 be selected from N_2. Show that the sample of $n_1 + n_2$ elements can be treated as a simple random sample from the population of size $N_1 + N_2$.

37. (Section 1.2.7) Let $y_1, y_2, \ldots, y_n$ be independent random variables with $y_i \sim (\mu, \sigma_i^2)$. Show that

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

38. (Section 1.3) Prove:

Result 1.6.1. Let $y_1, y_2, \ldots,$ be a sequence of real numbers. Let $\mathcal{F}_N = (y_1, y_2, \ldots, y_N)$ be a sequence of populations, and let (1.3.20) and

(1.3.21) hold. Let a sequence of samples be selected with probabilities π_i and joint probabilities $\pi_{ij,N}$, where $\pi_{ij,N} \leq \pi_i \pi_j$ for all i and j in $U_N, i \neq j$, and all N. Then

$$V\{N^{-1}(\hat{T}_y - T_y) \mid \mathcal{F}_N\} = O(n_{BN}^{-1}),$$

where n_{BN} is the expected sample size for population N and $\hat{T}_y$ is the Horvitz–Thompson estimator of the total.

39. (Section 1.4) Let a sample of size n be selected in the following way. The first element is selected with probability p_i, where $\Sigma_{i=1}^{N} p_i = 1$. Then $n-1$ elements are selected as a simple random sample from the remaining $N-1$ elements. What is the total probability, π_i, that element i is included in the sample? What is the probability that elements i and j appear in the sample? What is the probability that the n elements $i_1, i_2, \ldots, i_n$ form the sample? See Midzuno (1952).

40. (Section 1.2.8) Assume simple random sampling at each of the two stages of a two-stage sample. Are there population configurations such that $\pi_{(ij)(km)}$ of (1.2.76) is the same for all ij and km, $ij \neq km$?

41. (Section 1.3.1) Consider a sequence of populations $\{\mathcal{F}_N\}$ created as the first N elements of the sequence $\{y_1, y_2, \ldots\}$. Assume that the $\mid y_i \mid$ are bounded and that S_{yN}^2 converges to a positive quantity. Let a systematic sample be selected from the Nth population with a rate of K^{-1} for all N. Is $\bar{y}_n$ design consistent for $\bar{y}_N$? Explain.

42. (Section 1.2.4) Assume that a population of size 10 is generated by the autoregressive model of (1.2.61) with $\rho = 0.9$. Assume that a systematic sample of size 2 is selected and that the selected elements are $A = [i, j]$. Give

$$V\{\bar{y}_2 - \bar{y}_N \mid A = [2, 7]\}$$

and

$$V\{\bar{y}_2 - \bar{y}_N \mid A = [3, 8]\}.$$

Derive the variance of the best linear unbiased predictor of $\bar{y}_N$ for each of the situations $A = [2, 7]$ and $A = [3, 8]$. Give the variance of the predictor.

43. (Section 1.2) Let a population of size $2N$ be divided into two groups of size N. Let a group be selected at random (with probability equal to one-half), and let one unit be selected at random from the selected group. Let a simple random sample of size 2 be selected from the other

group to form a sample of size 3 from the original population. Give the inclusion probabilities and joint inclusion probabilities for this selection scheme.

44. (Section 1.2.5) Assume that a replacement sample of size 2 is selected from a population of size 4 with probability $p_i = 0.25$ at each draw. What is the relative efficiency of estimator (1.2.66) to estimator (1.2.71)?

45. (Section 1.2.1) Let $y_i, i = 1, 2, \ldots, n$, be independent (μ, σ_i^2) random variables and let

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i.$$

Show that

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

is unbiased for $V\{\bar{y}\}$.

46. (Section 1.3.1) Let $\hat{\theta} = \Sigma_{i=1}^{n} w_i y_i$, where $\Sigma_{i=1}^{n} w_i^4 = O(n^{-3})$, $\Sigma_{i=1}^{n} w_i = 1$ for all n, the w_i are fixed, and the $y_i, i = 1, 2, \ldots$, are independent (μ, σ_i^2) random variables with bounded fourth moments. Show that

$$E\{\hat{V}(\hat{\theta})\} = V\{\hat{\theta}\} + O(n^{-2}),$$

where

$$\hat{V}\{\hat{\theta}\} = \left(1 - \sum_{i=1}^{n} w_i^2\right)^{-1} \sum_{i=1}^{n} w_i^2 (y_i - \hat{\theta})^2.$$

47. (Section 1.3) Let $(y_1, y_2, \ldots, y_n)$ be a simple random sample from a population with $y_i > 0$ for all i and finite fourth moment. Let an estimator of the coefficient of variation be

$$\hat{\theta} = \bar{y}^{-1} s_y,$$

where $\theta = \bar{y}_N S_{y,N}$ is the coefficient of variation. Using a Taylor expansion, find the variance of the approximate distribution of $n^{0.5}(\hat{\theta} - \theta)$.

48. (Section 1.2) Let $(y_1, y_2, \ldots, y_N) = \mathbf{y}'$ be a vector of $iid(\mu, \sigma^2)$ random variables. What are the conditional mean and covariance matrix of $\mathbf{y}$ conditional on $\Sigma_{i=1}^{N} y = T$? What are the conditional mean and covariance matrix of $\mathbf{y}$ conditional on (T, S^2), where

$$S^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{y}_N)^2?$$

49. (Section 1.3) To evaluate M_5 of Theorem 1.3.2, show that for a population with zero mean and finite fourth moment,

$$E\{\bar{y}^4\} = n^{-3}[E\{y^4\} + 3(n-1)\sigma^4]$$

and

$$C\{\sum_{i=1}^{n} y_i^2,\ n\bar{y}^2\} = E\{y^4\} - \sigma^4.$$

Hint: See Fuller (1996, p. 241).

50. (Section 1.3) In Exercise 13, procedure (b) consisted of a nonreplacement sample of n integers. Let an estimator of the total of y be

$$\hat{T}_{y,r} = n^{-1}T_m^{-1}\sum_{d=1}^{n} m_d^{-1}y_d,$$

where d is the index for draw, (m_d, y_d) is the (measure of size, y value) obtained on the dth draw, and the vector of totals is

$$(T_m, T_y) = \sum_{i=1}^{N}(m_i, y_i).$$

(a) Show that $\hat{T}_{y,r}$ is design unbiased for T_y and give an expression for $V\{\hat{T}_{y,r} - T_y \mid \mathcal{F}\}$. Give an estimator of $V\{\hat{T}_{y,r} - T_y \mid \mathcal{F}\}$. *Hint*: Let $z_i = m_i^{-1}y_i$ and consider the population composed of m_1 values of $m_1^{-1}y_1, m_2$ values of $m_2^{-1}y_2, \ldots, m_N$ values of $m_N^{-1}y_N$.

(b) Consider a sequence of pairs of real numbers $\{m_i, y_i\}$, where the m_i are positive integers. Assume that:

(i) The m_i are bounded.

(ii) $\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} y_i = \mu_y.$

(iii) $\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2 = \sigma_y^2 > 0.$

(iv) $\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} |\, y_i - \bar{y}_N \,|^{2+\delta} = K_3 < \infty.$

Show that

$$n^{1/2}(N^{-1}\hat{T}_{y,r} - \bar{y}_N) \xrightarrow{\mathcal{L}} N(0, V_{11}),$$

where

$$V_{11} = \lim_{N\to\infty} nT_m^{-2}V\{\hat{T}_{y,r} - T_y \mid \mathcal{F}_N\}.$$

1.7 APPENDIX 1A: SOME ORDER CONCEPTS

There are exact distributional results available for only a few of the statistics associated with survey sampling. For most, approximations based on large-sample theory are required. Concepts of relative magnitude or *order of magnitude* are useful in deriving those approximations. The following material is from Fuller (1996, Chapter 5). Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be sequences of real numbers, let $\{f_n\}_{n=1}^{\infty}$ and $\{g_n\}_{n=1}^{\infty}$ be sequences of positive real numbers, and let $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$ be sequences of random variables.

Definition 1.7.1. We say that a_n is of smaller order than g_n and write

$$a_n = o(g_n)$$

if

$$\lim_{N \to \infty} g_n^{-1} a_n = 0.$$

Definition 1.7.2. We say that a_n is at most of order g_n and write

$$a_n = O(g_n)$$

if there exists a real number M such that $g_n^{-1} \mid a_n \mid \leq M$ for all n.

Using the definitions of order and the properties of limits, one can prove:

1. If $a_n = o(f_n)$ and $b_n = o(g_n)$, then

$$\begin{aligned} a_n b_n &= o(f_n g_n), \\ \mid a_n \mid^s &= o(f_n^s) \quad \text{for } s > 0, \\ a_n + b_n &= o(\max\{f_n, g_n\}). \end{aligned}$$

2. If $a_n = O(f_n)$ and $b_n = O(g_n)$, then

$$\begin{aligned} a_n b_n &= O(f_n g_n), \\ \mid a_n \mid^s &= O(f_n^s) \quad \text{for } s \geq 0, \\ a_n + b_n &= O(\max\{f_n, g_n\}). \end{aligned}$$

3. If $a_n = o(f_n)$ and $b_n = O(g_n)$, then

$$a_n b_n = o(f_n g_n).$$

The concepts of order for random variables, introduced by Mann and Wald (1943), are closely related to *convergence in probability*.

Definition 1.7.3. The sequence of random variables $\{X_n\}$ converges in probability to the random variable X, written

$$p\lim X_n = X$$

(the *probability limit* of X_n is X), if for every $\epsilon > 0$

$$\lim_{n\to\infty} P\{|\, X_n - X \,| > \epsilon\} = 0.$$

Definition 1.7.4. We say that X_n is of smaller order in probability than g_n and write

$$X_n = o_p(g_n)$$

if

$$p\lim g_n^{-1} X_n = 0.$$

Definition 1.7.5. We say that X_n is at most of order in probability g_n (or bounded in probability by g_n) and write

$$X_n = O_p(g_n)$$

if for every $\epsilon > 0$ there exists a positive real number M_ϵ such that

$$P\{|\, X_n \,| \geq M_\epsilon g_n\} \leq \epsilon$$

for all n.

Analogous definitions hold for vectors.

Definition 1.7.6. If $\mathbf{X}_n$ is a k-dimensional random variable, $\mathbf{X}_n$ is at most of order in probability g_n and we write

$$\mathbf{X}_n = O_p(g_n)$$

if for every $\epsilon > 0$ there exists a positive real number M_ϵ such that

$$P\{|\ X_{jn}\ | \geq M_\epsilon g_n\} \leq \epsilon, \qquad j = 1, 2, \ldots, k,$$

for all n.

Definition 1.7.7. We say that X_n is of smaller order in probability than g_n and write

$$\mathbf{X}_n = o_p(g_n)$$

if for every $\epsilon > 0$ and $\delta > 0$ there exists an N such that for all $n > N$,

$$P\{|\ X_{jn}\ | > \epsilon g_n\} < \delta, \qquad j = 1, 2, \ldots, k.$$

Order operations for sequences of random variables are similar to those for sequences of real numbers; thus:

1. If $X_n = o_p(f_n)$ and $Y_n = o_p(g_n)$, then

$$\begin{aligned} X_n Y_n &= o_p(f_n g_n), \\ |\ X_n\ |^s &= o_p(f_n^s) \quad \text{for } s > 0, \\ X_n + Y_n &= o_p(\max\{f_n, g_n\}). \end{aligned}$$

2. If $X_n = O_p(f_n)$ and $Y_n = O_p(g_n)$, then

$$\begin{aligned} X_n Y_n &= O_p(f_n g_n), \\ |\ X_n\ |^s &= O_p(f_n^s) \quad \text{for } s \geq 0, \\ X_n + Y_n &= O_p(\max\{f_n, g_n\}). \end{aligned}$$

3. If $X_n = o_p(f_n)$ and $Y_n = O_p(g_n)$, then

$$X_n Y_n = o_p(f_n g_n).$$

One of the most useful tools for establishing the order in probability of random variables is *Chebyshev's inequality*.

Theorem 1.7.1. Let $r > 0$ and let X be a random variable such that $E\{|\ X\ |^r\} < \infty$. Then for every $\epsilon > 0$ and finite A,

$$P\{|\ X - A\ | \geq \epsilon\} \leq \frac{E\{|\ X - A\ |^r\}}{\epsilon^r}.$$

It follows from Chebyshev's inequality that any random variable with finite variance is bounded in probability by the square root of its second moment about the origin.

Corollary 1.7.1.1. If $\{X_n\}$ is a sequence of random variables such that

$$E\{X_n^2\} = O(a_n^2),$$

then

$$X_n = O_p(a_n).$$

This Page Intentionally Left Blank

CHAPTER 2

USE OF AUXILIARY INFORMATION IN ESTIMATION

Information available at the estimation stage beyond that in the sample is called *auxiliary information.* Such information can be placed into two categories: (1) knowledge of population totals, or means, of characteristics that are observed on the elements of the sample but not on all elements of the population, and (2) knowledge of characteristics for every element in the population. As an example of the first situation, the age distribution of the population of Iowa may be treated as known on the basis of a recent census, but the age of people in a sample of households is not known until the households are contacted, and the age of nonsampled persons is unknown. An example of a characteristic known for all households in the population is the geographic location of the households on an address list. Information available for every sampling unit can be used at both the design and estimation stages. Information available only at the population level, but not for the sampling frame, can be used only at the estimation stage.

Sampling Statistics. By Wayne A. Fuller

2.1 RATIO ESTIMATION

Ratio estimation of population totals is one of the oldest uses of auxiliary information in survey sampling. For example, if a government agency reports the acres planted to corn, and we conduct a survey of farmers later in the season, it is very natural to multiply the yield per acre obtained in our survey by the government report of acres planted to obtain an estimate of total production.

Assume that the vector $(y_i,\ x_i)$ is observed on a sample of elements and that the population mean of x is known and not zero. Then a ratio estimator of the mean of y is

$$\bar{y}_{rat} = \bar{y}_\pi \bar{x}_\pi^{-1} \bar{x}_N, \tag{2.1.1}$$

where $(\bar{y}_\pi,\ \bar{x}_\pi)$ is a design consistent estimator of the mean of $(y,\ x)$. Note that $\bar{y}_\pi \bar{x}_\pi^{-1} = \bar{y}_{HT} \bar{x}_{HT}^{-1}$. By the results of Theorem 1.3.7,

$$\begin{aligned} \bar{y}_{rat} - \bar{y}_N &= \bar{x}_\pi^{-1} \bar{x}_N (\bar{y}_\pi - R_N \bar{x}_\pi) \\ &= \bar{y}_\pi - R_N \bar{x}_\pi + O_p(n^{-1}), \end{aligned} \tag{2.1.2}$$

where $R_N = \bar{x}_N^{-1} \bar{y}_N$, we assume that $(\bar{y}_\pi,\ \bar{x}_\pi) - (\bar{y}_N,\ \bar{x}_N) = O_p(n^{-1/2})$, and assume that $\bar{x}_N \neq 0$. Thus, in large samples, the variance of the approximate distribution is $V\{\bar{y}_\pi - R_N \bar{x}_\pi \mid \mathcal{F}_N\}$ and the ratio estimator is superior to the design-consistent mean if

$$V\{\bar{y}_\pi - R_N \bar{x}_\pi \mid \mathcal{F}_N\} < V\{\bar{y}_\pi \mid \mathcal{F}_N\}. \tag{2.1.3}$$

For simple random sampling the inequality is

$$S_y^2 - 2R_N S_{xy} + R_N^2 S_x^2 < S_y^2 \tag{2.1.4}$$

or

$$R_N < 2\beta_N, \tag{2.1.5}$$

where $\beta_N = S_x^{-2} S_{xy}$ is the regression coefficient. It is easy to construct populations where the correlation is large and (2.1.5) is not satisfied.

It is possible to use a Taylor expansion to evaluate the first term in the bias of the estimated ratio. Assume that $x_i > 0$ for all i, that the moments of $(x,\ y)$ are bounded, and that the sample design is such that $V\{(\bar{x}_\pi,\ \bar{y}_\pi) \mid \mathcal{F}_N\} = O(n^{-1})$. Then, by a Taylor expansion about $(\bar{x}_N,\ \bar{y}_N)$,

$$\begin{aligned} \bar{x}_\pi^{-1} \bar{y}_\pi &= R_N + \bar{x}_N^{-1} (\bar{y}_\pi - R_N \bar{x}_\pi) \\ &\quad + \bar{x}_N^{-2} \left[R_N (\bar{x}_\pi - \bar{x}_N)^2 - (\bar{x}_\pi - \bar{x}_N)(\bar{y}_\pi - \bar{y}_N) \right] \\ &\quad + O_p(n^{-3/2}). \end{aligned} \tag{2.1.6}$$

The expected value of the ratio does not always exist, but it is still useful to evaluate the expected value of the leading terms on the right of (2.1.6) as a property of the approximating random variable. If the moments of $\bar{x}_\pi^{-1}\bar{y}_\pi$ are bounded, one can approximate the expectation to obtain

$$\begin{aligned} E\{\bar{x}_\pi^{-1}\bar{y}_\pi - R_N \mid \mathcal{F}_N\} &= \bar{x}_N^{-2}\left[R_N V\{\bar{x}_\pi \mid \mathcal{F}_N\} - C\{(\bar{y}_\pi,\ \bar{x}_\pi \mid \mathcal{F}_N)\}\right] \\ &\quad + O(n^{-2}) \quad \text{a.s.} \end{aligned} \tag{2.1.7}$$

If the regression coefficient for the sample means, defined by

$$\beta_N = [V\{\bar{x}_\pi \mid \mathcal{F}_N\}]^{-1} C\{\bar{y}_\pi,\ \bar{x}_\pi \mid \mathcal{F}_N\}, \tag{2.1.8}$$

is equal to R_N, the regression line passes through the origin, and the leading $O_p(n^{-1})$ term of (2.1.7) vanishes.

Mussa (1999) has reviewed modifications of the ratio estimator that have been suggested to reduce bias. If one is interested in the ratio and is concerned about bias, the estimator due to Beale (1962) often performs well. The estimator is

$$\hat{R}_B = [\bar{x}_\pi^2 + \hat{V}\{\bar{x}_\pi \mid \mathcal{F}_N\}]^{-1}[\bar{x}_\pi\bar{y}_\pi + \hat{C}\{\bar{x}_\pi,\ \bar{y}_\pi \mid \mathcal{F}_N\}]. \tag{2.1.9}$$

The estimator removes the $O(n^{-1})$ terms in the bias and performs well if $|\bar{x}_\pi|$ is small relative to the standard error of $\bar{x}_\pi$.

The implicit model underlying the ratio estimator is that the relationship between y and x is a straight line through the origin and that the variance about the line is proportional to x. Assume that the finite population is a sample from the infinite population satisfying the model

$$\begin{aligned} y_i &= x_i\beta + e_i, \\ e_i &\sim ind(0,\ x_i\sigma^2), \end{aligned} \tag{2.1.10}$$

where $x_i > 0$ for all $i \in U_N$. Then, given a sample, the best linear unbiased estimator of β is

$$\hat{\beta} = \left(\sum_{i\in A} x_i^2 x_i^{-1}\right)^{-1} \sum_{i\in A} x_i x_i^{-1} y_i = \left(\sum_{i\in A} x_i\right)^{-1} \sum_{i\in A} y_i. \tag{2.1.11}$$

Hence, under the model, the best linear unbiased estimator of $\beta\bar{x}_N$ is the ratio estimator $\hat{\beta}\bar{x}_N$. See also Exercise 18. If the sample is a simple random sample, the best linear unbiased estimator of the mean corresponds to estimator (2.1.11). This result has been a delight for theoreticians for decades. However, the application of ratio estimation in practice requires caution because inequality (2.1.5) fails to hold for many y-characteristics and will often fail for subpopulations.

The model (2.1.10) is most likely to be satisfied for cluster sampling or two-stage sampling. In cluster sampling the relevant variable for variance estimation is the sum of the element values for the cluster. If no elements are observed in the cluster, the cluster totals are zero for all variables. Similarly, if the element values of y and of x are modestly correlated with the number of elements in the cluster, the cluster total for y will be approximately proportional to the cluster total for x. Also, for a given cluster, the variance of the cluster total of y will be approximately proportional to the number of elements and hence approximately proportional to the total for x.

The ratio bias can become important in ratio estimation for stratified samples. Consider a stratified sample for a population in which the stratum means of x, denoted by $\bar{x}_{hN}$, are known. The *separate ratio estimator* is the estimator composed of the weighted sum of the stratum ratio estimators,

$$\bar{y}_{st,s} = \sum_{h=1}^{H} W_h \bar{y}_h \bar{x}_h^{-1} \bar{x}_{hN}, \tag{2.1.12}$$

where $W_h = N^{-1}N_h$. If H is large and the n_h small, the sum of the H stratum biases can make an important contribution to the mean square error. For large H and small n_h, it may be preferable to use the *combined ratio estimator*

$$\bar{y}_{st,c} = \bar{x}_{st}^{-1} \bar{x}_N \bar{y}_{st}, \tag{2.1.13}$$

where $(\bar{x}_{st}, \bar{y}_{st})$ is the vector of stratified estimators defined by (1.2.54). See Exercise 25.

Example 2.1.1. To illustrate ratio estimation we use a small sample of segments from the 1997 NRI sample for the state of Missouri. The segments in Table 2.8 of Appendix 2A are the slightly modified sample in a single county. The weight in the third column of the table is the inverse of the sampling rate and segment size of the fourth column is the area of the segment in acres. For this illustration the 80 segments are placed in three strata. The actual stratification is somewhat finer. All segments, except the next to last, have three points. That segment, with a size of 100 acres, has two points. The entry "federal" in the table is the fraction of points that are federally owned multiplied by segment size. The entries for the broaduses cultivated cropland and forest are defined in the same way.

The direct (Horvitz–Thompson) estimate of the total area in the county and the direct estimate of cultivated cropland in thousands of acres are

$$\begin{matrix} (\hat{T}_a, \hat{T}_{cc}) \\ \end{matrix} = \begin{matrix} (450.7, & 153.8), \\ (12.9) & (18.3) \end{matrix}$$

where the numbers in parentheses below the estimates are the estimated standard errors. The area of the county, $T_{a,N}$ is known to be 437,100 acres. Therefore, it is very natural to multiply all weights by the ratio 0.9692 so that the value for total acres calculated with the new weights is equal to the known value. This gives the ratio estimate for cultivated cropland,

$$\hat{T}_a^{-1} T_{aN} \hat{T}_{cc} = 0.9692(153.8) = 149.0,$$

with an estimated standard error of 17.9. The ratio model is reasonable for these data because the segment must have acres if it is to have acres of cultivated cropland. The estimated efficiency of the ratio estimator relative to the direct estimator is about 105%. The gain from ratio estimation is modest because of the small variation in segment sizes. The use of the ratio to adjust the weights has the advantage for the analyst that the modified weights give the correct total acreage for the county. ■■

In many sample surveys, there are potentially thousands of quantities to be estimated. For example, the questionnaires for the U.S. Current Population Survey contain about 100 items. When one realizes that responses are often classified by age, gender, race, and location, the number of potential estimates easily exceeds 1 million. Because it is not possible to develop an estimator that is optimal for each potential y-variable, it is standard survey practice to provide a general-purpose estimation scheme for totals. These estimators almost always are of the form

$$\hat{T}_y = \sum_{i \in A} w_i y_i, \tag{2.1.14}$$

where the weights w_i are not functions of the y_i's. Although linear in y, the estimators typically are not design linear because the w_i often depend on other attributes of the sample and of the population. The typical data set contains the y-variables and the weights. Estimated totals for any characteristic are given by (2.1.14), and estimated means are the ratios of two estimated totals where the denominator is the estimated number of units in the population. Estimators of the form (2.1.14) have the desirable property that they produce internally consistent estimators. That is, a two-way table with A and B as classifications has the same A marginals as a two-way table with A and C as classifications, and so on. The ratio estimator of this section and the regression estimator of the next section can be expressed in the form (2.1.14).

2.2 REGRESSION ESTIMATION

2.2.1 Simple random sampling under the normal model

To begin our discussion of regression estimation, assume that the finite population is composed of row vectors $(y_i,\ \mathbf{x}_i)$ that are realizations of random variables satisfying the model

$$\begin{aligned} y_i &= \mathbf{x}_i\boldsymbol{\beta} + e_i, \\ e_i &\sim NI(0,\ \sigma_e^2), \end{aligned} \tag{2.2.1}$$

where e_j is independent of $\mathbf{x}_i$ for all i and j. We assume that the first element of the $(k+1)$- dimensional vector $\mathbf{x}$ is identically equal to 1 and write

$$\mathbf{x}_i = (1,\ \mathbf{x}_{1,i}),$$

where the $\mathbf{x}_{1,i} \sim ii(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{xx})$, and $\sim ii$ denotes independent identically distributed.

We assume that $\boldsymbol{\Sigma}_{xx}$ is of rank k and that the mean of $\mathbf{x}$ for the finite population of N elements is known and denoted by $\bar{\mathbf{x}}_N$. The values of $\mathbf{x}$ for the individual elements are not known prior to sampling and after sampling are known only for the sampled elements. Assume that a simple random nonreplacement sample of size n is selected from N. Let $\mathbf{X}$ be the matrix of observations on $\mathbf{x}$,

$$\mathbf{X} = (\mathbf{x}_1',\ \mathbf{x}_2',\ \ldots,\ \mathbf{x}_n')',$$

let $\mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_n)'$, and let $\mathbf{e} = (e_1,\ e_2,\ \ldots,\ e_n)'$. Then the model for the sample can be written

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \\ \mathbf{e} &\sim N(\mathbf{0},\ \mathbf{I}\sigma_e^2), \end{aligned}$$

where we assume that

$$\mathbf{X}'\mathbf{X} = \sum_{i\in A} \mathbf{x}_i'\mathbf{x}_i$$

is positive definite. For $\mathbf{x}_i$ normally distributed and $n > k+1$, $\mathbf{X}'\mathbf{X}$ is positive definite except for a set of probability zero. To simplify the presentation, we assume throughout that inverses are defined.

Because the standard regression assumptions hold for the sample, given $\mathbf{X}$, the estimated regression vector

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \sum_{i\in A} \mathbf{x}_i' y_i \tag{2.2.2}$$

is the best estimator of the superpopulation parameter. The estimator $\hat{\boldsymbol{\beta}}$ has the properties

$$E\{\hat{\boldsymbol{\beta}} \mid \mathbf{X}\} = \boldsymbol{\beta} \tag{2.2.3}$$

and

$$V\{\hat{\boldsymbol{\beta}} \mid \mathbf{X}\} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i\right)^{-1} \sigma_e^2.$$

Under the normality assumption for $\mathbf{e}$, $\hat{\boldsymbol{\beta}}$ is normally distributed, conditional on $\mathbf{X}$. An unbiased estimator of σ_e^2 is

$$s_e^2 = (n-k-1)^{-1} \sum_{i \in A} (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}})^2 \tag{2.2.4}$$

and an unbiased estimator of $V\{\hat{\boldsymbol{\beta}} \mid \mathbf{X}\}$ is

$$\hat{V}\{\hat{\boldsymbol{\beta}} \mid \mathbf{X}\} = (\mathbf{X}'\mathbf{X})^{-1} s_e^2.$$

The mean of $\mathbf{x}$ for the unobserved elements is

$$\bar{\mathbf{x}}_{N-n} = (N-n)^{-1}(N\bar{\mathbf{x}}_N - n\bar{\mathbf{x}}_n)$$

and the best predictor of the mean of y for the unobserved $N-n$ elements is

$$\bar{y}_{N-n,reg} = \bar{\mathbf{x}}_{N-n} \hat{\boldsymbol{\beta}}. \tag{2.2.5}$$

Hence, the best predictor of the overall mean is

$$\bar{y}_{reg} = N^{-1} \left(\sum_{i \in A} y_i + (N-n) \bar{\mathbf{x}}_{N-n} \hat{\boldsymbol{\beta}} \right). \tag{2.2.6}$$

By least squares theory,

$$\sum_{i \in A} (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) \mathbf{x}_i = \mathbf{0},$$

and because the first element of $\mathbf{x}_i$ is always 1,

$$\sum_{i \in A} (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) = 0. \tag{2.2.7}$$

It follows that the predictor

$$\begin{aligned} \bar{y}_{reg} &= N^{-1}[n\bar{\mathbf{x}}_n \hat{\boldsymbol{\beta}} + (N-n)\bar{\mathbf{x}}_{N-n} \hat{\boldsymbol{\beta}}] \\ &= \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}. \end{aligned} \tag{2.2.8}$$

Often, the unit element of $\mathbf{x}_i$ is isolated and the model is written

$$y_i = \beta_0 + \mathbf{x}_{1,i}\boldsymbol{\beta}_1 + e_i, \tag{2.2.9}$$

where, as before, $\mathbf{x}_i = (1, \mathbf{x}_{1,i})$. Then the estimator (2.2.6) takes the form

$$\bar{y}_{reg} = \bar{y}_n + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})\hat{\boldsymbol{\beta}}_1, \tag{2.2.10}$$

where

$$\hat{\boldsymbol{\beta}}_1 = \left(\sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})'(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})\right)^{-1} \sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})'(y_i - \bar{y}_n).$$

The conditional variance of the predictor of $\bar{y}_{N-n}$, conditional on the sample $\mathbf{x}$-values and $\bar{\mathbf{x}}_N$, is

$$\begin{aligned} V\{\bar{y}_{N-n,reg} - \bar{y}_{N-n} \mid \mathbf{X}, \bar{\mathbf{x}}_N\} &= V\{\bar{\mathbf{x}}_{N-n}\hat{\boldsymbol{\beta}} - \bar{y}_{N-n} \mid \mathbf{X}, \bar{\mathbf{x}}_N\} \\ &= V\{\bar{\mathbf{x}}_{N-n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \bar{e}_{N-n} \mid \mathbf{X}, \bar{\mathbf{x}}_N\} \\ &= \bar{\mathbf{x}}_{N-n}\left(\sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\bar{\mathbf{x}}_{N-n}'\sigma_e^2 \\ &\quad + (N-n)^{-1}\sigma_e^2 \end{aligned} \tag{2.2.11}$$

and $V\{\bar{y}_{reg} - \bar{y}_N \mid \mathbf{X}, \bar{\mathbf{x}}_N\}$ is expression (2.2.11) multiplied by $(1 - f_N)^2$.

For an alternative derivation of the variance, we write

$$\begin{aligned} \bar{y}_{reg} - \bar{y}_N &= N^{-1}(N-n)\,(\bar{y}_{N-n,reg} - \bar{y}_{N-n}) \\ &= N^{-1}(N-n)[(\bar{\mathbf{x}}_{1,N-n} - \bar{\mathbf{x}}_{1,n})(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) + \bar{e}_n - \bar{e}_{N-n}] \\ &= (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) + N^{-1}(N-n)\,(\bar{e}_n - \bar{e}_{N-n}). \end{aligned} \tag{2.2.12}$$

It follows that an expression for the conditional variance of the predictor of $\bar{y}_N$ is

$$\begin{aligned} V\{\bar{y}_{reg} - \bar{y}_N \mid \mathbf{X}, \bar{\mathbf{x}}_N\} &= (1 - f_N)n^{-1}\sigma_e^2 \\ &\quad + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})V\{\hat{\boldsymbol{\beta}}_1 \mid \mathbf{X}\}(\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})', \end{aligned} \tag{2.2.13}$$

where

$$V\{\hat{\boldsymbol{\beta}}_1 \mid \mathbf{X}\} = \left(\sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})'(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})\right)^{-1}\sigma_e^2. \tag{2.2.14}$$

Because s_e^2 of (2.2.4) is an unbiased estimator of σ_e^2, an unbiased estimator of the conditional prediction variance is obtained by substituting s_e^2 of (2.2.4) for σ_e^2 in (2.2.13).

The unconditional variance of $\bar{y}_{reg} - \bar{y}_N$ is the expected value of (2.2.13), because the estimator is conditionally unbiased. Assume that $\mathbf{x}_{1i}$ is normally distributed with covariance matrix $\mathbf{\Sigma}_{xx}$. Then the expected value of the last term of (2.2.13) is

$$E\left\{(\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})\left(\sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})'(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})\right)^{-1}(\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})'\sigma_e^2\right\}$$
$$= (1 - f_N)n^{-1}(n - k - 2)^{-1}k\sigma_e^2, \quad (2.2.15)$$

where k is the dimension of $\mathbf{x}_{1,i}$. Thus, under the normal model, the unconditional variance of the regression predictor is

$$V\{\bar{y}_{reg} - \bar{y}_N\} = (1 - f_N)n^{-1}\left[1 + k(n - k - 2)^{-1}\right]\sigma_e^2. \quad (2.2.16)$$

If $\mathbf{x}_{1,i}$ is not normally distributed, expression (2.2.15) furnishes an approximation that is correct through terms of order n^{-2}. It follows that if the multiple correlation $R^2 = (1 - \sigma_y^{-2}\sigma_e^2)$ is greater than $k(n-2)^{-1}$, the regression estimator is superior to the simple mean as an estimator of the population mean. Because $k(n-2)^{-1}$ is small for fixed k and large n, it is often stated that in large samples, the regression estimator is never inferior to the sample mean. In practice, there is a temptation to increase the dimension of the x-vector until the term $k(n-k-2)^{-1}$ becomes important.

The estimator $\bar{y}_{reg}$ is linear in y and can be written

$$\bar{y}_{reg} = \sum_{i\in A} w_i y_i, \quad (2.2.17)$$

where the weights are

$$w_i = \bar{\mathbf{x}}_N\left(\sum_{j\in A}\mathbf{x}_j'\mathbf{x}_j\right)^{-1}\mathbf{x}_i'$$
$$= n^{-1} + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})\left(\sum_{j\in A}(\mathbf{x}_{1,j} - \bar{\mathbf{x}}_{1,n})'(\mathbf{x}_{1,j} - \bar{\mathbf{x}}_{1,n})\right)^{-1}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})'.$$

Given the regression estimator for the mean, the regression estimator for the total is the estimator for the mean multiplied by N. Thus, the estimated total

$$\hat{T}_{y,reg} = \sum_{i\in A} N w_i y_i, \quad (2.2.18)$$

where the w_i are defined in (2.2.17), is also linear in y. The linearity in y is very important in practice because it means that the w_i can be computed once and used to compute estimates for any characteristic.

An additional advantage of the regression estimator is that, by construction, the weight applied to the vector $\mathbf{x}_i$ gives the population mean, $\bar{\mathbf{x}}_N$. See equation (2.2.20). Thus, if one has a survey of the residents of a state and uses the official census number of persons by age and gender as control totals in the vector $N\bar{\mathbf{x}}_N$, any table with gender or age as columns or rows will give the official numbers for the margins. This is a considerable asset to researchers who are comparing their analyses to those of others.

A derivation of the regression estimator under alternative assumptions will serve to illustrate its properties. Assume that we desire the estimator of $\bar{\mathbf{x}}_N\boldsymbol{\beta}$ with the smallest variance in the class of linear estimators that are unbiased under the model

$$\begin{aligned} y_j &= \mathbf{x}_j\boldsymbol{\beta} + e_j, \\ e_j &\sim ii(0,\ \sigma_e^2), \end{aligned} \tag{2.2.19}$$

where $\mathbf{x}_j = (1,\ \mathbf{x}_{1,j})$, and e_j is independent of $\mathbf{x}_i$ for all i and j. To calculate the variance, we only require the e_j to be uncorrelated with common variance, but we retain the independence assumption. Now

$$\begin{aligned} E\left\{\sum_{i\in A} w_i y_i \mid \mathbf{X}\right\} &= E\left\{\sum_{i\in A} w_i\mathbf{x}_i\boldsymbol{\beta} + \sum_{i\in A} w_i e_i \mid \mathbf{X}\right\} \\ &= \sum_{i\in A} w_i\mathbf{x}_i\boldsymbol{\beta}. \end{aligned}$$

If $\Sigma_{i\in A} w_i y_i$ is to be unbiased for $\bar{\mathbf{x}}_N\boldsymbol{\beta}$, we must have $\Sigma_{i\in A} w_i\mathbf{x}_i = \bar{\mathbf{x}}_N$. Thus, we desire the w_i, $i = 1,\ 2,\ \ldots,\ n$, that minimize

$$V\left\{\sum_{i\in A} w_i y_i \mid \mathbf{X}\right\} = \sum_{i\in A} w_i^2\sigma_e^2$$

subject to the condition that

$$\sum_{i\in A} w_i\mathbf{x}_i = \bar{\mathbf{x}}_N. \tag{2.2.20}$$

The problem can be formulated as the minimization of the Lagrangian

$$\sum_{i\in A} w_i^2\sigma_e^2 + \boldsymbol{\lambda}\left(\sum_{i\in A} w_i\mathbf{x}_i - \bar{\mathbf{x}}_N\right)', \tag{2.2.21}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{k+1})$ is a row vector of Lagrange multipliers. The solution is

$$w_i = -0.5\boldsymbol{\lambda}\mathbf{x}_i',$$

where

$$\boldsymbol{\lambda} = 2\bar{\mathbf{x}}_N(\mathbf{X}'\mathbf{X})^{-1}$$

and $\mathbf{X}$ was defined following (2.2.1). Of course, these w_i are precisely those defined in (2.2.17).

The regression estimator defined with $\mathbf{x}_j = (1, \mathbf{x}_{1,j})$ and the weights of (2.2.17) has the important property that it is scale and location invariant. Thus, for arbitrary α_0 and α_1, the regression estimator of the mean of $u = \alpha_0 + \alpha_1 y$ is $\bar{u}_{reg} = \alpha_0 + \alpha_1 \bar{y}_{reg}$. Some have taken linearity, scale invariance, and location invariance as defining a regression estimator. See Mickey (1959).

A reparameterization of the regression equation provides another useful representation of the regression estimator. A linear transformation can be constructed in which the regression estimator is the coefficient for the first element of $\mathbf{x}$, the element that is identically equal to 1. Express the other x-variables as deviations from the population mean and write

$$\mathbf{z}_i = (1, x_{1,i} - \bar{x}_{1,N}, x_{2,i} - \bar{x}_{2,N}, \ldots, x_{k,i} - \bar{x}_{k,N}).$$

The population mean of $\mathbf{z}$ is $\bar{\mathbf{z}}_N = (1, \mathbf{0})$ and the transformed regression model is

$$y_i = \mathbf{z}_i\boldsymbol{\gamma} + e_i,$$

where $\boldsymbol{\gamma}' = (\gamma_0, \gamma_1, \ldots, \gamma_k) = (\gamma_0, \boldsymbol{\beta}_1')$. The vector regression coefficient for the regression of y on $\mathbf{z}$ is

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \left(\sum_{i\in A}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\sum_{i\in A}\mathbf{z}_i'y_i \\ &= \left(\bar{y}_n - \sum_{j=1}^{k}(\bar{x}_{j,n} - \bar{x}_{j,N})\hat{\beta}_j,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \ldots,\ \hat{\beta}_k\right)', \qquad (2.2.22)\end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is defined in (2.2.2). In expression (2.2.22) we see that the regression estimator of the mean of y is the regression coefficient $\hat{\gamma}_0$. That is,

$$\bar{y}_{reg} = \bar{\mathbf{z}}_N\hat{\boldsymbol{\gamma}} = \hat{\gamma}_0$$

and

$$V\{\hat{\gamma}_0 \mid \mathbf{X}\} = n^{-1}\sigma_e^2 + \bar{\mathbf{z}}_{1,n}V\{\hat{\boldsymbol{\beta}}_1 \mid \mathbf{X}\}\bar{\mathbf{z}}_{1,n}', \qquad (2.2.23)$$

where $\bar{\mathbf{z}}_{1,n} = (\bar{\mathbf{x}}_{1,n} - \bar{\mathbf{x}}_{1,N})$. Variance expression (2.2.23) is expression (2.2.13) without the finite population correction.

Under the model of this section, the expected value of e is zero for every x. If this is not true, the regression estimator has a bias of $O(n^{-0.5})$. To see the nature of the bias, consider the estimator for a simple random sample with a single x-variable. Then

$$\begin{aligned}
\bar{y}_{reg} - \bar{y}_N &= \bar{y}_n - \bar{y}_N + (\bar{x}_N - \bar{x}_n)\beta + (\bar{x}_N - \bar{x}_n)(\hat{\beta} - \beta) \\
&= \bar{e}_n + (\bar{x}_N - \bar{x}_n)\left(\sum_{i\in A}(x_i - \bar{x}_n)^2\right)^{-1}\sum_{i\in A}(x_i - \bar{x}_n)(e_i - \bar{e}_n) \\
&= \bar{e}_n + (\bar{x}_N - \bar{x}_n)S_x^{-2}(n-1)^{-1}\sum_{i\in A}(x_i - \bar{x}_n)(e_i - \bar{e}_n) \\
&\quad + O_p(n^{-3/2})
\end{aligned}$$

and the mean of the approximate distribution of $\bar{y}_{reg} - \bar{y}_N$ is

$$-N^{-1}(N-n)n^{-1}E\{S_x^{-2}(x_i - \bar{x}_N)^2 e_i\}.$$

Thus, the approximate bias is a function of the covariance between e_i and $(x_i - \bar{x}_N)^2$.

2.2.2 General populations and complex samples

Let us now relax our assumptions on the population. Define the $(k+1)$-dimensional vector $\mathbf{q}_j = (y_j,\ \mathbf{x}_{1,j})'$ and assume that

$$\mathbf{q}_j \sim ii\left[(\mu_y,\ \boldsymbol{\mu}_x)',\ \boldsymbol{\Sigma}_{qq}\right].$$

Assume that $\mathbf{q}_i$ has finite fourth moments and that the rank of $\boldsymbol{\Sigma}_{qq}$ is $k+1$. If we define $\boldsymbol{\beta}$ to be the value of $\boldsymbol{\gamma}$ that minimizes

$$E\{(y_i - \mathbf{x}_i\boldsymbol{\gamma})^2\}, \tag{2.2.24}$$

where $\mathbf{x}_i = (1,\ \mathbf{x}_{1,i})$, we obtain

$$\boldsymbol{\beta} = \left[E\{\mathbf{x}_i'\mathbf{x}_i\}\right]^{-1} E\left\{\mathbf{x}_i' y_i\right\}. \tag{2.2.25}$$

Consider a sequence of populations and estimators as described in Section 1.3, where the Nth population is a realization of vectors $\mathbf{q}_i$, $i = 1,\ 2,\ \ldots,\ N$. Let $\bar{\mathbf{x}}_N$ be known for the Nth population. Let a sequence of simple random nonreplacement samples be selected, where the sample selected from the Nth population is of size n_N, $n_N \geq n_{N-1}$,

$$\lim_{N\to\infty} n_N = \infty,$$

and

$$\lim_{N\to\infty} f_N = f < 1,$$

where $f_N = N^{-1}n_N$. Because β is a differentiable function of second moments, the least squares estimator $\hat{\beta}$ of (2.2.2) satisfies

$$\hat{\beta} - \beta = O_p(n_N^{-1/2}).$$

Also, under these assumptions,

$$\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n} = O_p(n_N^{-1/2}).$$

Therefore, from (2.2.12), the regression estimator (2.2.10) constructed with $\hat{\beta}$ of (2.2.2) satisfies

$$\bar{y}_{reg} - \bar{y}_N = (1 - f_N)(\bar{e}_n - \bar{e}_{N-n}) + O_p(n_N^{-1}), \tag{2.2.26}$$

where $e_i = y_i - \mathbf{x}_i\beta$ and β is defined in (2.2.25). It follows that the variance of the limiting distribution of $n_N^{1/2}(\bar{y}_{reg} - \bar{y}_N)$ is

$$\lim_{N\to\infty} V\{n_N^{1/2}(1 - f_N)(\bar{e}_n - \bar{e}_{N-n})\} = (1 - f_N)\sigma_e^2, \tag{2.2.27}$$

where $\sigma_e^2 = V\{y_i - \mathbf{x}_i\beta\}$. Because β minimizes (2.2.24), it minimizes σ_e^2 and there is no estimator of $\bar{y}_N$, linear in y_i, whose limit distribution has a smaller variance. This result is extended to general designs and estimators in Theorem 2.2.3.

Under simple random sampling, the estimator $\hat{\beta}$ is also design consistent for the finite population regression coefficient in that

$$\hat{\beta} - \beta_N \mid \mathcal{F}_N = O_p(n_N^{-1/2}) \text{ a.s.}, \tag{2.2.28}$$

where

$$\beta_N = \left(\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i'y$$

and $\beta_N - \beta = O_p(N^{-1/2})$.

Given the definition of the finite population regression coefficient, the error in the regression estimator of the finite population mean can be written in terms of deviations from the finite population regression. The error in $\hat{\beta}$ as an estimator of the finite population parameter is

$$\hat{\beta} - \beta_N = \left(\sum_{i\in A_N} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \sum_{i\in A_N} \mathbf{x}_i'a_i \tag{2.2.29}$$

and

$$\begin{aligned}\bar{y}_{reg} - \bar{y}_N &= \bar{a}_n + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_n)\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \\ &= \bar{a}_n + O_p(n_N^{-1}) \quad \text{a.s.},\end{aligned}$$

where $a_i = a_{Ni} = y_i - \mathbf{x}_i\boldsymbol{\beta}_N$. It follows that the variance of the approximating distribution, conditional on $\mathcal{F}_N$, is

$$V\{n_N^{1/2}(\bar{y}_{reg} - \bar{y}_N) \mid \mathcal{F}_N\} = (1 - f_N)S_{aN}^2, \tag{2.2.30}$$

where

$$S_{aN}^2 = (N-1)^{-1} \sum_{i \in U_N} a_{Ni}^2.$$

An estimator of the approximate variance of the regression estimator of the mean of y under simple random sampling is

$$\hat{V}\{\bar{y}_{reg} - \bar{y}_N\} = (1 - f_N)n_N^{-1}(n_N - k - 1)^{-1} \sum_{i \in A_N} \hat{a}_i^2, \tag{2.2.31}$$

where the divisor is chosen by analogy to (2.2.4) and

$$\hat{a}_i = \hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}.$$

Two things are noteworthy. First, the estimator (2.2.31) is an estimator of (2.2.27), the variance expression in terms of the superpopulation variance, and of (2.2.30), the variance expression in terms of the finite population variance. Second, the simple estimator underestimates the true variance because it contains no term for the estimation error in $\hat{\boldsymbol{\beta}}$. However, the variance contribution from estimating $\boldsymbol{\beta}$ is order n_N^{-2}. See, for example, (2.2.13) and (2.2.16).

By (2.2.22), we can express the regression estimator of the mean as a regression coefficient, and the limiting properties of the regression estimator follow from the limiting properties of the vector of regression coefficients. We give a general result for the coefficients in Theorem 2.2.1. We extend the definition of the regression coefficient to the generalized regression coefficient and give the result for designs in which $V\{\bar{y}_{HT}|\mathcal{F}_N\} = O(n_{BN}^{-1})$ almost surely and in which the limiting distribution of $n^{1/2}(\bar{y}_{HT} - \bar{y}_N)$ is normal almost surely.

Theorem 2.2.1. Let $\mathcal{F}_N = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$ for $N = k+3, k+4, \ldots$, where $\mathbf{z}_j = (y_j, \mathbf{x}_j), j = 1, 2, \ldots$, and $\{\mathbf{z}_j\}$ is a sequence of $(k+2)$-dimensional, independent random vectors with bounded eighth moments. Let $\mathbf{Z}$ be the $n \times (k+2)$ matrix of observations for the sample from the Nth population. Let

$$\hat{\mathbf{M}}_{z\phi z} = n^{-1}\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z} \tag{2.2.32}$$

for a positive definite $n \times n$ matrix $\mathbf{\Phi}$ that may be a function of $\mathbf{x}$ but not of y. If $\mathbf{\Phi}$ is random, assume that the rows of $\mathbf{\Phi}^{-1}\mathbf{Z}$ have bounded fourth moments. Assume:

(i) The sample design is such that for any $\mathbf{z}$ with bounded fourth moments

$$V\{\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N \mid \mathcal{F}_N\} = O_p(n_{BN}^{-1}) \text{ a.s.,} \tag{2.2.33}$$

where

$$\bar{\mathbf{z}}_{HT} = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{z}_i,$$

π_i are the selection probabilities, and $\bar{\mathbf{z}}_N$ is the finite population mean of $\mathbf{z}$.

(ii) There is a sequence $\{\mathbf{M}_{z\phi z,N}\}$ such that

$$\hat{\mathbf{M}}_{z\phi z} - \mathbf{M}_{z\phi z,N} \mid \mathcal{F}_N = O_p(n_{BN}^{-1/2}) \text{ a.s.,} \tag{2.2.34}$$

$\mathbf{M}_{z\phi z,N}$ is positive definite almost surely, the limit of $\mathbf{M}_{z\phi z,N}$ is positive definite, and $\hat{\mathbf{M}}_{z\phi z}$ is positive definite almost surely.

(iii) The selection probabilities satisfy

$$K_1 < N n_{BN}^{-1} \pi_i < K_2$$

for positive K_1 and K_2.

(iv) The design is such that

$$[V\{\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N \mid \mathcal{F}_N\}]^{-1/2} (\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \text{ a.s.,} \tag{2.2.35}$$

as $n \to \infty$ for any $\mathbf{z}$ with finite fourth moments, where $V\{\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N \mid \mathcal{F}_N\} =: \mathbf{V}_{zz,N}$ is the positive definite covariance matrix of $\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N$, and $\mathbf{V}_{zz,N}^{1/2}$ is the symmetric square root of $\mathbf{V}_{zz,N}$.

(v) The design admits an estimator $\hat{\mathbf{V}}_{zz}$ such that

$$n(\hat{\mathbf{V}}_{zz} - \mathbf{V}_{zz,N}) \mid \mathcal{F}_N = o_p(1) \text{ a.s.} \tag{2.2.36}$$

for any z with bounded fourth moments.

Let

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{y}. \tag{2.2.37}$$

Then

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N \mid \mathcal{F}_N \;=\; \mathbf{M}_{x\phi x,N}^{-1}\bar{\mathbf{b}}_{HT}' \;+\; O_p(n_{BN}^{-1}) \quad \text{a.s.}, \tag{2.2.38}$$

where

$$\begin{aligned}
\boldsymbol{\beta}_N &= \mathbf{M}_{x\phi x,N}^{-1}\mathbf{M}_{x\phi y,N}, \\
\bar{\mathbf{b}}_{HT}' &= N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{b}_i', \\
\mathbf{M}_{z\phi z,N} &= \begin{pmatrix} M_{y\phi y,N} & \mathbf{M}_{y\phi x,N} \\ \mathbf{M}_{x\phi y,N} & \mathbf{M}_{x\phi x,N} \end{pmatrix},
\end{aligned}$$

$\mathbf{b}_i' = n_{BN}^{-1}N\pi_i\boldsymbol{\zeta}_i' a_{Ni}$, $a_{Ni} = y_i - \mathbf{x}_i\boldsymbol{\beta}_N$, and $\boldsymbol{\zeta}_i'$ is column i of $\mathbf{X}'\boldsymbol{\Phi}^{-1}$.

Furthermore,

$$[\hat{V}\{\hat{\boldsymbol{\beta}} \mid \mathcal{F}_N\}]^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0},\, \mathbf{I}) \quad \text{a.s.}, \tag{2.2.39}$$

where

$$\hat{V}\{\hat{\boldsymbol{\beta}} \mid \mathcal{F}_N\} \;=\; \hat{\mathbf{M}}_{x\phi x}^{-1}\hat{\mathbf{V}}_{\bar{b}\bar{b}}\hat{\mathbf{M}}_{x\phi x}^{-1},$$

$\hat{\mathbf{V}}_{\bar{b}\bar{b}} = \hat{V}\{\bar{\mathbf{b}}_{HT} \mid \mathcal{F}_N\}$ is the estimated sampling variance of $\bar{\mathbf{b}}_{HT}$ calculated with $\hat{\mathbf{b}}_i' = n_{BN}^{-1}N\pi_i\hat{\boldsymbol{\zeta}}_i'\hat{a}_i$, and $\hat{a}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$.

Proof. The error in $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N &= (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y} - \mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\boldsymbol{\beta}_N) \\
&= \hat{\mathbf{M}}_{x\phi x}^{-1}(n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{a}).
\end{aligned}$$

By the definition of $\boldsymbol{\beta}_N$,

$$\mathbf{M}_{x\phi y,N} - \mathbf{M}_{x\phi x,N}\boldsymbol{\beta}_N \;=\; \mathbf{M}_{x\phi a,N} \;=\; \mathbf{0} \quad \text{a.s.},$$

and by assumption (2.2.34),

$$\hat{\mathbf{M}}_{x\phi a} \;=\; n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{a} \;=\; O_p(n^{-1/2}).$$

It follows that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N \mid \mathcal{F}_N \;=\; \mathbf{M}_{x\phi x,N}^{-1}\left(N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{b}_i'\right) + O_p(n^{-1}) \quad \text{a.s.}$$

The $\mathbf{b}_i$ have bounded fourth moments by the moment assumptions and by the bounds on $Nn_{BN}^{-1}\pi_i$. Thus, by assumption (2.2.35),

$$\mathbf{V}_{\beta\beta,N}^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \quad \text{a.s.}, \tag{2.2.40}$$

where

$$\mathbf{V}_{\beta\beta,N} = \mathbf{M}_{x\phi x,N}^{-1}\mathbf{V}_{\bar{b}\bar{b},N}\mathbf{M}_{x\phi x,N}^{-1}$$

and $\mathbf{V}_{\bar{b}\bar{b},N} = V\{\bar{\mathbf{b}}_{HT} \mid \mathcal{F}_N\}$.

For variance estimation, consider

$$\begin{aligned} n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\hat{\mathbf{a}} &= n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{a} - n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \\ &=: N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{b}_i' - N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i', \end{aligned}$$

where

$$\mathbf{h}_i' = n^{-1}N\pi_i\boldsymbol{\zeta}_i'\mathbf{x}_i\boldsymbol{\delta}_\beta$$

and $\boldsymbol{\delta}_\beta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N$. For any fixed $\boldsymbol{\delta}$, by (2.2.36), the estimated variance of $N^{-1}\Sigma_{i\in A}\pi_i^{-1}(\mathbf{b}_i' + \mathbf{h}_i')$ is consistent for the variance of the estimator of the mean of $\mathbf{b} + \mathbf{h}$. By assumption, the elements of $\boldsymbol{\zeta}_i'\mathbf{x}_i$ have fourth moments, and for a fixed $\boldsymbol{\delta}$, the variance of $\bar{\mathbf{h}}_{HT}$ is $O_p(n^{-1})$ almost surely. Therefore, for $\boldsymbol{\delta} = \boldsymbol{\delta}_\beta$,

$$\hat{V}\{\bar{\mathbf{h}}_{HT}'|\mathcal{F}_N\} \mid \mathcal{F}_N = o_p(n^{-1}) \quad \text{a.s.}$$

and

$$\hat{V}\{\bar{\mathbf{b}}_{HT}'|\mathcal{F}_N\} \mid \mathcal{F}_N = V\{\bar{\mathbf{b}}_{HT}'|\mathcal{F}_N\} + o_p(n^{-1}) \quad \text{a.s.}$$

Result (2.2.39) then follows from (2.2.40). ■

In Theorem 2.2.1 it is assumed that $\hat{\mathbf{M}}_{z\phi z}$ converges to some positive definite matrix $\mathbf{M}_{z\phi z,N}$. If $\boldsymbol{\Phi}$ is the $n \times n$ portion of an $N \times N$ diagonal matrix $\boldsymbol{\Phi}_N$, then

$$\mathbf{M}_{z\phi z,N} = n^{-1}\mathbf{Z}_N'\boldsymbol{\Phi}_N^{-1}\mathbf{D}_{\pi,N}\mathbf{Z}_N$$

and

$$\boldsymbol{\beta}_N = (\mathbf{X}_N'\boldsymbol{\Phi}_N^{-1}\mathbf{D}_{\pi,N}\mathbf{X}_N)^{-1}\mathbf{X}_N'\boldsymbol{\Phi}_N^{-1}\mathbf{D}_{\pi,N}\mathbf{y}_N, \tag{2.2.41}$$

where $\mathbf{X}_N$ is the $N \times k$ matrix of the N explanatory vectors, $\mathbf{y}_N$ is the vector of N values of y_i, and $\mathbf{D}_{\pi,N} = \text{diag}(\pi_1, \pi_2, \ldots, \pi_N)$. If $\boldsymbol{\Phi}_N = \mathbf{D}_{\pi,N}$, then $\boldsymbol{\beta}_N$ is the ordinary least squares coefficient for the finite population.

The conditions of Theorem 2.2.1 require the variables to have moments and require the orders of the error in the estimators and in the estimated variances to be the same as those of a simple random sample. This will be true for stratified samples and for stratified two-stage samples under mild restrictions on the sequence of populations.

In Theorem 2.2.1 we defined $\mathbf{M}_{z\phi z,N}$ as the matrix being estimated by $n^{-1}\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}$. In the case of diagonal $\boldsymbol{\Phi}$, where ϕ_i is defined for all N elements,

$$\mathbf{M}_{z\phi z,N} = n^{-1} \sum_{i \in U} \mathbf{z}_i' \phi_i^{-1} \pi_i \mathbf{z}_i.$$

We give the limiting distribution for the regression coefficients for simple random sampling in Theorem 2.2.2, where the results are for samples from all possible sequences.

Theorem 2.2.2. Let $\{U_N, \mathcal{F}_N : N = k+3, k+4, \ldots\}$ be a sequence of finite populations, where U_N is the set of indices identifying the elements. Let $\mathcal{F}_N = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$, where $\mathbf{z}_i = (y_i, \mathbf{x}_i)$, $\{\mathbf{z}_i\}$ is a sequence of iid random variables with $E\{\mathbf{z}_i, \mathbf{z}_i'\mathbf{z}_i\} = (\boldsymbol{\mu}_z, \mathbf{M}_{zz})$, and $\mathbf{M}_{zz}$ is of full rank. Assume that $\mathbf{q}$ has finite fourth moments, where

$$\mathbf{q}_i = \left[y_i, \mathbf{x}_i, (vech\mathbf{x}_i'\mathbf{x}_i)'\right],$$

$vech\mathbf{A}$ of the $p \times p$ symmetric matrix $\mathbf{A}$ is

$$vech\mathbf{A} = (a_{11}, a_{21}, \ldots, a_{p1}, a_{22}, a_{32}, \ldots, a_{p2}, \ldots, a_{pp})',$$

and a_{ij} is the ijth element of $\mathbf{A}$.

Let A_N be a set of $n_N \geq n_{N-1}$ indices selected as a simple random nonreplacement sample from U_N. Assume that $n_N \to \infty$ as $N \to \infty$ and

$$\lim_{N \to \infty} f_N = f, \qquad 0 \leq f < 1.$$

Let the quantities of Theorem 2.2.1 be defined with $\boldsymbol{\Phi} = \mathbf{I}$ so that

$$\hat{\boldsymbol{\beta}} = \left(n_N^{-1} \sum_{i \in A_N} \mathbf{x}_i'\mathbf{x}_i \right)^{-1} n_N^{-1} \sum_{i \in A_N} \mathbf{x}_i' y_i =: \hat{\mathbf{M}}_{xx}^{-1} \hat{\mathbf{M}}_{xy}, \tag{2.2.42}$$

$$\boldsymbol{\beta}_N = \left(N^{-1} \sum_{i \in U_N} \mathbf{x}_i'\mathbf{x}_i \right)^{-1} N^{-1} \sum_{i \in U_N} \mathbf{x}_i' y_i =: \mathbf{M}_{xx,N}^{-1} \mathbf{M}_{xy,N}, \tag{2.2.43}$$

and

$$\beta = [E\{\mathbf{x}_i'\mathbf{x}_i\}]^{-1}E\{\mathbf{x}_i'y_i\} =: \mathbf{M}_{xx}^{-1}\mathbf{M}_{xy}.$$

Then

$$[V\{\hat{\beta} - \beta_N \mid \mathcal{F}_N\}]^{1/2}(\hat{\beta} - \beta_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \quad \text{a.s.}, \tag{2.2.44}$$

where

$$V\{\hat{\beta} - \beta_N \mid \mathcal{F}_N\} = n_N^{-1}(1 - f_N)\mathbf{M}_{xx,N}^{-1}\mathbf{V}_{bb,N}\mathbf{M}_{xx,N}^{-1},$$

$$\mathbf{V}_{bb,N} = N^{-1}\sum_{i \in U_N} \mathbf{x}_i' e_{iN}^2 \mathbf{x}_i,$$

and $e_{iN} = y_i - \mathbf{x}_i\beta_N$.

Also,

$$n_N^{1/2}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{M}_{xx}^{-1}\mathbf{V}_{bb}\mathbf{M}_{xx}^{-1}), \tag{2.2.45}$$

where $\mathbf{V}_{bb} = E\{\mathbf{x}_i'e_i^2\mathbf{x}_i\}$ and $e_i = y_i - \mathbf{x}_i\beta$.

Proof. We can write

$$\hat{\beta} - \beta = \hat{\mathbf{M}}_{xx}^{-1}\hat{\mathbf{M}}_{xe}$$

and

$$\beta_N - \beta = \mathbf{M}_{xx,N}^{-1}\mathbf{M}_{xe,N},$$

where $\hat{\mathbf{M}}_{xe}$ and $\mathbf{M}_{xe,N}$ are defined by analogy to (2.2.42) and (2.2.43).

By the moment assumptions,

$$\lim_{N \to \infty} (\hat{\mathbf{M}}_{xx}, \hat{\mathbf{M}}_{xe}) = (\mathbf{M}_{xx}, \mathbf{M}_{xe}) \quad \text{a.s.}$$

and

$$(\hat{\mathbf{M}}_{xx}, \hat{\mathbf{M}}_{xe}) - (\mathbf{M}_{xx,N}, \mathbf{M}_{xe,N}) \mid \mathcal{F}_N = O_p(n_N^{-1/2}) \quad \text{a.s.} \tag{2.2.46}$$

Now $\mathbf{x}_i'e_i$ is a characteristic of the ith observation in a simple random sample from a finite population and the finite population variance is converging to a finite positive number almost surely. Therefore, by Corollary 1.3.3.1,

$$[(1 - f_N)n_N^{-1}\mathbf{V}_{bb,N}]^{-1/2}\sum_{i \in A} \mathbf{x}_i'e_i \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \quad \text{a.s.}$$

Result (2.2.44) follows by (2.2.46).

Arguments supporting result (2.2.45) are given in the first part of this section. See (2.2.27). Alternatively, one can use result (2.2.46), the limiting normality of $\bar{e}_N$, and Theorem 1.3.6 to prove (2.2.45). ■

An estimator of the variance of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N$ of Theorem 2.2.2 can be constructed by replacing e_i with $\hat{e}_i$ in $\hat{\mathbf{M}}_{xe}$, where $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$. Thus, an estimator of the variance of the regression vector for a simple random sample is

$$\hat{V}\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N \mid \mathcal{F}\} = \hat{\mathbf{M}}_{xx}^{-1}\hat{\mathbf{V}}_{\bar{b}\bar{b}}\hat{\mathbf{M}}_{xx}^{-1}, \tag{2.2.47}$$

where $\hat{\mathbf{M}}_{xx}$ is defined in (2.2.42) and $\hat{\mathbf{V}}_{\bar{b}\bar{b}}$ is a design-consistent estimator of the variance of the mean vector

$$\tilde{\mathbf{b}}_\pi = n^{-1}\sum_{i\in A}\mathbf{x}_i'e_i.$$

The estimator (2.2.47) is generally an underestimate because it contains no degrees-of-freedom adjustment for the use of estimators in constructing $\hat{e}_i$. Also, $\hat{e}_i^2$ is often negatively correlated with x_i^2, so that $x_i^2\hat{e}_i^2$ often underestimates $x_i^2e_i^2$.

If the regression vector can be written $\mathbf{x}_i = (1, \mathbf{x}_{1,i})$ and if $\mathbf{x}_i$ is coded in the deviation form used in (2.2.22), the estimated variance of the regression estimator of the mean (2.2.17) is the first element of (2.2.47). Unlike expression (2.2.31), the estimated variance so constructed contains a contribution due to estimating the last k elements of $\boldsymbol{\beta}$. The estimated variance (2.2.47) for the first element of the reparameterized model, denoted by $\hat{\gamma}_0$, can be written in the form

$$\hat{V}\{\bar{y}_{reg}\} = \hat{V}\{\hat{\gamma}_0\} = \sum_{i\in A}(w_i\hat{e}_i)^2, \tag{2.2.48}$$

where

$$w_i = n^{-1} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_n)\hat{\mathbf{M}}_{xx}^{-1}\mathbf{x}_i'.$$

Many of the samples encountered in practice are more complicated than the simple random sample of Theorem 2.2.2. It does not follow from Theorem 2.2.1 that an estimator of the population mean using the estimator $\hat{\boldsymbol{\beta}}$ of (2.2.37) will be a design-consistent estimator of the finite population mean. In Theorem 2.2.3 we give conditions such that the regression estimator of the population mean is design consistent.

Theorem 2.2.3. Let $\mathbf{z}_1, \mathbf{z}_2, \ldots,$ be a sequence of real vectors and let $\mathcal{F}_N = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)$ be a sequence of finite populations. Let $\bar{\mathbf{z}}_N = (\bar{y}_N, \bar{\mathbf{x}}_N)$ be the mean of $\mathbf{z}_i = (y_i, \mathbf{x}_i)$ for the Nth population. Let a sequence of samples be selected from the sequence $\{\mathcal{F}_N\}$. Define the regression estimator of $\bar{y}_N$

by

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N \tilde{\boldsymbol{\beta}},$$

where $\tilde{\boldsymbol{\beta}}$ is a design-consistent estimator of a parameter denoted by $\boldsymbol{\beta}_N$. Then

$$p \lim_{N\to\infty} \{(\bar{y}_{reg} - \bar{y}_N) \mid \mathcal{F}_N\} = 0$$

if and only if

$$p \lim_{N\to\infty} \{\bar{a}_N \mid \mathcal{F}_N\} = 0, \tag{2.2.49}$$

where $a_i = y_i - \mathbf{x}_i \boldsymbol{\beta}_N$.

Proof. Because $\tilde{\boldsymbol{\beta}}$ is design consistent for $\boldsymbol{\beta}_N$,

$$\begin{aligned} p \lim_{N\to\infty} \{(\bar{y}_N - \bar{\mathbf{x}}_N \tilde{\boldsymbol{\beta}}) \mid \mathcal{F}_N\} &= p \lim_{N\to\infty} \{(\bar{y}_N - \bar{\mathbf{x}}_N \boldsymbol{\beta}_N) \mid \mathcal{F}_N\} \\ &= p \lim_{N\to\infty} \{\bar{a}_N \mid \mathcal{F}_N\}, \end{aligned}$$

and we have the conclusion. ■

Our most used condition for design consistency is given in Corollary 2.2.3.1.

Corollary 2.2.3.1. Let $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ and $\mathbf{X}' = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')$. Let $\{\boldsymbol{\Phi}_n\}$ be a sequence of nonsingular symmetric $n \times n$ matrices. Let $\bar{y}_{HT}$, $\bar{\mathbf{x}}_{HT}$, $n_N^{-1}(\mathbf{X}'\boldsymbol{\Phi}_n^{-1}\mathbf{X})$, and $n_N^{-1}\mathbf{X}'\boldsymbol{\Phi}_n^{-1}\mathbf{y}$ be design-consistent estimators for finite population characteristics $\bar{y}_N$, $\bar{\mathbf{x}}_N$, $\mathbf{M}_{x\phi x,N}$, and $\mathbf{M}_{x\phi y,N}$, respectively. Assume that there is a sequence of $(k+1)$-dimensional column vectors $\{\boldsymbol{\gamma}_n\}$ such that

$$\mathbf{X}\boldsymbol{\gamma}_n = \boldsymbol{\Phi}_n \mathbf{D}_\pi^{-1} \mathbf{J}_n \tag{2.2.50}$$

for all possible samples, where $\mathbf{D}_\pi = \operatorname{diag}(\pi_1, \pi_2, \ldots, \pi_n)$ and $\mathbf{J}_n$ is a column vector of 1's. Then the regression estimator

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}, \tag{2.2.51}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Phi}_n^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\boldsymbol{\Phi}_n^{-1}\mathbf{y}, \tag{2.2.52}$$

is a design-consistent estimator of $\bar{y}_N$.

Proof. The estimator $\hat{\boldsymbol{\beta}}$ is design consistent for

$$\boldsymbol{\beta}_N = \mathbf{M}_{x\phi x,N}^{-1}\mathbf{M}_{x\phi y,N},$$

by the design consistency of the sample moments. If $\hat{\boldsymbol{\beta}}$ is defined by (2.2.52), then by the properties of generalized least squares estimators,

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Phi}_n^{-1}\mathbf{X} = \mathbf{0}.$$

If (2.2.50) holds,

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{D}_\pi^{-1}\mathbf{J}_n = N(\bar{y}_{HT} - \bar{\mathbf{x}}_{HT}\hat{\boldsymbol{\beta}}) = 0.$$

It follows that $\bar{y}_{reg}$ is design consistent because

$$\begin{aligned} 0 &= p\lim_{N\to\infty}\{(\bar{y}_{HT} - \bar{\mathbf{x}}_{HT}\hat{\boldsymbol{\beta}}) \mid \mathcal{F}_N\} \\ &= p\lim_{N\to\infty}\{(\bar{y}_{HT} - \bar{\mathbf{x}}_{HT}\boldsymbol{\beta}_N) \mid \mathcal{F}_N\} \\ &= p\lim_{N\to\infty}\{(\bar{y}_N - \bar{x}_N\boldsymbol{\beta}_N) \mid \mathcal{F}_N\} \end{aligned}$$

and condition (2.2.49) of Theorem 2.2.3 is satisfied. ■

To describe an estimator meeting the requirements of Corollary 2.2.3.1, we reparameterize the regression problem by defining $\mathbf{z}_0 = \boldsymbol{\Phi}_n\mathbf{w}$ and

$$\mathbf{Z}_1 = \mathbf{X}_1 - \mathbf{z}_0(\mathbf{z}_0'\boldsymbol{\Phi}_n^{-1}\mathbf{z}_0)^{-1}\mathbf{z}_0'\boldsymbol{\Phi}_n^{-1}\mathbf{X}_1,$$

where $\mathbf{w} = \mathbf{D}_\pi^{-1}\mathbf{J}_n$, $\mathbf{Z} = (\mathbf{z}_0, \mathbf{Z}_1)$, and $\mathbf{X} = (\mathbf{z}_0, \mathbf{X}_1)$. Then

$$\begin{aligned} \mathbf{Z}'\boldsymbol{\Phi}_n^{-1}\mathbf{Z} &= \begin{pmatrix} \mathbf{z}_0'\boldsymbol{\Phi}_n^{-1}\mathbf{z}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_1'\boldsymbol{\Phi}_n^{-1}\mathbf{Z}_1 \end{pmatrix} \\ &=: \text{blockdiag}(\mathbf{z}_0'\boldsymbol{\Phi}_n^{-1}\mathbf{z}_0,\ \mathbf{Z}_1'\boldsymbol{\Phi}_n^{-1}\mathbf{Z}_1), \end{aligned}$$

where $\text{blockdiag}(\mathbf{B}, \mathbf{G})$ is a matrix composed of four submatrices with $\mathbf{B}$ and $\mathbf{G}$ as the main diagonal matrices and zero matrices as the off-diagonal matrices. It follows that the regression estimator (2.2.51) can be written

$$\begin{aligned} \bar{y}_{reg} &= \bar{\mathbf{z}}_N\hat{\boldsymbol{\gamma}} \\ &= \bar{y}_{dc} + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,dc})\hat{\boldsymbol{\beta}}_1, \end{aligned} \tag{2.2.53}$$

where $\hat{\boldsymbol{\beta}}$ is defined in (2.2.52), $\hat{\boldsymbol{\gamma}}' = (\hat{\gamma}_0, \hat{\boldsymbol{\beta}}_1')$,

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= (\mathbf{Z}'\boldsymbol{\Phi}_n^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Phi}_n^{-1}\mathbf{y}, \\ (\bar{y}_{dc}, \bar{\mathbf{x}}_{1,dc}) &= \bar{z}_{0,N}(\mathbf{z}_0'\boldsymbol{\Phi}_n^{-1}\mathbf{z}_0)^{-1}\mathbf{z}_0'\boldsymbol{\Phi}_n^{-1}(\mathbf{y}, \mathbf{X}_1) \\ &= \bar{z}_{0,N}(\mathbf{w}'\mathbf{z}_0)^{-1}\mathbf{w}'(\mathbf{y}, \mathbf{X}_1),\end{aligned}$$

and

$$\bar{\mathbf{z}}_N = [\bar{z}_{0,N},\ \bar{\mathbf{x}}_{1,N} - \bar{z}_{0,N}(\mathbf{w}'\mathbf{z}_0)^{-1}\mathbf{w}'\mathbf{X}_1].$$

The elements of $(\bar{y}_{dc}, \bar{\mathbf{x}}_{1,dc})$ are design-consistent estimators under the assumptions of Corollary 2.2.3.1 because they are ratio estimators of the form (2.1.1) with the x-variable of (2.1.1) equal to z_0. For example, if $\boldsymbol{\Phi} = \mathbf{I}$ and $\mathbf{x}_i = (N^{-1}nw_i,\ \mathbf{x}_{1,i})$, the regression estimator (2.2.53) is

$$\bar{y}_{reg} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}, \tag{2.2.54}$$

where

$$(\bar{y}_\pi, \bar{\mathbf{x}}_\pi) = \hat{N}^{-1}\sum_{i\in A}\pi_i^{-1}(y_i, \mathbf{x}_i),$$

$$\hat{N} = \sum_{i\in A}\pi_i^{-1},$$

and

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1')' = \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The estimator (2.2.53) with $\bar{y}_{dc} = \bar{y}_\pi$ can be written

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}} + \hat{N}^{-1}\sum_{i\in A}\pi_i^{-1}(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}).$$

In this form it is sometimes called the *generalized regression estimator* (GREG) . See Cassel, Särndal, and Wretman (1976) and Särndal (1980).

To study the large-sample properties of the regression estimator, assume that

$$[(\bar{y}_{dc},\ \bar{\mathbf{x}}_{dc},\ \hat{\boldsymbol{\beta}}_1') - (\bar{y}_N,\ \bar{\mathbf{x}}_N,\ \boldsymbol{\beta}_{1,N}')] \mid \mathcal{F}_N = O_p(k_N),$$

where $k_N \to 0$ as $N \to \infty$. Then the regression estimator (2.2.53) satisfies

$$\begin{aligned}\bar{y}_{reg} - \bar{y}_N &= \bar{y}_{dc} - \bar{y}_N + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,dc})\hat{\boldsymbol{\beta}}_1 \\ &= \bar{y}_{dc} - \bar{y}_N + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,dc})\boldsymbol{\beta}_{1,N} + O_p(k_N^2) \\ &= \bar{a}_{dc} + O_p(k_N^2),\end{aligned} \tag{2.2.55}$$

where $a_i = y_i - \bar{y}_N - (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})\boldsymbol{\beta}_{1,N}$. Observe that the finite population mean of the a_i of (2.2.55) is zero for any $\boldsymbol{\beta}_{1,N}$ and that very mild conditions are placed on $\hat{\boldsymbol{\beta}}_1$. In many situations k_N is $n^{-1/2}$.

A natural $\hat{\boldsymbol{\beta}}$ to use in constructing a regression estimator is the estimator weighted with the inverses of the selection probabilities,

$$\hat{\boldsymbol{\beta}}_\pi = \left(\sum_{i \in A} \mathbf{x}_i' \pi_i^{-1} \mathbf{x}_i \right)^{-1} \sum_{i \in A} \mathbf{x}_i' \pi_i^{-1} y_i. \tag{2.2.56}$$

We call $\hat{\boldsymbol{\beta}}_\pi$ the *probability weighted estimator* even though the weights are the inverses of the probabilities. The estimator (2.2.56) is of the form (2.2.52) with $\boldsymbol{\Phi} = \mathbf{D}_\pi$, where $\mathbf{D}_\pi$ is a diagonal matrix with the π_i on the diagonal. The estimator (2.2.56) is design consistent for

$$\boldsymbol{\beta}_N = \left(\sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i' y_i.$$

An estimator of the variance of $\hat{\boldsymbol{\beta}}_\pi$ is, by Theorem 2.2.1,

$$\hat{V}\{\hat{\boldsymbol{\beta}}_\pi \mid \mathcal{F}\} = \hat{\mathbf{M}}_{xx}^{-1} \hat{\mathbf{V}}_{\bar{b}\bar{b}} \hat{\mathbf{M}}_{xx}^{-1}, \tag{2.2.57}$$

where

$$\hat{\mathbf{M}}_{xx} = N^{-1} \sum_{i \in A} \mathbf{x}_i' \pi_i^{-1} \mathbf{x}_i,$$

$$\hat{\mathbf{V}}_{\bar{b}\bar{b}} = \hat{V} \left\{ N^{-1} \sum_{i \in A} \mathbf{x}_i' \pi_i^{-1} a_i \right\},$$

$a_i = y_i - \mathbf{x}_i \boldsymbol{\beta}_N$, $\hat{\mathbf{V}}_{\bar{b}\bar{b}}$ is a consistent estimator of variance calculated with $\hat{a}_i$ replacing a_i, and $\hat{a}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_\pi$.

The variance estimation approach of (2.2.57) is applicable for all regression estimators of type (2.2.51). The estimator of variance is

$$\begin{aligned} \hat{V}\{\bar{y}_{reg} \mid \mathcal{F}\} &= \bar{\mathbf{x}}_N \hat{V}\{\hat{\boldsymbol{\beta}} \mid \mathcal{F}\} \bar{\mathbf{x}}_N' \\ &= \hat{V} \left\{ \sum_{i \in A} \tilde{w}_i a_i \mid \mathcal{F} \right\}, \end{aligned} \tag{2.2.58}$$

where $\hat{V}\{\Sigma_{i \in A} \tilde{w}_i a_i \mid \mathcal{F}\}$ is a consistent estimator, such as the Horvitz–Thompson estimator, computed with $\hat{a}_i$ replacing a_i, and

$$\tilde{\mathbf{w}} = (\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_n)' = \bar{\mathbf{x}}_N (\mathbf{X}' \boldsymbol{\Phi}_n^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Phi}_n^{-1}.$$

Multiplying (2.2.58) by $n(n-k-1)^{-1}$ will often reduce the small-sample bias of the variance estimator.

Example 2.2.1. The NRI data in Table 2.8 were used to illustrate ratio estimation in Example 2.1.1. It happens that the area of federally owned land and the total area are available from external sources. Therefore, the variables "segment size" and "federal acres" can be used as auxiliary variables in a regression estimator. We include indicators for strata in the regression so that the "estimated" stratum sizes are equal to the known sizes. The stratum variables have little effect on the estimated variance for the estimation of acres of cultivated cropland, but it is felt that they might have importance for variables related to geography. Let the indicator for stratum h be

$$\begin{aligned} x_{hi} &= 1 \quad \text{if segment } i \text{ is in stratum } h \\ &= 0 \quad \text{otherwise} \end{aligned}$$

for $h = 1, 2, 3$. Let x_{4i} be total acres in segment i and let x_{5i} be federal acres for segment i, where federal acres is the segment acres multiplied by the fraction of points that are federally owned. The estimated regression with acres of cultivated cropland as the dependent variable is

$$\begin{array}{rcccccccc} \hat{y}_i = & 26.48x_{1i} & + & 39.59x_{2i} & + & 4.47x_{3i} & + & 0.22x_{4i} & - & 0.49x_{5i}, \\ & (51.75) & & (48.57) & & (48.70) & & (0.31) & & (0.10) \end{array}$$

where (x_1, x_2, x_3) is the vector of stratum indicators, x_4 is segment size, and x_5 is federal acres. The estimated standard errors are calculated as the square roots of the diagonal elements of the estimated covariance matrix defined in (2.2.57). For example, the computations can be carried out in SAS or in STATA. The sample is a two-stage sample, and ownership is determined only at the points observed. Therefore, the value of federal acres for a segment is an estimated value, and the expected values for the coefficients are affected by this sampling error. See Section 5.6. For the purposes of estimating the total y, we can ignore the within-segment sampling effect. See Section 2.6.

The estimated total for cultivated cropland is

$$\hat{T}_{y,reg} = \mathbf{T}_x\hat{\boldsymbol{\beta}} = (990, 1155, 442, 437.1, 272)\hat{\boldsymbol{\beta}} = 156.9,$$

where all values are thousands of acres. An estimated variance is

$$\hat{V}\{\hat{T}_{y,reg} \mid \mathcal{F}\} = \mathbf{T}_x\hat{V}\{\hat{\boldsymbol{\beta}} \mid \mathcal{F}\}\mathbf{T}_x' = 279.8.$$

The estimated total of cultivated cropland calculated as the direct-expansion stratified estimator is

$$\hat{T}_y = \sum_{i \in A} \pi_i^{-1} y_i = 153.8,$$

Table 2.1 Alternative Estimates for Missouri County

Characteristic	Stratified Estimator	Ratio Estimator	Regression Estimator
Cultivated cropland	153.7	149.0	156.9
	(18.3)	(17.9)	(16.7)
Forest	76.4	74.1	74.7
	(14.6)	(13.8)	(13.7)
Other nonfederal	181.1	175.5	178.3
	(18.5)	(17.2)	(17.2)
Federal	39.7	38.5	27.2
	(13.1)	(12.6)	(0)
Total	450.7	437.1	437.1
	(12.8)	(0)	(0)

and the estimated variance for the stratified estimator is 336.2.

Table 2.1 contains the stratified estimator, the ratio estimator, and the regression estimator for four broaduses. The table is a simple form of the type of table produced with the NRI data. Typically, the reports focus on nonfederal lands. All three estimators are linear in y, so the sum of the four broaduse estimates is the estimate of the total. It is interesting that the addition of the known acres of federally owned land to the estimation procedure has a differential effect on the estimates for the different broaduses. The regression estimate for cultivated cropland is larger than the stratified estimate, whereas the regression estimate for forest is smaller.

The gains estimated for the regression estimator relative to the stratified estimator range from 14% for forest to 20% for cultivated cropland. In general, gains will be larger for estimates of large quantities because the variance of the total area is zero for the regression estimator. The regression estimator has a large-sample variance that is never greater than the large-sample variance of the ratio estimator. In this example the efficiency estimated for the regression estimator relative to the ratio estimator ranges from 100% to 115%. ■■

The regression estimator $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}_\pi$ with $\hat{\boldsymbol{\beta}}_\pi$ of (2.2.56) is design consistent for $\bar{y}_N$ if the column of 1's is in the column space of $\mathbf{X}$. While the estimator is design consistent, it is not necessarily the minimum variance estimator. In Theorem 2.2.4 we define a regression estimator in the class (2.2.54) with the

smallest large-sample design variance. Note that the estimator (2.2.63) of Theorem 2.2.4 is linear in y when $\hat{C}\{\bar{\mathbf{x}}_{1,\pi},\ \bar{y}_\pi \mid \mathcal{F}_N\}$ is linear in y. We define the vector of auxiliary variables so that the covariance matrix of the estimated mean vector is nonsingular.

Theorem 2.2.4. Let $\{\mathbf{z}_j\}$ be a sequence of real vectors where $\mathbf{z}_j = (y_i, \mathbf{x}_{1j})$. Let $\mathcal{F}_N = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)$ for $N > N_0$. Assume:

(i) Positive K_U and K_L exist such that

$$K_L < n\boldsymbol{\gamma}' V\{\bar{\mathbf{z}}_\pi \mid \mathcal{F}_N\}\boldsymbol{\gamma} < K_U \tag{2.2.59}$$

for all N and any vector $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| = 1$.

(ii) The sequence is such that

$$[V\{\bar{\mathbf{z}}_\pi \mid \mathcal{F}_N\}]^{-1/2}(\bar{\mathbf{z}}_\pi - \bar{\mathbf{z}}_N) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{I}). \tag{2.2.60}$$

(iii) There is a quadratic estimator of the variance of $\bar{\mathbf{z}}_\pi$, denoted by $\hat{V}\{\bar{\mathbf{z}}_\pi|\mathcal{F}_N\}$, satisfying

$$[V\{\bar{\mathbf{z}}_\pi \mid \mathcal{F}_N\}]^{-1}\hat{V}\{\bar{\mathbf{z}}_\pi \mid \mathcal{F}_N\} - \mathbf{I} = O_p(n^{-\kappa}) \tag{2.2.61}$$

for some $\kappa > 0$.

(iv) The variance $\hat{V}\{\bar{\mathbf{z}}_\pi|\mathcal{F}_N\}$ is nonsingular with probability 1 for all N greater than some N_0.

Define

$$\bar{y}_{D,reg} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_{1,D}, \tag{2.2.62}$$

where

$$\hat{\boldsymbol{\beta}}_{1,D} = [\hat{V}\{\bar{\mathbf{x}}_{1,\pi} \mid \mathcal{F}_N\}]^{-1}\hat{C}\left\{\bar{\mathbf{x}}'_{1,\pi},\ \bar{y}_\pi \mid \mathcal{F}_N\right\}. \tag{2.2.63}$$

Then

$$n_N^{1/2}(\bar{y}_{D,reg} - \bar{y}_N) \xrightarrow{\mathcal{L}} N(0,\ V_{e,\infty}), \tag{2.2.64}$$

where $V_{e,\infty}$ is the minimum variance for the limiting distribution of design-consistent estimators of the form (2.2.54).

Also,

$$[\tilde{V}\{\bar{a}_\pi \mid \mathcal{F}_N\}]^{-1/2}(\bar{y}_{D,reg} - \bar{y}_N) \xrightarrow{\mathcal{L}} N(0,\ 1), \tag{2.2.65}$$

where $\tilde{V}\{\bar{a}_\pi \mid \mathcal{F}_N\}$ is the estimator of (2.2.61) constructed with

$$\hat{a}_i = y_i - \bar{y}_\pi - (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})\hat{\boldsymbol{\beta}}_{1,D}$$

in place of $a_i = y_i - \bar{y}_N - (\mathbf{x}_{1,i} - \bar{x}_{1,\pi})\boldsymbol{\beta}_{1,DN}$ and

$$\boldsymbol{\beta}_{1,DN} = [V\{\bar{\mathbf{x}}_{1,\pi} \mid \mathcal{F}_N\}]^{-1} C\{\bar{\mathbf{x}}'_{1,\pi}, \bar{y}_\pi \mid \mathcal{F}_N\}.$$

Proof. By assumption (2.2.61), the estimated covariance matrix for sample means is design consistent for the true covariance matrix. Therefore,

$$\hat{\boldsymbol{\beta}}_{1,D} - \boldsymbol{\beta}_{1,DN} \mid \mathcal{F}_N = O_p(n_N^{-\kappa}) \tag{2.2.66}$$

and

$$\bar{y}_{D,reg} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\boldsymbol{\beta}_{1,DN} + O_p(n^{-0.5-\kappa}).$$

It follows from (2.2.60) that $n^{1/2}(\bar{y}_{D,reg} - \bar{y}_N)$ has a limiting normal distribution. Now $\boldsymbol{\beta}_{1,DN}$ is the $\boldsymbol{\beta}$ that minimizes

$$V\{\bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\boldsymbol{\beta} \mid \mathcal{F}_N\}$$

and there is no $\boldsymbol{\beta}$ that will give a smaller variance for the limit distribution.

To prove (2.2.65), we write the quadratic estimator of the variance of $\bar{a}_\pi$ as

$$\hat{V}\{\bar{a}_\pi \mid \mathcal{F}_N\} = \sum_{i\in A}\sum_{j\in A} \omega_{ij} a_i a_j,$$

where ω_{ij} are the coefficients. If a_i is replaced with $\hat{a}_i$, where $\hat{a}_i$ is defined for (2.2.65),

$$\tilde{V}\{\bar{a}_\pi \mid \mathcal{F}_N\} = \sum_{i\in A}\sum_{j\in A} \omega_{ij}(a_i a_j + \Delta_{\pi,ij}),$$

where

$$\Delta_{\pi,ij} = a_i \mathbf{c}_j \boldsymbol{\delta}_\pi + a_j \mathbf{c}_i \boldsymbol{\delta}_\pi + \boldsymbol{\delta}'_\pi \mathbf{c}'_i \mathbf{c}_j \boldsymbol{\delta}_\pi,$$

$\hat{a}_i = a_i - \mathbf{c}_i\boldsymbol{\delta}_\pi$, $\mathbf{c}_i = (1,\ \mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})$, and $\boldsymbol{\delta}'_\pi = (\bar{a}_\pi,\ \hat{\boldsymbol{\beta}}'_{1,D} - \boldsymbol{\beta}'_{1,DN})$. Now

$$\sum_{i\in A}\sum_{j\in A} \omega_{ij}\Delta_{\pi,ij} = O_p(n^{-1-\kappa})$$

because, for example,

$$\sum_{i\in A}\sum_{j\in A} \omega_{ij} a_i (\mathbf{x}_{1,j} - \bar{\mathbf{x}}_{1,N})(\hat{\boldsymbol{\beta}}_{1,D} - \boldsymbol{\beta}_{1,DN}) = \hat{C}\{\bar{a}_\pi,\ \bar{\mathbf{x}}_\pi \mid \mathcal{F}_N\}(\hat{\boldsymbol{\beta}}_{1,D} - \boldsymbol{\beta}_{1,DN})$$

is $O_p(n^{-1-\kappa})$ by assumptions (2.2.59) and (2.2.61). Therefore, by (2.2.66) and (2.2.61),

$$[V\{\bar{a}_\pi \mid \mathcal{F}_N\}]^{-1}\tilde{V}\{\bar{a}_\pi \mid \mathcal{F}_N\} - 1 = O_p(n^{-1-\kappa})$$

and (2.2.65) follows. ■

In a large-sample sense, Theorem 2.2.4 answers the question of how to construct a regression estimator with the minimum design variance for a member of a particular class. Given estimators of $(\bar{y}_N, \bar{\mathbf{x}}_N)$, the theorem gives the β that minimizes the large-sample variance. We used the estimator $(\bar{y}_\pi, \bar{\mathbf{x}}_\pi)$ because these estimators are location invariant. One could use other estimators, such as $(\bar{y}_{HT}, \bar{\mathbf{x}}_{HT})$, but the resulting estimator may not be location invariant. In practice, a number of questions remain. Theorem 2.2.4 assumes a large sample and a vector $\mathbf{x}$ of fixed dimension. If a large number of auxiliary variables are included in the regression, terms excluded in the large-sample approximation become important. This can be true if the number of primary sampling units in the sample is small and hence the number of degrees of freedom in $\hat{V}\{\bar{\mathbf{z}}_{1,n} \mid \mathcal{F}\}$ is small. With a small number of degrees of freedom, terms ignored in the approximations can become important. The variance contribution due to estimating β can be large, and it is possible for some of the weights defined by the regression procedure to be negative. If weights are negative, estimates of quantities known to be positive can be negative. Negative weights are discussed in Section 2.8.

We have introduced regression estimation for the mean, but it is often the totals that are estimated and totals that are used as controls. Consider the regression estimator of the total of y defined by

$$\hat{T}_{y,reg} = \hat{T}_{y,\pi} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi})\hat{\beta}_{y\cdot x}, \tag{2.2.67}$$

where $\mathbf{T}_{x,N}$ is the known total of x, and $(\hat{T}_{y,\pi}, \hat{\mathbf{T}}_{x,\pi})$ is the vector of design unbiased estimators of $(T_{y,N}, \mathbf{T}_{x,N})$. By analogy to Theorem 2.2.4, the estimator of the optimum β is

$$\hat{\beta}_{y\cdot x} = [\hat{V}\{\hat{\mathbf{T}}_{x,\pi} \mid \mathcal{F}\}]^{-1}\hat{C}\{\hat{\mathbf{T}}_{x,\pi}, \hat{T}_{y,\pi} \mid \mathcal{F}\}, \tag{2.2.68}$$

where $\hat{V}\{\hat{\mathbf{T}}_{x,\pi} \mid \mathcal{F}\}$ is a design-consistent estimator of the variance of $\hat{\mathbf{T}}_{x,\pi}$ and $\hat{C}\{\hat{\mathbf{T}}_{x,\pi}, \hat{T}_{y,\pi} \mid \mathcal{F}\}$ is a design-consistent estimator of the covariance of $\hat{\mathbf{T}}_{x,\pi}$ and $\hat{T}_{y,\pi}$.

The estimator of the total is $N\bar{y}_{reg}$ for simple random sampling, but the exact equivalence may not hold for other designs. In more complicated samples, the mean is a ratio estimator. However, if the regression estimator of the two totals is constructed using Theorem 2.2.4, the ratio of the two estimated totals has large-sample variance equal to that of the regression-estimator of the ratio. To see this, write the error in the regression-estimated totals of y and u as

$$\hat{T}_{y,reg} - T_{y,N} = \hat{T}_{y,\pi} - T_{y,N} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi})\beta_{y\cdot x} + O_p(Nn_N^{-1})$$

and

$$\hat{T}_{u,reg} - T_{u,N} = \hat{T}_{u,\pi} - T_{u,N} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi})\boldsymbol{\beta}_{u\cdot x} + O_p(Nn_N^{-1}), \quad (2.2.69)$$

where we are assuming that $\hat{T}_{y,\pi} - T_{y,N}$, $\hat{\boldsymbol{\beta}}_{y\cdot x} - \boldsymbol{\beta}_{y\cdot x}$, and the corresponding quantities for u are $O_p(Nn_N^{-1/2})$ and $O_p(n_N^{-1/2})$, respectively. Then the error in $\hat{T}_{u,reg}^{-1}\hat{T}_{y,reg}$ is

$$\hat{T}_{u,reg}^{-1}\hat{T}_{y,reg} - T_{u,N}^{-1}T_{y,N} = T_{u,N}^{-1}[(\hat{T}_{y,\pi} - T_{y,N}) - R_N(\hat{T}_{u,\pi} - T_{u,N}) + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi})(\boldsymbol{\beta}_{y\cdot x} - R_N\boldsymbol{\beta}_{u\cdot x})] + O_p(Nn_N^{-1}), \quad (2.2.70)$$

where $R_N = T_{u,N}^{-1}T_{y,N}$. If we construct the regression estimator for R_N starting with $\hat{R} = \hat{T}_{u,\pi}^{-1}\hat{T}_{y,\pi}$, we have

$$\hat{R}_{reg} = \hat{R} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi})\hat{\boldsymbol{\beta}}_{R\cdot x}, \quad (2.2.71)$$

where

$$\hat{\boldsymbol{\beta}}_{R\cdot x} = [\hat{V}\{\hat{\mathbf{T}}_{x,\pi} \mid \mathcal{F}\}]^{-1}\hat{C}\{\hat{\mathbf{T}}'_{x,\pi}, \hat{R} \mid \mathcal{F}\}$$

and

$$\hat{C}\{\hat{\mathbf{T}}_{x,\pi}, \hat{R} \mid \mathcal{F}\} = \hat{C}\{\hat{\mathbf{T}}_{x,\pi}, T_{u,N}^{-1}(\hat{T}_{y,\pi} - R_N\hat{T}_{u,\pi}) \mid \mathcal{F}\}.$$

It follows that the large-sample design-optimum coefficient for the ratio is $T_{u,N}^{-1}(\boldsymbol{\beta}_{y\cdot x} - R_N\boldsymbol{\beta}_{u\cdot x})$, and the ratio of design-optimum regression estimators is the large-sample design-optimum regression estimator of the ratio.

2.2.3 Poststratification

One type of regression estimation is so important that it deserves special discussion. Assume that after the sample is observed the elements of the sample can be assigned to mutually exclusive and exhaustive categories for which the population totals are known. Assume that the categories are defined prior to sample selection. When the categories are used to construct an estimator, they are called *poststrata*. If the original sample is a simple random sample, the sample observed in each poststratum is a simple random sample of elements in that poststratum. Given a particular sample realization of simple random sampling, a very natural procedure is to use the stratified estimator,

$$\bar{y}_{ps} = \sum_{h=1}^{H} N^{-1}N_h\bar{y}_h, \quad (2.2.72)$$

where $\bar{y}_h$ is the sample mean for the hth poststratum. Conditional on the sample sizes $(n_1, n_2, \ldots, n_H)$, all $n_h > 0$, the estimator is unbiased for the population mean. Also, if all $n_h > 0$, the conditional variance of the estimator is the variance of the stratified estimator,

$$V\{\bar{y}_{ps} - \bar{y}_N \mid \mathcal{F}, (n_1, n_2, \ldots, n_H)\} = \sum_{h=1}^{H} (N^{-1}N_h)^2 N_h^{-1}(N_h - n_h)n_h^{-1}S_h^2 \qquad (2.2.73)$$

as defined in equation (1.2.55).

Because the sample in each stratum is a simple random sample of elements in that stratum,

$$s_h^2 = (n_h - 1)^{-1} \sum_{j \in A_h} (y_{hj} - \bar{y}_h)^2$$

is unbiased for S_h^2 and

$$\hat{V}\{\bar{y}_{ps} - \bar{y}_N \mid \mathcal{F}, (n_1, n_2, \ldots, n_H)\} = \sum_{h=1}^{H} (N^{-1}N_h)^2 N_h^{-1}(N_h - n_h)n_h^{-1}s_h^2 \qquad (2.2.74)$$

is unbiased for the conditional variance, given that all $n_h \geq 2$.

To see that the poststratified estimator for simple random sampling is a special case of regression estimation, let $\mathbf{x}_i$ be the vector with H elements x_{hi}, where

$$\begin{aligned} x_{hi} &= 1 \quad \text{if element } i \text{ is in poststratum } h \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Note that the vector of 1's is in the column space of the matrix $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_h')'$ because $\Sigma_h x_{hi} = 1$ for all i. The vector of regression coefficients for the regression of y on $\mathbf{x}$ is

$$\hat{\boldsymbol{\beta}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_H)$$

and the sample mean of $\mathbf{x}_i$ is

$$\bar{\mathbf{x}}_n = n^{-1}(n_1, n_2, \ldots, n_H),$$

where $n^{-1}n_h$ is the fraction of the sample that falls in stratum h. The population mean of $\mathbf{x}_i$ is

$$\bar{\mathbf{x}}_N = N^{-1}(N_1, N_2, \ldots, N_H).$$

Thus, the regression estimator is

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}} = \sum_{h=1}^{H} N^{-1} N_h \bar{y}_h = \bar{y}_{ps}. \tag{2.2.75}$$

Using the variance estimator (2.2.47) in the variance for the linear combination $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ gives an alternative variance estimator for the regression estimator (2.2.75). The difference between (2.2.47) and (2.2.74) depends on the degrees-of-freedom adjustment made in (2.2.47). Generally, (2.2.74) is the preferred estimator.

The regression form of the poststratified estimator can be used for any probability sample and the variance estimator of the regression estimator remains appropriate.

2.2.4 Residuals for variance estimation

We noted that estimator (2.2.47) generally underestimates the variance of regression coefficients. For many situations this bias is not serious, but it can be important in small samples when the regression coefficients are of subject matter interest. An improved estimator of variance can be obtained by using alternative regression residuals. Consider the model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \\ \mathbf{e} &\sim \left(\mathbf{0}, \mathbf{I}\sigma^2\right). \end{aligned}$$

Let an estimator of $\boldsymbol{\beta}$ be constructed by giving the ith observation a weight of g_i and write the estimator as

$$\tilde{\boldsymbol{\beta}}_{(i)} = \left(\mathbf{X}'\mathbf{G}_i\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{G}_i\mathbf{y}, \tag{2.2.76}$$

where $\mathbf{G}_i = \text{diag}(1,\ 1,\ \ldots,\ 1,\ g_i,\ 1,\ \ldots,\ 1)$ and g_i is in the ith position. Let the residual for observation i be

$$\tilde{e}_i = y_i - \mathbf{x}_i\tilde{\boldsymbol{\beta}}_{(i)} = e_i - \mathbf{x}_i\left(\mathbf{X}'\mathbf{G}_i\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{G}_i\mathbf{e}. \tag{2.2.77}$$

Then

$$\begin{aligned} \sigma^{-2}V\left\{\tilde{e}_i \mid \mathbf{X}\right\} &= 1 - 2\mathbf{x}_i\mathbf{A}_i^{-1}\mathbf{x}_i' g_i + \mathbf{x}_i\mathbf{A}_i^{-1}\mathbf{X}'\mathbf{G}_i^2\mathbf{X}\mathbf{A}_i^{-1}\mathbf{x}_i' \\ &= 1 - 2\mathbf{x}_i\mathbf{A}_i^{-1}\mathbf{x}_i' g_i + \mathbf{x}_i\mathbf{A}_i^{-1}\mathbf{x}_i' - \left(\mathbf{x}_i\mathbf{A}_i^{-1}\mathbf{x}_i'\right)^2\left(g_i - g_i^2\right), \end{aligned} \tag{2.2.78}$$

where $\mathbf{X}'\mathbf{G}_i\mathbf{X} = \mathbf{A}_i$. By replacing $\mathbf{I}\sigma^2$ with $\text{diag}(\sigma_i^2)$, one can see the importance of the assumption that $\mathbf{e} \sim \left(\mathbf{0}, \mathbf{I}\sigma^2\right)$ in obtaining (2.2.78). Setting

the right side of (2.2.78) equal to 1, it is possible to use iterative methods to determine a g_i so that the variance of $\tilde{e}_i$ is σ^2. Because $\left(\mathbf{x}_i \mathbf{A}_i^{-1} \mathbf{x}_i'\right) = O_p\left(n^{-1}\right)$, for many problems, 0.5 furnishes a good approximation to the desired g_i. A computational form for $\left(\mathbf{X}'\mathbf{G}_i\mathbf{X}\right)^{-1}$ that requires no additional matrix inverse calculations is

$$\left(\mathbf{X}'\mathbf{G}_i\mathbf{X}\right)^{-1} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} + h_i \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_i'\mathbf{x}_i \left(\mathbf{X}'\mathbf{X}\right)^{-1}, \qquad (2.2.79)$$

where $h_i = g_i(1 - g_i\mathbf{x}_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_i)^{-1}$.

2.3 MODELS AND REGRESSION ESTIMATION

In Section 2.2.1 we introduced the regression estimator of the finite population mean using the normal regression model and simple random nonreplacement sampling. It was demonstrated in Section 2.2.2 that a vector of regression coefficients can be used to construct an estimator of the population mean that has minimum design variance, in large samples, for estimators in a class of design-consistent regression estimators. In this section we explore the use of models to construct regression estimators.

Given information about the population, it is natural to formulate a model relating the characteristics of interest to the information available, but the way in which the model is used to create estimators remains an area of discussion in the survey literature. This is because estimators constructed under a model need not be design consistent. If the estimator is consistent for the parameter under the randomization distribution, it can be asserted that the estimator is consistent even if the model is incorrect. In some cases, the estimator constructed to be optimal under the model will automatically satisfy the design-consistency requirement. We call a model for which the standard model estimator is design consistent a *full model.* A model for which the model estimator is not design consistent is called a *restricted model* or *reduced model.* It is understood that the definition of full model depends on the sample design. Estimators that are not design consistent are called *model dependent* in the survey sampling literature.

2.3.1 Linear models

In this section we consider linear models and show how to construct estimators with good model properties that also satisfy a condition for design consistency.

Assume that a sample is selected using a design such that the linear-in-y estimator

$$\bar{y}_\pi = \sum_{i \in A} a_i y_i, \qquad (2.3.1)$$

where the a_i are weights, is design consistent for the mean. Assume that the elements of the finite population were generated by the model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i, \tag{2.3.2}$$

where $\mathbf{x}_i = (1, \mathbf{x}_{1,i})$, $\boldsymbol{\beta}' = (\beta_0, \boldsymbol{\beta}_1')$, y_i is the observation on the characteristic of interest for the ith element, $\mathbf{x}_{1,i}$ is the vector of observations on an auxiliary variable, and e_i is a zero mean error that is independent of $\mathbf{x}_{1,j}$ for all i and j. Assume that the finite population mean of $\mathbf{x}_{1,i}$ is known and equal to $\bar{\mathbf{x}}_{1,N}$.

Given that the weighted sample mean $(\bar{y}_\pi, \bar{\mathbf{x}}_{1,\pi})$ is consistent for the population mean, we have shown that the estimator

$$\bar{y}_{reg} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_1 \tag{2.3.3}$$

is a design-consistent estimator of the population mean of y for any $\hat{\boldsymbol{\beta}}_1$ that is a design-consistent estimator of a constant. The $\hat{\boldsymbol{\beta}}_1$ that minimizes the large-sample design variance was defined in Theorem 2.2.4.

Assume now that a superpopulation model is postulated for the data. Assume also that the sample is an unequal probability sample or (and) the specified error covariance structure is not a multiple of the identity matrix. Only in special cases will the estimator of Theorem 2.2.4 agree with the best estimator constructed under the model, conditioning on the sample $\mathbf{x}$-values. To investigate this possible conflict, write the model for the population in matrix notation as

$$\begin{aligned} \mathbf{y}_N &= \mathbf{X}_N\boldsymbol{\beta} + \mathbf{e}_U, \\ \mathbf{e}_N &\sim (\mathbf{0}, \boldsymbol{\Sigma}_{eeNN}), \end{aligned} \tag{2.3.4}$$

where $\mathbf{y}_N = (y_1, y_2, \ldots, y_N)'$, $\mathbf{e}_N = (e_1, e_2, \ldots, e_N)'$, and $\mathbf{e}_N$ is independent of

$$\mathbf{X}_N = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_N')'.$$

It is assumed that $\boldsymbol{\Sigma}_{eeNN}$ is known or known up to a multiple.

If the selection indicators are independent of the vector $\mathbf{e}_N$, the model for a sample of n observations is of the same form as (2.3.4) and can be written

$$\begin{aligned} \mathbf{y}_A &= \mathbf{X}_A\boldsymbol{\beta} + \mathbf{e}_A, \\ \mathbf{e}_A &\sim (\mathbf{0}, \boldsymbol{\Sigma}_{eeAA}), \end{aligned} \tag{2.3.5}$$

where $\mathbf{e}_A$ is independent of $\mathbf{X}_A$, $\mathbf{y}_A = (y_1, y_2, \ldots, y_n)'$, $\mathbf{e}_A = (e_1, e_2, \ldots, e_n)'$,

$$\mathbf{X}_A = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')',$$

and we index the sample elements by $1, 2, \ldots, n$, for convenience. We have used the subscript N to identify population quantities and the subscript A to identify sample quantities, but we will often omit the subscript A to simplify the notation. For example, we may sometimes write the $n \times n$ covariance matrix of $\mathbf{e}_A$ as $\boldsymbol{\Sigma}_{ee}$.

The unknown finite population mean of y is

$$\bar{y}_N = \bar{\mathbf{x}}_N \boldsymbol{\beta} + \bar{\mathbf{e}}_N,$$

and under model (2.3.5), the best linear-in-y, conditionally unbiased predictor of $\theta_N = \bar{y}_N$, conditional on $\mathbf{X}_A$, is

$$\hat{\theta} = N^{-1}\left(\sum_{i \in A} y_i + (N-n)\bar{\mathbf{x}}_{N-n}\hat{\boldsymbol{\beta}} + \mathbf{J}'_{N-n}\boldsymbol{\Gamma}_{\bar{A}A}(\mathbf{y}_A - \mathbf{X}_A\hat{\boldsymbol{\beta}})\right), \tag{2.3.6}$$

where $\boldsymbol{\Gamma}_{\bar{A}A} = \boldsymbol{\Sigma}_{ee\bar{A}A}\boldsymbol{\Sigma}^{-1}_{eeAA}$, $\bar{\mathbf{x}}_{N-n} = (N-n)^{-1}(N\bar{\mathbf{x}}_N - n\bar{\mathbf{x}}_n)$,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'_A\boldsymbol{\Sigma}^{-1}_{eeAA}\mathbf{X}_A)^{-1}\mathbf{X}'_A\boldsymbol{\Sigma}^{-1}_{eeAA}\mathbf{y}_A,$$

$\boldsymbol{\Sigma}_{ee\bar{A}A} = E\{\mathbf{e}_{\bar{A}}\mathbf{e}'_A\}$, $\mathbf{e}_{\bar{A}} = (e_{n+1}, e_{n+2}, \ldots, e_N)'$, $\mathbf{J}_{N-n}$ is an $(N-n)$-dimensional column vector of 1's, $\bar{\mathbf{x}}_n$ is the simple sample mean, and $\bar{A}$ is the set of elements in U that are not in A.

Under model (2.3.5),

$$\hat{\theta} - \bar{y}_N = \mathbf{C}_{x\bar{A}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + N^{-1}\mathbf{J}'_{N-n}(\boldsymbol{\Gamma}_{\bar{A}A}\mathbf{e}_A - \mathbf{e}_{\bar{A}})$$

and

$$\begin{aligned} V\{\hat{\theta} - \bar{y}_N \mid \mathbf{X}_A\} &= \mathbf{C}_{x\bar{A}}V\{\hat{\boldsymbol{\beta}} \mid \mathbf{X}_A\}\mathbf{C}'_{x\bar{A}} \\ &\quad + N^{-2}\mathbf{J}'_{N-n}\left(\boldsymbol{\Sigma}_{ee\bar{A}\bar{A}} - \boldsymbol{\Gamma}_{\bar{A}A}\boldsymbol{\Sigma}_{eeA\bar{A}}\right)\mathbf{J}_{N-n}, \end{aligned} \tag{2.3.7}$$

where $\boldsymbol{\Sigma}_{eeA\bar{A}} = \boldsymbol{\Sigma}'_{ee\bar{A}A}$ and

$$\mathbf{C}_{x\bar{A}} = N^{-1}\left[(N-n)\bar{\mathbf{x}}_{N-n} - \mathbf{J}'_{N-n}\boldsymbol{\Gamma}_{\bar{A}A}\mathbf{X}_A\right].$$

The conditional variance (2.3.7) based on the assumption that selection probabilities are independent of $\mathbf{e}_A$ is sometimes called model variance in the survey literature. The predictor (2.3.6) will be design consistent for the finite population mean if the design probabilities, the matrix $\boldsymbol{\Sigma}_{eeNN}$, and the matrix $\mathbf{X}_N$ meet certain conditions.

Theorem 2.3.1. Let the superpopulation model be (2.3.4), where the $\mathbf{x}_i$, $i = 1, 2, \ldots, N$, are *iid* random variables with finite fourth moments. Assume

a sequence of populations, designs, and estimators such that

$$N^{-1}\left(\sum_{i\in A}\pi_i^{-1}(y_i,\ \mathbf{x}_i) - (T_{yN},\ \mathbf{T}_{xN})\right)\Big|\mathcal{F}_N \ = \ O_p(n_N^{-\alpha}) \quad \text{a.s.}, \tag{2.3.8}$$

where the π_i are the selection probabilities, T_{yN} is the total of y for the Nth population, and $\alpha > 0$. Let $\hat{\boldsymbol{\beta}}$ be defined by (2.3.6) and let $\{\boldsymbol{\beta}_N\}$ be a sequence of finite population parameters such that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \mid \mathcal{F}_N \ = \ O_p(n_N^{-\alpha}) \quad \text{a.s.} \tag{2.3.9}$$

(i) Assume that there is a sequence $\{\boldsymbol{\gamma}_N\}$ such that

$$\mathbf{X}_A\boldsymbol{\gamma}_N \ = \ \boldsymbol{\Sigma}_{eeAA}\mathbf{D}_\pi^{-1}\mathbf{J}_n, \tag{2.3.10}$$

where $\mathbf{D}_\pi = \text{diag}(\pi_1,\ \pi_2,\ \ldots,\ \pi_n)'$, for every sample from U_N that is possible under the design. Then

$$(\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}} - \bar{y}_N) \mid \mathcal{F}_N \ = \ O_p(n_N^{-\alpha}) \quad \text{a.s.} \tag{2.3.11}$$

(ii) Assume that there is a sequence $\{\boldsymbol{\eta}_N\}$ such that

$$\mathbf{X}_A\boldsymbol{\eta}_N \ = \ \boldsymbol{\Sigma}_{eeAA}\mathbf{J}_n + \boldsymbol{\Sigma}_{eeA\bar{A}}\mathbf{J}_{N-n} \tag{2.3.12}$$

for all samples with positive probability. Then $\hat{\theta}$ of (2.3.6) satisfies

$$\hat{\theta} \ = \ \bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}. \tag{2.3.13}$$

(iii) Assume that there is a sequence $\{\boldsymbol{\zeta}_N\}$ such that

$$\mathbf{X}_A\boldsymbol{\zeta}_N \ = \ \boldsymbol{\Sigma}_{eeAA}(\mathbf{D}_\pi^{-1}\mathbf{J}_n - \mathbf{J}_n) - \boldsymbol{\Sigma}_{eeA\bar{A}}\mathbf{J}_{N-n} \tag{2.3.14}$$

for all samples with positive probability. Then $\hat{\theta}$ of (2.3.6) is expressible as

$$\hat{\theta} \ = \ \bar{y}_{HT} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})\hat{\boldsymbol{\beta}} \tag{2.3.15}$$

and

$$(\hat{\theta} - \bar{y}_N) \mid \mathcal{F}_N \ = \ \bar{a}_{HT} + O_p(n^{-2\alpha}) \quad \text{a.s.}, \tag{2.3.16}$$

where $a_{iN} = y_i - \bar{y}_N - (\mathbf{x}_i - \bar{\mathbf{x}}_N)\boldsymbol{\beta}_N$.

Proof. We sometimes omit subscripts on the sample quantities. Given (2.3.10), by Corollary 2.2.3.1, the regression estimator $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ is design consistent. By assumption (2.3.9),

$$\begin{aligned}\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}} \mid \mathcal{F}_N &= \bar{\mathbf{x}}_N\boldsymbol{\beta}_N + O_p(n_N^{-\alpha}) \quad \text{a.s.} \\ &= \bar{y}_N + O_p(n_N^{-\alpha}) \quad \text{a.s.}\end{aligned}$$

and result (2.3.11) is established.

The predictor (2.3.6) can be written

$$\hat{\theta} = N^{-1}[N\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}} + \mathbf{J}_n'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \mathbf{J}_{N-n}'\boldsymbol{\Sigma}_{ee\bar{A}A}\boldsymbol{\Sigma}_{eeAA}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})],$$

and if (2.3.12) is satisfied,

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{eeAA}^{-1}(\boldsymbol{\Sigma}_{eeAA}\mathbf{J}_n + \boldsymbol{\Sigma}_{ee\bar{A}\bar{A}}\mathbf{J}_{N-n}) = 0.$$

Hence, $\hat{\theta} = \bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$, and (2.3.13) holds.

If (2.3.14) is satisfied, we have

$$\begin{aligned}0 &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\left[(\mathbf{D}_\pi^{-1}\mathbf{J}_n - \mathbf{J}_n) - \boldsymbol{\Sigma}_{eeAA}^{-1}\boldsymbol{\Sigma}_{ee\bar{A}\bar{A}}\mathbf{J}_{N-n}\right] \\ &= (N-n)(\bar{y}_c - \bar{\mathbf{x}}_c\hat{\boldsymbol{\beta}}) - (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{eeAA}^{-1}\boldsymbol{\Sigma}_{eeAA}\mathbf{J}_{N-n},\end{aligned}$$

where

$$(\bar{y}_c,\ \bar{\mathbf{x}}_c) = (N-n)^{-1}\sum_{i\in A}(\pi_i^{-1} - 1)\ (y_i,\ \mathbf{x}_i).$$

It follows that $\hat{\theta}$ of (2.3.6) is

$$\begin{aligned}\hat{\theta} &= N^{-1}\left(\sum_{i\in A} y_i + (N-n)\bar{y}_c + (N-n)(\bar{\mathbf{x}}_{N-n} - \bar{\mathbf{x}}_c)\hat{\boldsymbol{\beta}}\right) \\ &= \bar{y}_{HT} + N^{-1}(N-n)(\bar{\mathbf{x}}_{N-n} - \bar{\mathbf{x}}_c)\hat{\boldsymbol{\beta}} \\ &= \bar{y}_{HT} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})\hat{\boldsymbol{\beta}}\end{aligned}$$

and equality (2.3.15) is established. By (2.3.8) and (2.3.9),

$$(\hat{\theta} - \bar{y}_N) \mid \mathcal{F}_N = \bar{y}_{HT} - \bar{y}_N + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})\boldsymbol{\beta}_N + O_p(n_N^{-2\alpha}) \quad \text{a.s.}$$

and (2.3.16) follows. Note that $\bar{a}_N = 0$. ■

If (2.3.15) holds, the error in predictor (2.3.6) is

$$\hat{\theta} - \bar{y}_N = (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) + \bar{a}_{HT}$$

and the conditional variance is

$$
\begin{aligned}
V\{\hat{\theta} - \bar{y}_N \mid \mathbf{X}_A, \bar{\mathbf{x}}_N\} &= N^{-2}(\mathbf{C}_\pi', -\mathbf{J}_{N-n}')\boldsymbol{\Sigma}_{eeUU}(\mathbf{C}_\pi', -\mathbf{J}_{N-n}')' \\
&\quad + 2N^{-2}(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})\boldsymbol{\Sigma}_{eeAA}(\mathbf{C}_\pi', -\boldsymbol{\Gamma}_{A\bar{A}}\mathbf{J}_{N-n}) \\
&\quad + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})V\{\hat{\boldsymbol{\beta}} \mid \mathbf{X}_A\}(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})',
\end{aligned}
$$

where $\mathbf{C}_\pi = (\mathbf{D}_\pi^{-1} - \mathbf{I})\mathbf{J}_n$ and a_{iN} is defined in (2.3.16).

A regression model of the form (2.3.4) for which (2.3.10) or (2.3.14) holds is a full model. If (2.3.10) or (2.3.14) does not hold, the model is a reduced model.

We cannot expect condition (2.3.10) or condition (2.3.14) for a full model to hold for every y in a general-purpose survey because $\boldsymbol{\Sigma}_{ee}$ will be different for different y's. Therefore, given a reduced model, one might search for a good model estimator in the class of design-consistent estimators defined by (2.3.3). Given model (2.3.4), we choose the weights w_i to give small model variance for a particular $\boldsymbol{\Sigma}_{ee}$ subject to constraints that guarantee design consistency for any y-characteristic.

To construct a design-consistent estimator of the form $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ when (2.3.4) is a reduced model, we add a vector satisfying (2.3.10) to the X-matrix to create a full model from the original reduced model. There are two possible situations associated with this approach. In the first, the population mean (or total) of the added x-variable is known. With known mean, one can construct the usual regression estimator, and the usual variance estimation formulas are appropriate.

To describe an estimation procedure for the situation in which the population mean of the added variable is not known, we use a transformation similar to that used to obtain (2.2.54). Let z_k denote the added variable, where the vector $\mathbf{z}_k$ satisfies

$$
\mathbf{z}_k = \boldsymbol{\Sigma}_{ee}\mathbf{w}, \tag{2.3.17}
$$

$\mathbf{w} = (w_1, w_2, \ldots, w_n)'$, and

$$
w_i = \left(\sum_{j\in A} \pi_j^{-1}\right)^{-1} \pi_i^{-1}.
$$

Let $\mathbf{Z} = (\mathbf{X}, \mathbf{z}_k)$, where $\mathbf{X}$ is the matrix of auxiliary variables with known population mean $\bar{\mathbf{x}}_N$. For the model

$$
\begin{aligned}
\mathbf{y} &= \mathbf{Z}\boldsymbol{\beta}_{y\cdot z} + \mathbf{e}, \\
\mathbf{e} &\sim (\mathbf{0}, \boldsymbol{\Sigma}_{ee}),
\end{aligned} \tag{2.3.18}
$$

the best linear-in-y, conditional-on-Z unbiased estimator of $\boldsymbol{\beta}_{y\cdot z}$ is

$$\hat{\boldsymbol{\beta}}_{y\cdot z} = (\mathbf{Z}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{y}. \tag{2.3.19}$$

It is not possible to construct an unbiased estimator of $\bar{\mathbf{z}}_N\boldsymbol{\beta}_{y\cdot z}$, conditional on $\mathbf{Z}$, because the $\bar{z}_{k,N}$ of $\bar{\mathbf{z}}_N$ is unknown. It is natural to replace the unknown $\bar{z}_{k,N}$ with an estimator of $\bar{z}_{k,N}$ and we use the regression estimator

$$\bar{z}_{k,reg} = \bar{z}_{k,\pi} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}_{zk\cdot x}, \tag{2.3.20}$$

where

$$\hat{\boldsymbol{\beta}}_{zk\cdot x} = (\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{z}_k. \tag{2.3.21}$$

Then, replacing $\bar{z}_{k,N}$ with $\bar{z}_{k,reg}$ in a regression estimator for the mean of y, we obtain

$$\begin{aligned}
\bar{y}_{reg} &= \bar{y}_\pi + [(\bar{\mathbf{x}}_N, \bar{z}_{k,reg}) - (\bar{\mathbf{x}}_\pi, \bar{z}_{k,\pi})]\hat{\boldsymbol{\beta}}_{y\cdot z} \\
&= \bar{y}_\pi + [(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi), (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}_{zk\cdot x}]\hat{\boldsymbol{\beta}}_{y\cdot z} \\
&= \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}_{y\cdot x},
\end{aligned} \tag{2.3.22}$$

where

$$\hat{\boldsymbol{\beta}}_{y\cdot x} = (\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{y}. \tag{2.3.23}$$

Thus, a regression estimator of the finite population mean based on the full model, but with the mean of z_k unknown and estimated, is the regression estimator with $\boldsymbol{\beta}_{y\cdot x}$ estimated by the generalized least squares regression of y on x using the covariance matrix $\boldsymbol{\Sigma}_{ee}$.

Expression (2.2.50) of Corollary 2.2.3.1 can be used to construct other design-consistent estimators. Letting $\boldsymbol{\Phi}_n$ be diagonal, we obtain a consistent estimator of the form

$$\bar{y}_{reg,\phi} = \bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}_\phi \tag{2.3.24}$$

by setting the ith element of $\boldsymbol{\Phi}_n$ equal to $\phi_i = \pi_i\mathbf{x}_i\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is a fixed vector and, as in (2.2.52),

$$\hat{\boldsymbol{\beta}}_\phi = (\mathbf{X}'\boldsymbol{\Phi}_n^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}_n^{-1}\mathbf{y}. \tag{2.3.25}$$

It is possible that the model specifies the error covariance matrix to be diagonal and the error variances σ_{ei}^2 to be a linear function of $\mathbf{x}_i$. Then setting $\phi_i = \sigma_{ei}^2\pi_i$ gives a design-consistent estimator. See Särndal, Swensson and Wretman (1992, Section 6.5).

Example 2.3.1. Table 2.2 contains an example constructed to compare the model properties of several estimators under the model for the sample observations,

$$\begin{aligned} y_i &= x_{1,i}\beta + e_i, \\ e_i &\sim ind(0,\ x_{1,i}\sigma^2). \end{aligned} \tag{2.3.26}$$

Table 2.2 Weights for Alternative Estimators

Original Weight	$x_{1,i}$	Model-Optimal Weight	Design-Consistent F.M. Weight	Design-Consistent Ratio Wt.	Location-Invariant Regr. Wt.
0.07	0.1	0.10377	0.05636	0.06187	0.15242
0.08	0.3	0.10377	0.06636	0.07071	0.09533
0.08	0.6	0.10377	0.06636	0.07071	0.07856
0.09	1.0	0.10377	0.07636	0.07955	0.08185
0.09	1.7	0.10377	0.07636	0.07955	0.07771
0.10	2.1	0.10377	0.08636	0.08839	0.08658
0.10	3.1	0.10377	0.08636	0.08839	0.08504
0.11	3.5	0.10377	0.09636	0.09722	0.09466
0.13	4.2	0.10377	0.11636	0.11490	0.11419
0.15	5.4	0.10377	0.13636	0.13258	0.13365

and e_i independent of $x_{1,i}$. We assume that the finite population correction can be ignored. The first column of Table 2.2 contains the sampling weights for a mean where the weights are proportional to the inverses of the selection probabilities. The weighted mean of x_1 for the sample is $\bar{x}_{1,\pi} = 2.583$. Assume that $\bar{x}_{1,N} = 2.283$ is known.

Under model (2.3.26) the best linear-in-y conditionally unbiased estimator of β is

$$\hat{\beta}_m = (\mathbf{X}_1'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{y} = \left(\sum_{i\in A} x_{1,i}\right)^{-1}\sum_{i\in A} y_i, \tag{2.3.27}$$

where $\Sigma_{ee} = \sigma^2\ \mathrm{diag}(x_{1,1}, x_{1,2}, \ldots, x_{1,n})$ and $\mathbf{X}_1' = (x_{1,1}, x_{1,2}, \ldots, x_{1,n})$. The associated estimator of the mean of y is

$$\bar{y}_{rat,m} = \bar{x}_{1,N}\hat{\beta}_m = \sum_{i\in A} w_{mi}y_i, \tag{2.3.28}$$

where $w_{mi} = (\Sigma_{i\in A}x_{1,i})^{-1}\bar{x}_{1,N}$. The weights are given in Table 2.2 in the column "model optimal weight." This estimator is the simple ratio estima-

tor and is not design consistent for our unequal probability sample. The conditional model variance of the simple ratio estimator is

$$V\{\bar{y}_{rat,m} \mid \mathbf{X}_1, \bar{x}_{1,N}\} = \mathbf{w}_m'\boldsymbol{\Sigma}_{ee}\mathbf{w}_m = 0.23691\sigma^2,$$

where $\mathbf{w}_m' = (w_{m1}, w_{m2}, \ldots, w_{mn})$.

Construction of a design-consistent estimator of the form $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is as defined in (2.3.6), requires a variable proportional to $x_{1,i}\pi_i^{-1}$ to be in the column space of the matrix of explanatory variables. Because the population mean of $x_{1,i}\pi_i^{-1}$ is not known, we are led to consider estimator (2.3.22),

$$\bar{y}_{reg,fm} = \bar{y}_\pi + (\bar{x}_{1,N} - \bar{x}_{1,\pi})\hat{\beta}_m, \tag{2.3.29}$$

where $\hat{\beta}_m$ is defined in (2.3.27). The weights associated with this estimator are given in the column "design-consistent F.M. weight." The conditional model variance of this estimator is

$$V\{\bar{y}_{reg,fm} \mid \mathbf{X}_1, \bar{x}_{1,N}\} = 0.24860\sigma^2. \tag{2.3.30}$$

Model (2.3.26) is a special model in that the error variance is a linear function of the x-variable. Setting $\phi_i = \pi_i x_i$ in estimator (2.2.52) gives the estimator

$$\bar{y}_{rat,\pi} = \bar{x}_{1,N}(\mathbf{X}_1'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{D}_\pi^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{D}_\pi^{-1}\mathbf{y} = \bar{x}_{1,N}\bar{x}_{1,\pi}^{-1}\bar{y}_\pi. \tag{2.3.31}$$

The weights associated with this estimator are given in the column "design-consistent ratio wt." The conditional model variance of estimator (2.3.31) is

$$V\{\bar{y}_{rat,\pi} \mid \mathbf{X}_1, \bar{x}_{1,N}\} = 0.24604\sigma^2. \tag{2.3.32}$$

For this sample, imposing the design consistency requirement by either of the two methods increases the conditional model variance by less than 5%.

If we are seeking robustness against model failure for ratio model (2.3.26), it is reasonable to consider the alternative model,

$$\begin{aligned} y_i &= \beta_0 + x_{1,i}\beta_1 + e_i, \\ e_i &\sim ind(0, x_{1,i}\sigma^2). \end{aligned} \tag{2.3.33}$$

Under model (2.3.33) with e_i independent of the selection probabilities, the conditional model biases for estimators (2.3.28), (2.3.29), and (2.3.31) are

$$E\{\bar{y}_{rat,m} - \bar{y}_N \mid \mathbf{X}_1, \bar{x}_{1,N}\} = 0.0377\beta_0,$$

$$E\{\bar{y}_{reg,fm} - \bar{y}_N \mid \mathbf{X}_1, \bar{x}_{1,N}\} = -0.1364\beta_0,$$

and

$$E\{\bar{y}_{rat,\pi} - \bar{y}_N \mid \mathbf{X}_1, \bar{x}_{1,N}\} = -0.1161\beta_0,$$

respectively.

The weights in the column "location-invariant regr. wt." are the weights for the estimator

$$\begin{aligned} \bar{y}_{reg,g} &= \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}_{LI} \\ &= \mathbf{w}_g\mathbf{y}, \end{aligned} \tag{2.3.34}$$

where $\mathbf{x}_i = (1,\ x_{1,i})$, $\bar{\mathbf{x}}_N = (1,\ 2.283)$, $\bar{\mathbf{x}}_\pi = (1,\ 2.583)$,

$$\mathbf{w}_g = \mathbf{w} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)(\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1},$$

$$\hat{\boldsymbol{\beta}}_{LI} = (\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{y},$$

$\mathbf{X}' = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')$, and $\mathbf{w}$ is the vector of "original weight" of Table 2.2. The estimator (2.3.34) is conditionally model unbiased under both model (2.3.26) and model (2.3.33), and the conditional model variance is

$$\begin{aligned} V\{\bar{y}_{reg,g} \mid \mathbf{X}, \bar{\mathbf{x}}_N\} &= \mathbf{w}_g'\boldsymbol{\Sigma}_{ee}\mathbf{w}_g \\ &= 0.24646\sigma^2. \end{aligned}$$

The estimator (2.3.34) is location invariant because the column of 1's is the first column of the $\mathbf{X}$-matrix and the original weights sum to 1. That the conditional model variance of the estimator using $\mathbf{x}_i = (1, x_{1,i})$ is less than the conditional model variance of the estimator (2.3.29) using only $x_{1,i}$ may be counterintuitive. However, recall that neither estimator is minimum variance under model (2.3.29) or under model (2.3.33).

The model variances change considerably if the sample mean is less than the population mean. The conditional model variances of the estimators computed under the assumption that the population mean is 2.883 are $0.3778\sigma^2$ for the model optimal estimator, $0.3895\sigma^2$ for design-consistent estimator (2.3.29), $0.3924\sigma^2$ for design-consistent estimator (2.3.31), and $0.3944\sigma^2$ for design-consistent location invariant estimator (2.3.34). Note that the variance of estimator (2.3.29) is smaller than the variance of estimator (2.3.31) when $\bar{x}_{1,N} > \bar{x}_{1,\pi}$ and larger when $\bar{x}_{1,N} < \bar{x}_{1,\pi}$. ■■

There are many estimators that are design consistent and that use the model in construction. Only if the $\mathbf{X}$ matrix satisfies a condition such as (2.3.14) and the population mean of all x-variables is known can we claim model optimality. When the population mean of a variable in the full model is unknown, estimators (2.3.22) and (2.3.24) have appeal but neither is guaranteed to have a smaller conditional model variance than other design-consistent estimators.

The illustration demonstrates that the difference in efficiency between the model variance of an estimator constructed under the reduced model and the model variance of an estimator constructed to be design consistent need not be great. Given a small difference, one might well choose the design-consistent procedure to protect against model failure. If one is considering a *single* y-variable and is *confident* in the reduced model, one might choose the estimator based on the reduced model. It is prudent to test the coefficient of the $z = \sigma_i^2 \pi_i^{-1}$ of the extended model before proceeding with the estimator for the reduced model. See Chapter 6.

A model with no intercept, such as (2.3.26), fails for many common variables. Thus, if the weights are to be used for many y-variables, weights based on model (2.3.33) are preferred. See Exercise 6.

2.3.2 Nonlinear models

Typically, nonlinear models are proposed for estimation of the finite population mean only when the population of x values is known. We consider that case and assume that $(y_i, \mathbf{x}_i), i = 1, 2, \ldots, N$, is an *iid* sample from a superpopulation with finite fourth moments. Let the model for the superpopulation be

$$y_i = \alpha(\mathbf{x}_i, \boldsymbol{\theta}) + e_i, \tag{2.3.35}$$

where $\alpha(\mathbf{x}, \boldsymbol{\theta})$ is a continuous function of the unknown $\boldsymbol{\theta}$, $E\{e_i\} = 0$, and e_i is independent of $\mathbf{x}_j$ for all i and j. Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}$ such that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/2})$ for an *iid* sample. Given the model and $\hat{\boldsymbol{\theta}}$, a natural estimator for the finite population mean is

$$\bar{y}_{m,reg} = N^{-1} \left(\sum_{i \in A} y_i + \sum_{i \in A^c} \alpha(\mathbf{x}_j, \hat{\boldsymbol{\theta}}) \right), \tag{2.3.36}$$

where A^c is the complement of A. Let the selection probabilities for the sample be π_i. If $\alpha(\mathbf{x}, \boldsymbol{\theta})$ has continuous derivatives, the estimator (2.3.36) will be design consistent provided that $\hat{\theta}$ is design consistent,

$$N^{-1} \sum_{i \in A} \pi_i^{-1} [y_i - \alpha(\mathbf{x}_i, \hat{\boldsymbol{\theta}})] = o_p(n^{-1/2}) \tag{2.3.37}$$

and

$$\sum_{i \in A} [y_i - \alpha(\mathbf{x}_i, \hat{\boldsymbol{\theta}})] = o_p(n^{-1/2}). \tag{2.3.38}$$

These conditions are satisfied for simple random sampling and least squares estimation of $\boldsymbol{\theta}$, but need not be satisfied for general sampling schemes. A

design-consistent version of estimator (2.3.36) analogous to estimator (2.2.53) is given in equation (2.3.40) of Theorem 2.3.2. See Firth and Bennett (1998).

Theorem 2.3.2. Let $\{y_i, \mathbf{x}_i\}$ be a sequence of iid random variables with finite fourth moments. Let $\{\mathcal{F}_N\}$ be the first N elements in the sequence $\{y_i, \mathbf{x}_i\}$. Let $\mathbf{x}_i, i = 1, 2, \ldots, N$, be known. Let a sampling design be given with selection probabilities $\pi_{i,N}$ and joint probabilities $\pi_{ik,N}$ such that

$$V\left\{\left(\sum_{i\in A}\pi_{i,N}^{-1}\right)^{-1}\sum_{i\in A}\pi_{i,N}^{-1}z_i - \bar{z}_N \mid \mathcal{F}_N\right\} = O(n^{-1}) \quad \text{a.s.}$$

for any characteristic z with finite second moment.

Assume that an estimator of $\boldsymbol{\theta}$ of (2.3.35), denoted by $\hat{\boldsymbol{\theta}}$, is such that there exists a sequence $\{\boldsymbol{\theta}_N\}$ satisfying

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \mathcal{F}_N = O_p(n^{-\eta}) \quad \text{a.s.} \tag{2.3.39}$$

for some $\eta > 0$, where $\hat{\boldsymbol{\theta}}$ is constructed from $(y_i, \mathbf{x}_i)$, $i \in A$ and $\boldsymbol{\theta}_N$ is an estimator of $\boldsymbol{\theta}$ constructed from $(y_i, \mathbf{x}_i)$, $i \in U_N$. Let $\alpha(\mathbf{x}, \boldsymbol{\theta})$ be a continuous differentiable function of $\boldsymbol{\theta}$ with derivative uniformly continuous in $(\mathbf{x}_j, \boldsymbol{\theta})$, $j \in U_N$, for $\boldsymbol{\theta}$ in a closed set $\mathcal{B}$ containing $\boldsymbol{\theta}_N$ as an interior point.

Let

$$\bar{y}_{c,reg} = \bar{y}_\pi + N^{-1}\sum_{i\in U}\alpha(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) - \hat{N}^{-1}\sum_{i\in A}\pi_{i,N}^{-1}\alpha(\mathbf{x}_i, \hat{\boldsymbol{\theta}}), \tag{2.3.40}$$

where

$$\hat{N} = \sum_{i\in A}\pi_{i,N}^{-1}.$$

Then

$$(\bar{y}_{c,reg} - \bar{y}_N) \mid \mathcal{F} = \bar{a}_\pi - \bar{a}_N + O_p(n^{-0.5-\eta}) \quad \text{a.s.}, \tag{2.3.41}$$

where

$$\bar{a}_\pi = \hat{N}^{-1}\sum_{i\in A}\pi_{i,N}^{-1}a_{i,N}$$

and $a_{i,N} = y_i - \alpha(\mathbf{x}_i, \boldsymbol{\theta}_N)$.

Proof. Expanding $\hat{\boldsymbol{\theta}}$ about $\boldsymbol{\theta}_N$, we have

$$\bar{y}_{c,reg} - \bar{y}_N = \bar{y}_\pi - \bar{y}_N + N^{-1}\sum_{i=1}^{N}\alpha(\mathbf{x}_i, \boldsymbol{\theta}_N)$$

$$
\begin{aligned}
&+ N^{-1}\sum_{i=1}^{N}\mathbf{h}(\mathbf{x}_i,\ \ddot{\boldsymbol{\theta}}_{i,N})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_N)\\
&- \hat{N}^{-1}\sum_{i\in A}\pi_{i,N}^{-1}\alpha(\mathbf{x}_i,\boldsymbol{\theta}_N)\\
&- \hat{N}^{-1}\sum_{i\in A}\pi_{i,N}^{-1}\mathbf{h}(\mathbf{x}_i,\ \ddot{\boldsymbol{\theta}}_{i,N})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_N)\\
= \ &\bar{a}_\pi - \bar{a}_N + (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_N)N^{-1}\sum_{i=1}^{N}\mathbf{h}(\mathbf{x}_i,\ \ddot{\boldsymbol{\theta}}_{i,N})\\
&- (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_N)\hat{N}^{-1}\sum_{i\in A}\pi_{i,N}^{-1}\mathbf{h}(\mathbf{x}_i,\ \ddot{\boldsymbol{\theta}}_{i,N}),
\end{aligned}
$$

where $\ddot{\boldsymbol{\theta}}_{i,N}$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_N$ and $\mathbf{h}(\mathbf{x}_i,\ \ddot{\boldsymbol{\theta}}_{j,N})$ is the vector of first derivatives of $\alpha(\mathbf{x}_i,\ \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at $\ddot{\boldsymbol{\theta}}_{j,N}$. Now $\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_N \to 0$ in probability, and hence the probability that $\ddot{\boldsymbol{\theta}}_{i,N}$ is in $\mathcal{B}$ goes to 1 as $N\to\infty$. Because $\mathbf{h}(x_i,\boldsymbol{\theta})$ is bounded on $\mathcal{B}$,

$$
\left(N^{-1}\sum_{i=1}^{N}\mathbf{h}(\mathbf{x}_i,\ \ddot{\boldsymbol{\theta}}_{i,N}) - \hat{N}^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{h}(\mathbf{x}_i,\ \ddot{\boldsymbol{\theta}}_{i,N})\right)\Bigg|\mathcal{F}_N = O_p(n^{-0.5})
$$

for almost all sequences and we have result (2.3.41). ■

If all x_i are known, a great many estimators of the mean of y are available. For example, given a particular y of interest, nonparametric procedures can be considered. See Breidt and Opsomer (2000). For another method of creating a design-consistent estimator, see Wu and Sitter (2001). A powerful general-purpose estimation procedure for approximating nonlinear relationships is to use the known x values to form poststrata. See Section 2.2.3.

2.4 REGRESSION AND STRATIFICATION

In Section 2.2 we presented the regression estimator for a general design and for a general vector of auxiliary variables. In Theorem 2.2.4 we gave conditions under which a particular regression estimator is the large-sample best with respect to the design variance. Because stratified sampling is so important, we describe the construction of the large-sample best estimator for stratified samples.

By Theorem 2.2.4, to find an estimator with minimum limit variance we choose $\hat{\boldsymbol{\beta}}$ to be the vector $\boldsymbol{\gamma}$ that minimizes the estimated variance

$$
\hat{V}\left\{\bar{y}_{st} - \bar{\mathbf{x}}_{st}\boldsymbol{\gamma}\right\}, \tag{2.4.1}
$$

where $(\bar{y}_{st},\ \bar{\mathbf{x}}_{st})$ is the stratified estimator of the mean of $(y,\ \mathbf{x})$. There are a number of possible estimators for β because there are a number of consistent estimators of the variance. If one uses the unbiased estimator of variance, the resulting estimator for stratified sampling is

$$\bar{y}_{reg} = \bar{y}_{st} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{st})\hat{\boldsymbol{\beta}}_{opt}, \tag{2.4.2}$$

where

$$\hat{\boldsymbol{\beta}}_{opt} = \left(\sum_{h=1}^{H} W_h^2(1-f_h)n_h^{-1}\hat{\mathbf{S}}_{xxh}\right)^{-1} \sum_{h=1}^{H} W_h^2(1-f_h)n_h^{-1}\hat{\mathbf{S}}_{xyh},$$

$$(\hat{\mathbf{S}}_{xxh},\ \hat{\mathbf{S}}_{xyh}) = (n_h-1)^{-1}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\bar{\mathbf{x}}_h)'(\mathbf{x}_{hj}-\bar{\mathbf{x}}_h,\ y_{hj}-\bar{y}_h),$$

$$(\bar{\mathbf{x}}_h,\ \bar{y}_h) = n_h^{-1}\sum_{j=1}^{n_h}(\mathbf{x}_{hj},\ y_{hj}),$$

$$(\bar{\mathbf{x}}_{st}, \bar{y}_{st}) = \sum_{h=1}^{H} W_h(\bar{\mathbf{x}}_h, \bar{y}_h),$$

$f_h = N_h^{-1}n_h$, $W_h = N^{-1}N_h$, N_h is the population stratum size, and n_h is the sample stratum size. The stratified weight for an individual element in the sample estimator of the mean is $W_h n_h^{-1}$. Therefore,

$$\hat{\boldsymbol{\beta}}_{opt} = \mathbf{M}_{xx}^{-1}\sum_{h=1}^{H}\sum_{j=1}^{n_h} K_h(\mathbf{x}_{hj}-\bar{\mathbf{x}}_h)'\ (y_{hj}-\bar{y}_h), \tag{2.4.3}$$

where

$$\mathbf{M}_{xx} = \sum_{h=1}^{H}\sum_{j=1}^{n_h} K_h(\mathbf{x}_{hj}-\bar{\mathbf{x}}_h)'\ (\mathbf{x}_{hj}-\bar{\mathbf{x}}_h),$$

and $K_h = W_h^2(1-f_h)n_h^{-1}(n_h-1)^{-1}$. A simpler form for the regression estimator is obtained if we replace $n_h - 1$ with n_h to obtain

$$K_h^* = W_h^2(1-f_h)n_h^{-2} = N^{-2}N_h^2 n_h^{-2}(1-f_h), \tag{2.4.4}$$

so that the weight for an observation in the regression is proportional to $\pi_{hi}^{-2}(1-f_h)$.

The estimator (2.4.2) is a linear-in-y estimator with weights

$$\omega_{hj} = W_h n_h^{-1} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\mathbf{M}_{xx}^{-1} K_h (\mathbf{x}_{hj} - \bar{\mathbf{x}}_h)',$$

where the weights minimize the Lagrangian

$$\sum_{h=1}^{H}\sum_{j=1}^{n_h} \omega_{hj}^2 K_h^{-1} + \sum_{h=1}^{H} \lambda_h \left(\sum_{j=1}^{n_h} \omega_{hj} - W_h \right) + \sum_{m=H+1}^{H+k} \lambda_i \left(\sum_{h=1}^{H}\sum_{j=1}^{n_h} \omega_{hj} x_{m,hj} - \bar{x}_{m,N} \right). \quad (2.4.5)$$

To compare alternative estimators, we write a regression representation for the jth observation in stratum h as

$$y_{hj} = \mathbf{x}_{hj}\boldsymbol{\beta}_h + e_{hj}, \quad (2.4.6)$$

where

$$e_{hj} \sim ind(0, \sigma_{e,h}^2),$$

and e_{hj} is independent of $\mathbf{x}_{hj}$.

Let the regression estimator be

$$\bar{y}_{st,reg} = \bar{y}_{st} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{st})\hat{\boldsymbol{\beta}}$$

and let $\boldsymbol{\beta}_N$ be the population regression parameter defined as the probability limit of $\hat{\boldsymbol{\beta}}$. Given $\boldsymbol{\beta}_N$, the large-sample variance of the regression estimator is

$$V\{\bar{y}_{st,reg}\} = \sum_{h=1}^{H} W_h^2 (1 - f_h) n_h^{-1} \sigma_{a,h}^2, \quad (2.4.7)$$

where

$$\sigma_{a,h}^2 = \sigma_{e,h}^2 + (\boldsymbol{\beta}_h - \boldsymbol{\beta}_N)' \boldsymbol{\Sigma}_{xx,h} (\boldsymbol{\beta}_h - \boldsymbol{\beta}_N),$$

W_h is the fraction of the population in stratum h, and $\boldsymbol{\Sigma}_{xx,h}$ is the population covariance matrix of $\mathbf{x}_{hj}$ for stratum h. Thus, under model (2.4.6), the presence of different within-stratum variances and (or) different population slopes in different strata leads to different large-sample variances for different $\boldsymbol{\beta}_N$. By construction, $\hat{\boldsymbol{\beta}}_{opt}$ is an estimator of the $\boldsymbol{\beta}_N$ that minimizes (2.4.7).

Example 2.4.1. To illustrate the possible difference between the variances of alternative estimators under model (2.4.6), consider two estimators of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_{wls} = (\mathbf{X}'\mathbf{D}_w\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_w\mathbf{y} \quad (2.4.8)$$

and

$$\hat{\beta}_{opt} = (\mathbf{X}'\mathbf{D}_w^2\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_w^2 y, \tag{2.4.9}$$

where $\mathbf{D}_w$ is a diagonal matrix with diagonal elements equal to W_h for units in stratum h. The probability limit of (2.4.8) is the population ordinary least squares coefficient

$$\boldsymbol{\beta}_{ols,N} = (\mathbf{X}_N'\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{y}_N, \tag{2.4.10}$$

and the probability limit of (2.4.9) is

$$\boldsymbol{\beta}_{opt,N} = (\mathbf{X}_N'\mathbf{D}_{w,N}\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{D}_{w,N}\mathbf{y}_N. \tag{2.4.11}$$

To construct a numerical example, assume a population divided into two strata with $W_1 = 0.15$ and $W_2 = 0.85$. Assume a single x variable with stratum population variances $\sigma_{x,1}^2 = 4.3$ for stratum 1 and $\sigma_{x,2}^2 = 0.6$ for stratum 2. Assume that the stratum coefficients for x are $\beta_1 = 3.0$ for stratum 1 and $\beta_2 = 1.0$ for stratum 2. Then the population least squares coefficient for x is

$$\beta_{ols,N} = \frac{(0.15)(4.3)(3.0) + (0.85)(0.6)(1.0)}{(0.15)(4.3) + (0.85)(0.6)} = 2.1169,$$

and the population coefficient (2.4.11) for x is

$$\beta_{opt,N} = \frac{(0.15)^2(4.3)(3.0) + (0.85)^2(0.6)(1.0)}{(0.15)^2(4.3) + (0.85)^2(0.6)} = 1.3649.$$

To complete the specification for model (2.4.6), assume that $\sigma_{e,1}^2 = 24$ and $\sigma_{e,2}^2 = 0.8$. For the ordinary least squares regression, the stratum variances of the population regression residuals are

$$\sigma_{a,1,ols}^2 = (3 - 2.1169)^2(4.3) + 24 = 27.3537$$

and

$$\sigma_{a,2,ols}^2 = (1 - 2.1169)^2(0.6) + 0.8 = 1.5485.$$

The stratum population variances of the regression residuals computed with $\boldsymbol{\beta}_{opt,N}$ of (2.4.11) are

$$\sigma_{a,1,opt}^2 = (3 - 1.3649)^2(4.3) + 24 = 35.4960$$

and

$$\sigma_{a,2,opt}^2 = (1 - 1.3649)^2(0.6) + 0.8 = 0.8106.$$

To compare the variances of the estimators we assume equal n_h and ignore the finite population correction. Then the large-sample variances of the regression estimators satisfy

$$n_h V\{\bar{y}_{st,reg,wls}\} = 1.7345$$

and

$$n_h V\{\bar{y}_{st,reg,opt}\} = 1.3845.$$

In this illustration, the stratum regression slopes and stratum variances differ considerably, and the optimum procedure is about 25% more efficient than the procedure that uses stratum weights to construct the regression. ■■

Example 2.4.2. Table 2.3 contains an illustrative stratified sample, where x_3 is the auxiliary variable. Let $\mathbf{x}_{hj} = (1, x_{2,hj}, x_{3,hj})$, where $x_{3,hi}$ is given in Table 2.3 and $x_{2,hj}$ is the indicator variable for stratum 1,

$$\begin{aligned} x_{2,hj} &= 1 \quad \text{if } h = 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Table 2.3 Regression Estimators for a Stratified Sample

Stratum	Original Weight	x_3	Stratified Regression Stratified Weight	Stratified Regression Optimum Weight
1	0.03	1.1	0.0034	0.0212
	0.03	2.5	0.0162	0.0254
	0.03	4.1	0.0309	0.0303
	0.03	5.2	0.0410	0.0337
	0.03	7.1	0.0584	0.0394
2	0.17	1.9	0.1128	0.0623
	0.17	2.6	0.1492	0.1309
	0.17	3.0	0.1700	0.1700
	0.17	3.3	0.1856	0.1994
	0.17	4.2	0.2324	0.2874

The column in Table 2.3 with the heading "stratified regression stratified weight" contains weights that can be used to construct the regression estimator with $\hat{\boldsymbol{\beta}}_{wls}$. The weights minimize the Lagrangian

$$\sum_{hj \in A} \omega_{wls,hj}^2 w_{hj}^{-1} + \sum_{r=1}^{3} \lambda_r \left(\sum_{hj \in A} \omega_{wls,hj} x_{r,hj} - \bar{x}_{r,N} \right), \quad (2.4.12)$$

where $(\bar{x}_{1,N}, \bar{x}_{2,N}, \bar{x}_{3,N}) = (1, 0.15, 3.50)$ and $w_{hj} = W_h$ is the original weight of Table 2.3. In matrix notation, the vector of weights is

$$\boldsymbol{\omega}_{wls} = \bar{\mathbf{x}}_N (\mathbf{X}'\mathbf{D}_w\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_w. \quad (2.4.13)$$

The set of weights associated with the coefficient (2.4.9) are given in the column "stratified regression optimum weight." The weights are obtained by minimizing (2.4.12), with w_{hj}^{-2} replacing w_{hj}^{-1}. Because the sample sizes are the same and the finite population correction is ignored, K_h of (2.4.5) is proportional to W_h^2. The vector of weights is

$$\boldsymbol{\omega}_{opt} = \bar{\mathbf{x}}_N(\mathbf{X}'\mathbf{D}_w^2\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_w^2. \tag{2.4.14}$$

We call the weights of (2.4.14) the optimal weights, recognizing that there are other weights that give the same limiting variance. Because the $\mathbf{x}$ vector contains a variable for stratum effect, the weights sum to the known stratum proportion for each stratum. ■■

Minimizing the approximate design variance of the regression estimator minimizes the variance under a model that permits different slopes in different strata. If the stratum means of x are known and if there is a possibility that the regression coefficients differ by stratum, it is natural to construct the regression estimator for each stratum and then merge the stratum estimators using the stratum proportions. The resulting estimator is called the *separate regression estimator*. The regression estimator constructed in Example 2.4.2 with a single coefficient is called the *combined regression estimator*. See Cochran (1977, p. 200).

In Theorem 2.2.3 the design consistency of the regression estimator required the vector $\tilde{\boldsymbol{\beta}}$ to converge to a population parameter $\boldsymbol{\beta}_N$. The practical implication of this requirement for the separate regression estimator is that the sample size should be large in each stratum. If some stratum sample sizes are small, the separate regression estimator can have a large bias. Thus, even if stratum means of the auxiliary variables are available, it may be preferable to use the combined estimator. A compromise estimator can be constructed by combining small strata on the basis of similar characteristics. Also, if there is a vector of auxiliary variables, a regression estimator can be constructed that permits some coefficients to vary by strata while using a single coefficient for other variables.

2.5 ESTIMATION WITH CONDITIONAL PROBABILITIES

Tillé (1998) suggested the computation of an estimator based on the conditional probabilities of selection, given the observed mean for an auxiliary variable. The basis for the estimator is the conditional probability that element i is included in the sample, given that the sample mean of x is equal to t,

$$P\{i \in A \mid \mathcal{F},\ \bar{x}_n = t\} = \frac{P\{i \in A\}\, P\{\bar{x}_n = t \mid \mathcal{F},\ i \in A\}}{P\{\bar{x}_n = t \mid \mathcal{F}\}}. \tag{2.5.1}$$

For some small samples and some finite populations, there may be no sample with sample mean exactly equal to t. Also, if only $\bar{x}_N$ is known, it is impossible to determine the probabilities for $\bar{x}_n$. Therefore, we consider approximations to the probabilities in (2.5.1). If n and N are large and the selection probabilities meet some regularity conditions, $\bar{x}_n$ is approximately normally distributed. Furthermore, the mean of the $n-1$ observations, other than x_i, is also approximately normally distributed. If the sample is a simple random sample and $N^{-1}n$ is small,

$$\bar{x}_n \dot{\sim} N(\bar{x}_N,\ n^{-1}\sigma_x^2) \tag{2.5.2}$$

and

$$\bar{x}_n \mid i \in A \dot{\sim} N(\bar{x}_{N,(i)},\ (n-1)n^{-2}\sigma_x^2), \tag{2.5.3}$$

where

$$\bar{x}_{N,(i)} = E\{\bar{x}_n \mid i \in A\} \doteq n^{-1}\left[(n-1)\bar{x}_N + x_i\right] \tag{2.5.4}$$

and $\dot{\sim}$ is used to denote "is approximately distributed as" and $\doteq$ is used to denote "is approximately equal to."

It follows that the approximate conditional density of the sample mean is

$$f(\bar{x}_n \mid i \in A) = \frac{exp\{-\left[2\sigma_x^2(n-1)n^{-2}\right]^{-1}(\bar{x}_n - \bar{x}_{N,(i)})^2\}}{\left[2\pi(n-1)n^{-2}\sigma_x^2\right]^{1/2}}. \tag{2.5.5}$$

The approximate unconditional density is

$$f(\bar{x}_n) = (2\pi n^{-1}\sigma_x^2)^{-1/2} exp\{-[2\sigma_x^2 n^{-1}]^{-1}(\bar{x}_n - \bar{x}_N)^2\} \tag{2.5.6}$$

and from (2.5.1),

$$\begin{aligned} Nn^{-1}&\left[(n-1)^{-1}n\right]^{-1/2} P\{i \in A \mid \mathcal{F},\ \bar{x}_n\} \\ &= exp\Big\{\left[2\sigma_x^2 n^{-1}\right]^{-1}\left[(\bar{x}_n - \bar{x}_N)^2 - (n-1)^{-1}n(\bar{x}_n - \bar{x}_{N,(i)})^2\right]\Big\} \\ &= exp\Big\{-(2\sigma_x^2)^{-1}n(n-1)^{-1}\left[(\bar{x}_n - \bar{x}_N)^2 + n^{-1}(x_i - \bar{x}_N)^2\right. \\ &\qquad \left.-2(\bar{x}_n - \bar{x}_N)\,(x_i - \bar{x}_N)\right]\Big\}. \end{aligned} \tag{2.5.7}$$

If we denote the probability in (2.5.7) by $\pi_{i|\bar{x}_n}$, a conditional Horvitz–Thompson ratio estimator of the mean is

$$\bar{y}_c = \left(\sum_{i\in A}\pi_{i|\bar{x}_n}^{-1}\right)^{-1}\sum_{i\in A}\pi_{i|\bar{x}_n}^{-1}y_i =: \sum_{i\in A}b_i y_i. \tag{2.5.8}$$

If only $\bar{x}_N$ is known, it is necessary to estimate (2.5.7) by replacing σ_x^2 with s_x^2.

Now $(\bar{x}_n - \bar{x}_N)^2 = O_p(n^{-1})$ and $(n-1)^{-1}(\bar{x}_n - x_i)^2 = O_p(n^{-1})$. Therefore, by a Taylor expansion,

$$\begin{aligned}(N\pi_{i|\bar{x}_n})^{-1} &= n^{-1} + n^{-1}(\bar{x}_N - \bar{x}_n)(\sigma_x^2)^{-1}(x_i - \bar{x}_n) + O_p(n^{-2}) \\ &= n^{-1} + (\bar{x}_N - \bar{x}_n)\left(\sum_{i\in A}(x_i - \bar{x}_n)^2\right)^{-1}(x_i - \bar{x}_n) \\ &\quad + O_p(n^{-2}). \qquad (2.5.9)\end{aligned}$$

Expression (2.5.9) is another demonstration that to the degree that $(\bar{y},\ \bar{x})$ is normally distributed, the regression estimator is an estimator of the conditional expected value of y given that the finite population mean of x is $\bar{x}_N$.

The estimator (2.5.8) is not a regression estimator because $\bar{x}_c$ is not necessarily equal to $\bar{x}_N$. It is natural to create a regression estimator using the b_i of (2.5.8) as initial weights. That regression estimator is

$$\bar{y}_{c,reg} = \sum_{i\in A} w_i y_i = \bar{y}_c + (\bar{x}_N - \bar{x}_c)\hat{\beta}_{c1}, \qquad (2.5.10)$$

where

$$\begin{aligned}(\hat{\beta}_{c0},\ \hat{\beta}_{c1})' &= \left(\sum_{i\in A}\mathbf{z}_i' b_i \mathbf{z}_i\right)^{-1}\sum_{i\in A} b_i \mathbf{z}_i' y_i, \\ w_i &= b_i + (\bar{\mathbf{z}}_N - \bar{\mathbf{z}}_c)\left(\sum_{j\in A} b_j \mathbf{z}_j'\mathbf{z}_j\right)^{-1} b_i \mathbf{z}_i',\end{aligned}$$

$\mathbf{z}_i = (1,\ x_i - \bar{x}_c)$, and $\bar{\mathbf{z}}_N = (0,\ \bar{x}_N - \bar{x}_c)$. Assuming that the existence of moments, by (2.5.9), $(\bar{x}_N - \bar{x}_c) = O_p(n^{-1})$ and estimator (2.5.10) differs from the usual regression estimator by a term that is $O_p(n^{-1})$.

The representation for the conditional inclusion probability extends immediately to a vector of auxiliary variables. Let $\boldsymbol{\Sigma}_{\bar{x}\bar{x}}$ be the covariance matrix of the sample vector of means, let $\bar{\mathbf{x}}_{N,(i)}$ be the expected value of the sample mean given that $\mathbf{x}_i$ is in the sample, let $\boldsymbol{\Sigma}_{\bar{x}\bar{x},\,(i)}$ be the covariance matrix of the sample mean given that $\mathbf{x}_i$ is in the sample, and let π_i be the original selection probability for $\mathbf{x}_i$. Then

$$P\{i \in A \mid \mathcal{F}, \bar{\mathbf{x}}_n\} = \pi_i \mid \boldsymbol{\Sigma}_{\bar{x}\bar{x}} \mid^{1/2} \mid \boldsymbol{\Sigma}_{\bar{x}\bar{x},(i)} \mid^{-1/2} exp\{Q_{xx}\}, \qquad (2.5.11)$$

where

$$\begin{aligned}Q_{xx} &= 0.5[(\bar{\mathbf{x}}_n - \bar{\mathbf{x}}_N)\boldsymbol{\Sigma}_{\bar{x}\bar{x}}^{-1}(\bar{\mathbf{x}}_n - \bar{\mathbf{x}}_N)' \\ &\quad - (\bar{\mathbf{x}}_n - \bar{\mathbf{x}}_{N,(i)})\boldsymbol{\Sigma}_{\bar{x}\bar{x},(i)}^{-1}(\bar{\mathbf{x}}_n - \bar{\mathbf{x}}_{N,(i)})'\,].\end{aligned}$$

In simple random sampling $\bar{\mathbf{x}}_{N,(i)}$ is given by the vector version of (2.5.4) and $\boldsymbol{\Sigma}_{xx,(i)}$ is $(n-1)n^{-2}\boldsymbol{\Sigma}_{xx}$.

To develop an approximation for general designs, let i be the index for primary sampling units. Let $\bar{\mathbf{x}}_\pi$ be a linear design unbiased, or nearly unbiased, estimator of the population mean. Consider an estimator of the population mean constructed by removing primary sampling unit i and increasing the weights of some of the remaining elements so that the resulting estimator is unbiased, or nearly unbiased, for the mean of the population with x_i deleted. For example, let the sample be a simple stratified sample with estimated mean

$$\bar{x}_\pi = \sum_{h=1}^{H} W_h \sum_{j \in A_h} n_h^{-1} x_j,$$

where A_h is the set of sample elements in stratum h, W_h is the fraction of the population in stratum h, and H is the total number of strata. Then the estimator of the population mean of x with element i in stratum m removed from the sample is

$$\bar{x}_\pi^{(i)} = \sum_{\substack{h=1 \\ h \neq m}}^{H} W_h \sum_{j \in A_h} n_h^{-1} x_j + W_m \sum_{\substack{j \in A_m \\ j \neq i}} (n_h - 1)^{-1} x_j. \quad (2.5.12)$$

See also jackknife variance estimation in Section 4.2.

By construction, the conditional expectation of $\bar{x}_\pi^{(i)}$ given that element i is in the sample is

$$E\{\bar{x}_\pi^{(i)} \mid i \in A, \mathcal{F}\} = (N-1)^{-1} \left(\sum_{j \in U} x_j - x_i \right) \doteq \bar{x}_N \quad (2.5.13)$$

for large N. Therefore,

$$E\{\bar{x}_\pi^{(i)} - \bar{x}_\pi \mid i \in A, \mathcal{F}\} \doteq \bar{x}_N - E\{\bar{x}_\pi \mid i \in A, \mathcal{F}\} \quad (2.5.14)$$

and an estimator of the conditional mean of $\bar{x}_\pi$ given that $x_i \in A$ is

$$\hat{\mu}_{\bar{x}|i \in A} = \bar{x}_N + \bar{x}_\pi - \bar{x}_\pi^{(i)}. \quad (2.5.15)$$

We assume that the estimators of the mean are such that

$$C\{\bar{x}_\pi^{(i)} - \bar{x}_\pi, \bar{x}_\pi \mid \mathcal{F}\} = 0. \quad (2.5.16)$$

This condition holds, for example, for the estimator (2.5.12). Then

$$V\{\bar{x}_\pi^{(i)} \mid i \in A, \mathcal{F}\} = V\{\bar{x}_\pi \mid i \in A, \mathcal{F}\} + V\{\bar{x}_\pi^{(i)} - \bar{x}_\pi \mid i \in A, \mathcal{F}\}. \quad (2.5.17)$$

An estimator of expression (2.5.11) can be constructed if it is possible to estimate $V\{\bar{x}_\pi^{(i)} \mid \mathcal{F}\}$. In the case of a stratified sample with element i in stratum m removed,

$$V\{\bar{x}_\pi^{(i)} \mid i \in A,\ \mathcal{F}\} = \sum_{\substack{h=1 \\ h \neq m}}^{H} W_h^2 n_h^{-1} S_h^2 + W_m^2 (n_m - 1)^{-1} S_{m(i)}^2$$

and

$$\hat{V}\{\bar{x}_\pi^{(i)} \mid i \in A,\ \mathcal{F}\} = \sum_{\substack{h=1 \\ h \neq m}}^{H} W_h^2 n_h^{-1} s_h^2 + W_m^2 (n_m - 1)^{-1} s_{m(i)}^2,$$

where $S_{m(i)}^2$ is the population variance of stratum m when element i is removed and $s_{m(i)}^2$ is the corresponding sample quantity. The stratified variance expressions are discussed in Section 1.2.2.

2.6 REGRESSION FOR TWO-STAGE SAMPLES

In the preceding sections we presented the regression estimator as if the sampling unit and the analysis unit were the same. In cluster or two-stage samples, this is no longer true. Consider a sample containing n primary sampling units with m_i second-stage units in the ith primary sampling unit. Let a vector of auxiliary variables with known population means be available. If, following Theorem 2.2.3, one constructs a regression estimator of the total by minimizing the estimated design variance, the estimator is

$$\hat{T}_{y,reg} = \hat{T}_y + (\mathbf{T}_{x1,N} - \hat{\mathbf{T}}_{x1})\hat{\boldsymbol{\beta}}_{do}, \tag{2.6.1}$$

where

$$\hat{\boldsymbol{\beta}}_{do} = [\hat{V}\{\hat{\mathbf{T}}_{x1} \mid \mathcal{F}\}]^{-1}\hat{C}\{\hat{\mathbf{T}}_{x1},\ \hat{T}_y \mid \mathcal{F}\}, \tag{2.6.2}$$

$$(\hat{\mathbf{T}}_{x1},\ \hat{T}_y) = \sum_{ij \in A} \pi_{(ij)}^{-1}(\mathbf{x}_{1,ij},\ y_{ij}),$$

and $\pi_{(ij)}$ is the probability of selecting secondary unit j in primary sampling unit i. If the covariance matrix $\hat{V}\{\hat{\mathbf{T}}_{x1} \mid \mathcal{F}\}$ is a quadratic function and the finite population corrections are ignored, estimator (2.6.1) assigns a weight to each primary sampling unit. If the number of primary units in the sample is large and the number of elements per primary unit relatively small, a single weight for each primary sampling unit is feasible. A good example

is a sample of households. Census information may be available for both household characteristics and personal characteristics. Because the number of persons per household is relatively small, the regression estimator can be constructed by assigning a weight to each household. If all members of each sampled household are observed, each member receives the household weight in tabulations of individuals. In many practical situations the number of primary sampling units is small and hence the degrees of freedom for the estimator $\hat{V}\{\hat{\mathbf{T}}_{x1}\}$ is small. It is also possible for the dimension of $\mathbf{x}_1$ to be relatively large. In such cases the $\hat{\boldsymbol{\beta}}$ of (2.6.2) may have a very large variance, or in extreme cases, $\hat{V}\{\hat{\mathbf{T}}_{x1}\}$ may be singular.

If the primary sampling units are relatively few in number, if little analysis of the primary units themselves is planned, and if the principal analysis unit is the element, it is reasonable to construct weights for elements. The most common procedure is to construct weights for secondary units ignoring the primary sampling units. Then a regression estimator for the total is

$$\hat{T}_{y,reg,su} = \sum_{ij \in A} \pi_{(ij)}^{-1} y_{ij} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_x)\hat{\boldsymbol{\beta}}_{su}, \tag{2.6.3}$$

where

$$\hat{\boldsymbol{\beta}}_{su} = \left(\sum_{ij \in A} \pi_{(ij)}^{-1} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_\pi)'(\mathbf{x}_{ij} - \bar{\mathbf{x}}_\pi) \right)^{-1} \sum_{ij \in A} \pi_{(ij)}^{-1} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_\pi)'(y_{ij} - \bar{y}_\pi) \tag{2.6.4}$$

and

$$(\bar{\mathbf{x}}_\pi, \bar{y}_\pi) = \left(\sum_{ij \in A} \pi_{(ij)}^{-1} \right)^{-1} \sum_{ij \in A} \pi_{(ij)}^{-1} (\mathbf{x}_{ij}, y_{ij}).$$

Under our usual conditions for consistency, the variance can be estimated by applying the two-stage variance estimator to $\hat{e}_{ij} = y_{ij} - \bar{y}_\pi - (\mathbf{x}_{ij} - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}_{su}$. See (1.2.82) and (1.2.87).

In constructing an element regression estimator for a two-stage sample there may be auxiliary information available for the primary sampling units as well as for elements. Also, it may be desirable to maintain control for the number of primary sampling units per stratum. One can include controls for primary sampling unit characteristics by defining the variable

$$z_{ij} = z_i m_i^{-1} \pi_{j|i}, \tag{2.6.5}$$

where z_i is the characteristic total for the ith primary sampling unit, m_i is the number of sample elements in the ith primary sampling unit, and $\pi_{j|i}$ is the probability of selecting element ij given that primary sampling unit i has

been selected. Observe that $\Sigma_j \pi_{j|i}^{-1} z_{ij} = z_i$, where the sum is over sample elements in the ith primary sampling unit. If the subsampling rate within a primary unit is constant, $z_{ij} = M_i^{-1} z_i$, where M_i is the population number of elements in the ith primary sampling unit. The restriction on the regression weights expressed in terms of element weights is

$$\sum_{ij \in A} w_{ij} z_{ij} = \bar{z}_N,$$

where w_{ij} is the final weight for element j in primary sampling unit i and $\bar{z}_N$ is the population mean of z_i.

In cases with a small number of primary sampling units, a design-consistent model-based estimator of β can be considered. A potential model for the observations y_{ij} in a two-stage, or cluster, sample is

$$\begin{aligned} y_{ij} &= \mathbf{x}_{ij}\boldsymbol{\beta} + u_{ij}, \\ u_{ij} &= b_i + e_{ij}, \end{aligned} \tag{2.6.6}$$

where $b_i \sim iid(0, \sigma_b^2)$, $e_{ij} \sim iid(0, \sigma_e^2)$, and e_{ij} is independent of b_k for all i, j, and k. With model (2.6.6) the error covariance matrix is block diagonal with n blocks, where the ith block is an $m_i \times m_i$ matrix

$$\boldsymbol{\Sigma}_{uuii} = \mathbf{I}\sigma_e^2 + \mathbf{J}\mathbf{J}'\sigma_b^2, \tag{2.6.7}$$

and $\mathbf{J}$ is an m_i-dimensional column vector of 1's.

Example 2.6.1. Table 2.4 contains an illustration data set for a stratified two-stage sample. The first column of Table 2.3 gives the stratum. The primary unit sampling rate is one in 40 for stratum 1 and one in 30 for stratum 2. Note that there are 200 primary sampling units in stratum 1 and 90 primary sampling units in stratum 2. The element weights are the stratum weights divided by the subsampling rates and are given in the column "w_0." An element control variable, identified as "x_1," is given in the table. The auxiliary variables used to maintain the stratum primary sampling unit totals, as defined in (2.6.5), are

$$(z_{1,ij}, z_{2,ij}) = m_i^{-1} \pi_{j|i} (\delta_{1,ij}, \delta_{2,ij}) = M_i^{-1} (\delta_{1,ij}, \delta_{2,ij}),$$

where $\delta_{h,ij}$ is the stratum indicator for stratum h. These variables are given in the two columns identified as "z_1" and "z_2."

To define regression weights, we assume that model (2.6.6) holds for both strata of the stratified sample and write the model as

$$\begin{aligned} y_{hij} &= \mathbf{x}_{hij}\boldsymbol{\beta} + u_{hij}, \\ u_{hij} &= b_{hi} + e_{hij}, \end{aligned} \tag{2.6.8}$$

Table 2.4 Stratified Two-Stage Sample

Str.	PSU	SSU	PSU Wt.	w_0	x_1	z_1	z_2	z_3	y
1	1	1	40	80	3.0	0.1667	0.0000	176	10.64
	1	2	40	80	3.3	0.1667	0.0000	176	11.40
	1	3	40	80	2.5	0.1666	0.0000	176	8.36
	2	1	40	80	5.3	0.5000	0.0000	112	19.75
	3	1	40	160	4.1	0.0625	0.0000	416	19.96
	3	2	40	160	5.0	0.0625	0.0000	416	19.92
	3	3	40	160	4.4	0.0625	0.0000	416	18.63
	3	4	40	160	3.7	0.0625	0.0000	416	16.31
	4	1	40	160	6.0	0.0500	0.0000	480	21.93
	4	2	40	160	5.1	0.0500	0.0000	480	19.78
	4	3	40	160	4.4	0.0500	0.0000	480	19.69
	4	4	40	160	3.3	0.0500	0.0000	480	15.99
	4	5	40	160	4.0	0.0500	0.0000	480	16.47
	5	1	40	80	6.4	0.2500	0.0000	144	23.27
	5	2	40	80	4.9	0.2500	0.0000	144	18.48
2	6	1	30	150	4.2	0.0000	0.0286	570	20.82
	6	2	30	150	4.5	0.0000	0.0286	570	20.17
	6	3	30	150	6.4	0.0000	0.0285	570	26.12
	6	4	30	150	3.6	0.0000	0.0286	570	18.78
	6	5	30	150	4.3	0.0000	0.0286	570	20.36
	6	6	30	150	3.3	0.0000	0.0286	570	19.37
	6	7	30	150	4.8	0.0000	0.0285	570	21.47
	7	1	30	60	5.1	0.0000	0.1667	132	21.64
	7	2	30	60	4.8	0.0000	0.1666	132	21.34
	7	3	30	60	4.9	0.0000	0.1667	132	22.73
	8	1	30	120	3.7	0.0000	0.0500	360	17.25
	8	2	30	120	3.8	0.0000	0.0500	360	19.41
	8	3	30	120	2.9	0.0000	0.0500	360	16.13
	8	4	30	120	5.4	0.0000	0.0500	360	23.76
	8	5	30	120	5.3	0.0000	0.0500	360	21.97

where b_{hi} is the primary sampling unit (PSU) effect and e_{hij} is the element effect. Assume that the b_{hi} are independent $(0, \sigma_b^2)$ random variables, the e_{hij} are independent $(0, \sigma_e^2)$ random variables, and b_{ht} is independent of e_{hij} for all h, i, j, and t. The covariance matrix of $\mathbf{u} = (u_{1,1,1}, u_{1,1,2}, \ldots, u_{2,8,5})'$, denoted by $\boldsymbol{\Sigma}_{uu}$, is block diagonal with blocks of dimension m_i where the ith

block is given by (2.6.7). Note that

$$\boldsymbol{\Sigma}_{uuii}^{-1} = \sigma_e^{-2}\mathbf{I} - \sigma_e^{-2}\sigma_b^2(\sigma_e^2 + m_i\sigma_b^2)^{-1}\mathbf{J}\mathbf{J}'. \tag{2.6.9}$$

To compute the regression estimator, only the ratio $\sigma_e^{-2}\sigma_b^2$ is required and we compute the regression weights under the assumption that $\sigma_e^{-2}\sigma_b^2 = 0.4$. The population total of x_1 is known to be 18,168.7 and the sample estimate of the total of x_1 is 16,517.0.

The regression estimator $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ for model (2.6.6) will be design consistent if the $\mathbf{X}$-matrix is such that

$$\boldsymbol{\Sigma}_{uu}^{-1}\mathbf{X}\boldsymbol{\gamma} = \mathbf{L}_w, \tag{2.6.10}$$

where $\boldsymbol{\gamma}$ is a fixed vector and $\mathbf{L}_w = (\pi_{1,1,1}^{-1}, \pi_{1,1,2}^{-1}, \ldots, \pi_{2,8,5}^{-1})'$ is the vector of sampling weights. The vector $\sigma_e^{-2}\boldsymbol{\Sigma}_{uu}\mathbf{L}_w$ is identified as "z_3" in Table 2.4. If the number of secondary units and the subsampling rate for each primary unit is known, the population total of z_3 is known and is

$$T_{z3} = \sum_{i\in U} M_i^2 m_i^{-1}(1 + \sigma_e^{-2}\sigma_b^2 m_i)\pi_i^{-1}. \tag{2.6.11}$$

For our illustration, the total of z_3 is 1,687,000 and the sample estimate of the total of z_3 is 1,562,740. Then the vector of weights for the model-optimal regression estimator $\mathbf{T}_\mathbf{x}\hat{\boldsymbol{\beta}}$, computed with $\mathbf{x}_{hij} = (x_{1,hij}, z_{1,hij}, z_{2,hij}, z_{3,hij})$, is

$$\mathbf{w}_f' = \mathbf{T}_\mathbf{x}(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}, \tag{2.6.12}$$

where the rows of $\mathbf{X}$ are $\mathbf{x}_{hij}$, the total of $\mathbf{x}$ is

$$\mathbf{T}_\mathbf{x} = (18.1687,\ 0.200,\ 0.090,\ 1687.000) \times 10^3,$$

and $\boldsymbol{\Phi} = \sigma_e^{-2}\boldsymbol{\Sigma}_{uu}$ with $\sigma_e^{-2}\sigma_b^2 = 0.4$. We use the weights of (2.6.12), given in Table 2.5 and the vector in the last column of Table 2.4, denoted by $\mathbf{y}$, to illustrate some of the computations for two-stage regression. The estimated total of y using weights (2.6.12) is

$$\hat{T}_{y,reg,f} = \mathbf{w}_f'\mathbf{y} = 78,432.$$

An estimated design variance, ignoring the finite population correction, is

$$\begin{aligned}\hat{V}\{\hat{T}_{y,reg,f} \mid \mathcal{F}\} &= c_{df}\sum_{h=1}^{2} n_h(n_h - 1)^{-1}\sum_{i\in A_{1h}}\pi_{hi}^{-2}(\hat{u}_{f,hi} - \bar{u}_{f,h})^2 \\ &= 2,276,381,\end{aligned} \tag{2.6.13}$$

Table 2.5 Regression Weights for Stratified Two-Stage Sample

Str.	PSU	SSU	w_0	w_f	w_p
1	1	1	80	80.4	77.3
	1	2	80	89.3	78.5
	1	3	80	65.6	75.3
	2	1	80	37.3	57.2
	3	1	160	172.5	181.5
	3	2	160	199.1	188.6
	3	3	160	181.3	183.9
	3	4	160	160.6	178.4
	4	1	160	220.0	198.7
	4	2	160	193.4	191.6
	4	3	160	172.7	186.1
	4	4	160	140.1	177.4
	4	5	160	160.8	182.9
	5	1	80	128.5	83.5
	5	2	80	84.2	77.5
2	6	1	150	151.2	169.4
	6	2	150	160.1	171.6
	6	3	150	216.3	185.7
	6	4	150	133.4	165.0
	6	5	150	154.1	170.2
	6	6	150	124.6	162.7
	6	7	150	168.9	173.9
	7	1	60	56.0	47.7
	7	2	60	47.1	46.8
	7	3	60	50.1	47.1
	8	1	120	115.8	125.5
	8	2	120	118.7	126.1
	8	3	120	92.1	120.8
	8	4	120	166.1	135.6
	8	5	120	163.1	135.0

where $c_{df} = (n_{su} - 4)^{-1}(n_{su} - 2)$, $\hat{u}_{f,hij} = y_{hij} - \mathbf{x}_{hij}\hat{\boldsymbol{\beta}}_f$,

$$\begin{aligned}
\hat{u}_{f,hi} &= \pi_{hi} \sum_{j \in B_{hi}} w_{f,hij} \hat{u}_{f,hij}, \\
\bar{u}_{f,h} &= n_h^{-1} \sum_{i \in A_{1h}} \hat{u}_{f,hi}, \\
\hat{\boldsymbol{\beta}}_f &= (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Phi}^{-1}\mathbf{y},
\end{aligned}$$

$n_{su} = 30$ is the number of secondary units in the regression, n_h is the number of primary sampling units in stratum h, B_{hi} is the set of indices for elements in primary sampling unit hi, $\pi_{hi} = \pi_h$ is the first-stage sampling rate, and A_{1h} is the set of indices for primary sampling units in stratum h. The ratio $(n_{su} - 4)^{-1}(n_{su} - 2)$ is the ratio of sample size less the number of strata to sample size less the dimension of $\mathbf{x}_{hij}$. The quantity $\pi_{hi} w_{f,hij}$ is the regression modification of the inverse of the conditional probability that element hij is in the sample given that primary sampling unit hi is in the sample. One could use w_o to form $\hat{u}_{f,hi}$ for an alternative consistent estimator.

The estimated variance of $\hat{T}_{y,reg,f}$ under the model is

$$\begin{aligned} \hat{V}\{\hat{T}_{y,reg,f} \mid \mathbf{X}, \bar{\mathbf{x}}_N\} &= \mathbf{w}_f' \mathbf{\Phi} \mathbf{w}_f \hat{\sigma}_e^2 \\ &= 1,843,943 \hat{\sigma}_e^2 \\ &= 1,915,857, \end{aligned} \tag{2.6.14}$$

where

$$\begin{aligned} \hat{\sigma}_e^2 &= (n_{su} - 4)^{-1} \left[\mathbf{y}'\mathbf{\Phi}^{-1}\mathbf{y} - \mathbf{y}'\mathbf{\Phi}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{y}\right] \\ &= 1.039. \end{aligned}$$

This sample is too small for reliable comparisons, but if the model is true and $\sigma_e^{-2}\sigma_b^2 = 0.4$, the estimators (2.6.13) and (2.6.14) are estimating the same quantity. The design estimator (2.6.13) remains an appropriate estimator if the model is not true.

The generalized least squares estimate of $\boldsymbol{\beta}$ for the regression of y on $(x_1,\ z_1,\ z_2,\ 0.01z_3)$ defined in (2.6.13) is

$$\hat{\boldsymbol{\beta}}_f = (2.76,\ 7.39,\ 40.89,\ 1.37)'.$$

The estimated model covariance matrix is

$$\begin{aligned} \hat{V}\{\hat{\boldsymbol{\beta}}_f \mid \mathbf{X}\} &= (\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{X})^{-1}\hat{\sigma}_e^2 \\ &= \begin{pmatrix} 0.0449 & -0.5105 & -1.0376 & -0.0350 \\ -0.5105 & 9.6950 & 12.9631 & 0.3408 \\ -1.0376 & 12.9631 & 51.4084 & 0.6347 \\ -0.0350 & 0.3408 & 0.6347 & 0.0356 \end{pmatrix}, \end{aligned} \tag{2.6.15}$$

where $\hat{\sigma}_e^2$ is given in (2.6.14). The t-statistic to test the hypothesis that the coefficient of z_3 is zero is 7.26, and the reduced model containing only (x_1, z_1, z_2) is rejected. Thus, we would expect the model estimator $\bar{\mathbf{z}}_N \hat{\boldsymbol{\beta}}_{y \cdot z_1}$ based on $\mathbf{z} = (x_1, z_1, z_2)$ to be seriously biased. Also see Section 6.3.

The design covariance matrix of $\hat{\boldsymbol{\beta}}_f$ is

$$\begin{aligned} V\{\hat{\boldsymbol{\beta}}_f \mid \mathcal{F}\} &= V\{(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{u} \mid \mathcal{F}\} \\ &= V\{\mathbf{G}_{f\beta}\mathbf{u} \mid \mathcal{F}\}, \end{aligned}$$

where $\mathbf{G}_{f\beta} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}$ and $\mathbf{u}$ is the vector with components u_{hij}. An estimator of the covariance matrix is

$$\hat{V}\{\hat{\boldsymbol{\beta}}_f \mid \mathcal{F}\} = c_{df} \sum_{h=1}^{2} n_h(n_h-1)^{-1} \sum_{i \in A_{1h}} (\hat{\boldsymbol{\zeta}}_{f\beta,hi} - \bar{\boldsymbol{\zeta}}_{f\beta,h})(\hat{\boldsymbol{\zeta}}_{f\beta,hi} - \bar{\boldsymbol{\zeta}}_{f\beta,h})', \tag{2.6.16}$$

where $c_{df} = (n-H)(n-H-2)^{-1}$, $H=2$ is the number of strata, there are two explanatory variables other than stratum indicators,

$$\hat{\boldsymbol{\zeta}}_{f\beta,hi} = \sum_{j \in B_{hi}} \mathbf{g}_{f\beta,hij}\hat{u}_{f,hij},$$

$$\bar{\boldsymbol{\zeta}}_{f\beta,h} = n_h^{-1} \sum_{i \in A_{1h}} \hat{\boldsymbol{\zeta}}_{f\beta,hi},$$

$\mathbf{g}_{f\beta,hij}$ is the hij column of $\mathbf{G}_{f\beta}$, and $\hat{u}_{f,hij}$ is defined in (2.6.13). The estimate is

$$\hat{V}\{\hat{\boldsymbol{\beta}}_f \mid \mathcal{F}\} = \begin{pmatrix} 0.0581 & -0.5334 & -1.4133 & -0.0337 \\ -0.5334 & 9.1527 & 12.5783 & 0.2748 \\ -1.4133 & 12.5783 & 44.6731 & 0.8074 \\ -0.0337 & 0.2748 & 0.8074 & 0.0238 \end{pmatrix}. \tag{2.6.17}$$

The regression vector and the estimated covariance matrix can be computed in SAS or STATA. The test statistics for the hypothesis that the coefficient of z_3 is zero based on the variance from (2.6.17) is 8.88, giving the same conclusion as the use of the variance from (2.6.15). The two covariance matrices (2.6.15) and (2.6.17) differ by a considerable amount because the model may not be true and because this is a very small sample.

If the mean of z_3 is not known and if z_3 is significant in the regression, the estimator (2.6.3) is a strong option for estimating T_y. The vector of regression estimation weights for the probability weighted estimator (2.6.3) is

$$\mathbf{w}_p' = \mathbf{w}_0' + (\mathbf{T}_z - \hat{\mathbf{T}}_z)(\mathbf{Z}'\mathbf{D}_\pi^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_\pi^{-1},$$

where $\mathbf{D}_\pi^{-1} = \text{diag}(\mathbf{w}_0)$ and the rows of $\mathbf{Z}$ are $\mathbf{z}_{hij} = (x_{1,hij},\ z_{1,hij},\ z_{2,hij})$. The weights $w_{p,hij}$ are given in the last column of Table 2.5. The estimated total using the $w_{p,hij}$ is

$$\hat{T}_{y,reg,p} = \mathbf{w}_p'\mathbf{y} = 79,368.$$

The estimated variance of the form (2.6.13) using $w_{p,hij}$ and $\hat{u}_{p,hij} = y_{hij} - \mathbf{z}_{1,hij}\hat{\boldsymbol{\beta}}_p$ is

$$\hat{V}\{\hat{T}_{y,reg,p} \mid \mathcal{F}\} = 3,163,440, \tag{2.6.18}$$

where

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{Z}'\mathbf{D}_\pi^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_\pi^{-1}\mathbf{y}.$$

The estimated vector of regression coefficients is

$$\hat{\boldsymbol{\beta}}_p' = \underset{(0.16)}{(4.36,} \quad \underset{(4.10)}{-10.58,} \quad \underset{(16.01)}{13.71),}$$

where the standard errors are the square roots of the estimated covariance matrix of the form (2.6.16). The primary reason that $\hat{V}\{\hat{T}_{y,reg,p} \mid \mathcal{F}\}$ is larger than $\hat{V}\{\hat{T}_{y,reg,f} \mid \mathcal{F}\}$ is that extra information, the z_3, is used for $\hat{T}_{y,reg,f}$. The estimator of the form (2.6.13) calculated with $w_{p,hij}$ and $\hat{u}_{f,hij}$ in place of $\hat{u}_{p,hij}$ gives 2,119,104, illustrating the effect of z_3 on the variance.

Table 2.6 Primary Sampling Unit Totals and Regression Weights

Str.	PSU	PSU Weight	PSU x_1 Total	z_1	z_2	$\hat{y}$	Regr. Design Weight	Residuals
1	1	40	17.6	1	0	60.80	36.13	1.2096
	2	40	10.6	1	0	39.50	35.02	11.8637
	3	40	68.8	1	0	299.28	44.20	5.9675
	4	40	91.2	1	0	375.44	47.73	–20.1259
	5	40	22.6	1	0	83.50	36.92	1.0851
2	6	30	155.5	0	1	735.45	36.99	15.3726
	7	30	29.6	0	1	131.42	23.59	–13.9385
	8	30	84.4	0	1	394.08	29.42	–1.4341

Table 2.6 contains the estimated PSU totals of x_1 and of y required to construct estimator (2.6.1) using x_1 as the auxiliary variable. The regression weights for the PSUs constructed by minimizing the estimated variance of the estimator are given in Table 2.6. The finite population correction was ignored, giving the vector of weights

$$\mathbf{w}_d = (T_{x1}, T_{z1}, T_{z2})(\tilde{\mathbf{Z}}_1'\mathbf{K}\tilde{\mathbf{Z}}_1)^{-1}\tilde{\mathbf{Z}}_1'\mathbf{K},$$

where the rows of $\tilde{\mathbf{Z}}_1$ are $\tilde{\mathbf{z}}_{1,hi} = (\hat{x}_{1,hi}, z_{1,hi}, z_{2,hi})$, the elements of $\hat{\mathbf{y}}$ are $\hat{y}_{hi}$, $\mathbf{K}$ is a diagonal matrix with $N_1^2 n_1^{-1}(n_1 - 1)^{-1} = 2000$ for PSUs in the first stratum and $N_2^2 n_2^{-1}(n_2 - 1)^{-1} = 1350$ for PSUs in the second stratum,

N_h is the population number of PSUs in stratum h, and n_h is the sample number of PSUs in stratum h.

The vector of estimated coefficients is

$$\hat{\boldsymbol{\beta}}_d' = (\hat{\beta}_{x1},\ \hat{\beta}_{z1},\ \hat{\beta}_{z2}) = \begin{matrix}(4.56, & -20.75, & 10.24),\\ (0.17) & (6.80) & (17.03)\end{matrix}$$

where

$$\hat{\boldsymbol{\beta}}_d = (\tilde{\mathbf{Z}}_1'\mathbf{K}\mathbf{Z}_1)^{-1}\tilde{\mathbf{Z}}_1'\mathbf{K}\hat{\mathbf{y}},$$

and the elements of $\hat{\mathbf{y}}$ are $\hat{y}_{hij}$. The standard errors are the square roots of the diagonal elements of

$$\hat{V}\{\hat{\boldsymbol{\beta}}_d \mid \mathcal{F}\} = (\tilde{\mathbf{Z}}_1'\mathbf{K}\tilde{\mathbf{Z}}_1)^{-1}\tilde{\mathbf{Z}}_1'\mathbf{K}\mathbf{D}_{cd}\hat{\mathbf{D}}_{uu}\mathbf{K}\tilde{\mathbf{Z}}_1(\tilde{\mathbf{Z}}_1'\mathbf{K}\tilde{\mathbf{Z}}_1)^{-1}, \qquad (2.6.19)$$

where $\hat{\mathbf{D}}_{uu} = \text{diag}(\hat{u}_{d,hi}^2)$,

$$\begin{aligned}\hat{u}_{d,hi} &= \hat{y}_{hi} - \bar{y}_{hi} - (\hat{x}_{hi} - \bar{x}_h)\hat{\beta}_{d,x1}\\ &= \hat{y}_{hi} - \tilde{\mathbf{z}}_{1,hi}\hat{\boldsymbol{\beta}}_d,\end{aligned}$$

$$\hat{\beta}_{d,x1} = [\hat{V}\{\hat{T}_x \mid \mathcal{F}\}]^{-1}\hat{\mathbf{C}}\{\hat{\mathbf{T}}_x, \hat{T}_y \mid \mathcal{F}\},$$

$$\mathbf{D}_{cd} = (n-H)(n-H-1)^{-1} \times \text{blockdiag}(1.2\mathbf{I}_5, 1.5\mathbf{I}_3),$$

$n = 8$ is the number of primary sampling units, $H = 2$ is the number of strata, and the multipliers for the identity matrices are $n_h(n_h - 1)^{-1}$ for $h = 1, 2$. Because the stratum means of the $\hat{u}_{d,hi}$ are zero, the estimated variance (2.6.19) is the estimated variance for a stratified sample.

The estimated total is

$$\hat{T}_{y,reg,d} = (T_{x1}, T_{z1}, T_{z2})\hat{\boldsymbol{\beta}}_d = \sum_{h=1}^{2}\sum_{j=1}^{n_h} w_{d,hj}y_{hj} = 79,709$$

with estimated variance

$$\begin{aligned}\hat{V}\{\hat{T}_{y,reg,d} \mid \mathcal{F}\} &= C_H \sum_{h=1}^{2}(n_h - 1)^{-1}n_h \sum_{j=1}^{n_h} w_{d,hj}^2\hat{u}_{d,hj}^2\\ &= 2,532,658,\end{aligned}$$

where $C_H = (n-H-1)^{-1}(n-H)$ and $w_{d,hj}$ is the regression design weight of Table 2.6. The estimated variance for the procedure using PSU totals is similar to that using the probability weighted estimator based on secondary units. This small sample was used to illustrate the computations and is not representative of the large samples typical of survey practice. However, samples with a large number of elements and a relatively small number of PSUs are common. ■■

2.7 CALIBRATION

One of the attributes of the regression estimator of the mean is the property that

$$\sum_{i \in A} w_i \mathbf{x}_i = \bar{\mathbf{x}}_N, \tag{2.7.1}$$

where the w_i are the regression weights. The property (2.7.1), called the *calibration property,* has been discussed by Deville and Särndal (1992) and Särndal (2007). One way to construct weights with the calibration property is to minimize a function of the weights subject to the restriction (2.7.1). See (2.2.21).

A procedure for constructing estimators discussed by Deville and Särndal (1992) is to minimize a function of the distance between an initial weight α_i and a final weight w_i subject to the calibration restriction. Let the distance function between α_i and w_i be denoted by $G(w_i, \alpha_i)$. Then the problem is to minimize

$$\sum_{i \in A} G(w_i, \alpha_i) \tag{2.7.2}$$

subject to the calibration constraint (2.7.1). It is assumed that $G(w, \alpha)$ is nonnegative, differentiable with respect to w, strictly convex, defined on an interval containing α, and such that $G(\alpha, \alpha) = 0$. It is also assumed that $\alpha_i = N^{-1}\pi_i^{-1}$ and

$$\sum_{i \in A} \alpha_i \mathbf{x}_i = \bar{\mathbf{x}}_N + O_p(n^{-1/2}). \tag{2.7.3}$$

The equations defining the weights are

$$g(w_i, \alpha_i) - \mathbf{x}_i \boldsymbol{\lambda} = 0, \tag{2.7.4}$$

where $\boldsymbol{\lambda}$ is a column vector of Lagrange multipliers and $g(w_i, \alpha_i) = \partial G(w_i, \alpha_i)/\partial w_i$. Let $g^{-1}(\cdot, \alpha_i)$ be the inverse function with respect to w, holding α fixed at α_i. Then

$$w_i = g^{-1}(\mathbf{x}_i \boldsymbol{\lambda}, \alpha_i). \tag{2.7.5}$$

From the calibration equation (2.7.1) and equation (2.7.5),

$$\sum_{i \in A} g^{-1}(\mathbf{x}_i \boldsymbol{\lambda}, \alpha_i) \mathbf{x}_i = \bar{\mathbf{x}}_N$$

defines $\boldsymbol{\lambda}$. Deville and Särndal (1992) prove, under regularity conditions, that the estimator with weights of (2.7.5) can be approximated by a regression estimator.

We give a result with a different proof than that of Deville and Särndal (1992).

Theorem 2.7.1. Let $G(w, \alpha)$ be a continuous convex function, with a first derivative that is zero for $w = \alpha$, with a continuous second derivative that is positive and bounded away from zero, with a continuous third derivative, and with $G(\alpha, \alpha) = 0$. Let a sequence of populations and sample designs be such that

$$E\{|\bar{\mathbf{x}}_{HT} - \bar{\mathbf{x}}_N|^4 \mid \mathcal{F}_N\} = O_p(n^{-2}), \tag{2.7.6}$$

$$E\left\{\max_{i\in A} |\hat{\mathbf{M}}_{xx}^{-1}\mathbf{x}_i'|^4 \mid \mathcal{F}_N\right\} = O_p(n^{0.5}), \tag{2.7.7}$$

and

$$E\{|\hat{\mathbf{M}}_{xx}^{-1}|^2 \mid \mathcal{F}_N\} = O_p(1), \tag{2.7.8}$$

where $\hat{\mathbf{M}}_{xx} = n^{-1}\Sigma_{j\in A}\mathbf{x}_j'\mathbf{x}_j$, $|\mathbf{x}_i| = (\mathbf{x}_i\mathbf{x}_i')^{0.5}$, and $|\hat{\mathbf{M}}_{xx}|$ is the determinant of $\hat{\mathbf{M}}_{xx}$. Assume that $K_1 < n\alpha_i < K_2$ for some positive K_1 and K_2, where $\alpha_i = N^{-1}\pi_i^{-1}$.

Then the w_i that minimize (2.7.2) subject to (2.7.1) satisfy

$$w_i = \alpha_i + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\left(\sum_{i\in A}\mathbf{x}_i'\phi_{ii}^{-1}\mathbf{x}_i\right)^{-1}\mathbf{x}_i'\phi_{ii}^{-1} + O_p(n^{-2}), \tag{2.7.9}$$

where

$$\phi_{ii} = \frac{\partial^2 G(\alpha_i, \alpha_i)}{\partial w^2} \tag{2.7.10}$$

is the second derivative of $G(w, \alpha)$ with respect to w evaluated at $(w, \alpha) = (\alpha_i, \alpha_i)$.

Proof. The equations associated with the Lagrangian are

$$G'(w_i, \alpha_i) + \mathbf{x}_i\boldsymbol{\lambda} = 0, \quad i = 1, 2, \ldots, n, \tag{2.7.11}$$

$$\sum_{i\in A} w_i\mathbf{x}_i' - \bar{\mathbf{x}}_N' = \mathbf{0},$$

where $G'(w_i, \alpha_i)$ is $\partial G(w, \alpha)/\partial w$ evaluated at $(w, \alpha) = (w_i, \alpha_i)$. Let

$$\epsilon_i = n^{-1}(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})\hat{\mathbf{M}}_{x\phi x}^{-1}\mathbf{x}_i'\phi_{ii}^{-1}, \tag{2.7.12}$$

where ϕ_{ii} is defined in (2.7.10) and

$$\hat{\mathbf{M}}_{x\phi x} = n^{-1}\sum_{i\in A}\mathbf{x}_i'\phi_{ii}^{-1}\mathbf{x}_i.$$

Then

$$\sum_{i\in A}(\alpha_i+\epsilon_i)\mathbf{x}_i = \bar{\mathbf{x}}_N$$

and

$$E\left\{\max_{i\in A}\epsilon_i^2\right\} \leq \left(E\left\{|\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_{HT}|^4\right\}E\left\{\max_{i\in A}|\mathbf{L}_i|^4\right\}\right)^{1/2} = O(n^{-2.75}), \tag{2.7.13}$$

where $\mathbf{L}_i = n^{-1}\hat{\mathbf{M}}_{x\phi x}^{-1}\mathbf{x}_i\phi_{ii}^{-1}$. Also,

$$\sum_{i\in A}\epsilon_i^2 = n^{-2}(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_{HT})\hat{\mathbf{M}}_{x\phi x}^{-1}\left(\sum_{i\in A}\mathbf{x}_i'\phi_{ii}^{-2}\mathbf{x}_i\right)\hat{\mathbf{M}}_{x\phi x}^{-1}(\bar{\mathbf{x}}_{HT}-\bar{\mathbf{x}}_N)' = O_p(n^{-2})$$

because $|\hat{\mathbf{M}}_{x\phi x}^{-1}| = O_p(1)$ and the ϕ_{ii}^{-1} are bounded.

A second-order expansion of the objective function is

$$\sum_{i\in A}G(\alpha_i+\epsilon_i,\ \alpha_i) = \sum_{i\in A}G''(\alpha_i+\epsilon_i^*,\ \alpha_i)\epsilon_i^2 \tag{2.7.14}$$

where ϵ_i^* is between ϵ_i and zero and we have used $G(\alpha_i,\ \alpha_i) = G'(\alpha_i,\ \alpha_i) = 0$. By (2.7.13), there is an n_o and a closed interval B_δ with zero as an interior point such that for $n > n_o$, the probability is greater than $1-\delta$ that all ϵ_i are in the interval. Then, because $|n\alpha_i|$ is bounded and $G''(w,\ \alpha)$ is continuous, $G''(\alpha_i+\epsilon_i^*,\ \alpha_i)$ is bounded for $\epsilon_i^* \in B_\delta$ and

$$\sum_{i\in A}G(\alpha_i+\epsilon_i,\ \alpha_i) = O_p(n^{-2}). \tag{2.7.15}$$

Now

$$\min_w\left(\sum_{i\in A}G(w_i,\ \alpha_i)\right) \leq \sum_{i\in A}G(\alpha_i+\epsilon_i,\ \alpha_i) = O_p(n^{-2}).$$

By the continuity and positivity of the second derivative of $G(w_i,\alpha_i)$,

$$\sum_{i\in A}G(w_i,\alpha_i) = \sum_{i\in A}G''(w_i^*,\alpha_i)(w_i-\alpha_i)^2 = O_p(n^{-2})$$

for the w_i that minimize (2.7.2), where w_i^* is between w_i and α_i. Therefore,

$$\sum_{i \in A} G(w_i,\ \alpha_i) \quad = \quad \sum_{i \in A} 0.5\phi_{ii}(w_i - \alpha_i)^2 \ + \ O_p(n^{-2})$$

and, as n increases, the w_i that minimize

$$\sum_{i \in A} \phi_{ii}(w_i - \alpha_i)^2 + \left(\sum_{i \in A} w_i \mathbf{x}_i - \bar{\mathbf{x}}_N \right) \boldsymbol{\lambda}$$

converge to the w_i that minimize (2.7.2) subject to (2.7.1).

To establish the order of the remainder in (2.7.9), we return to the defining equations (2.7.11). Expanding (2.7.11) about α_i yields

$$G''(\alpha_i,\ \alpha_i)\ (w_i - \alpha_i) + G'''(\alpha_i^*,\ \alpha_i)\ (w_i - \alpha_i)^2 + \mathbf{x}_i \boldsymbol{\lambda} \quad = \quad 0, \quad (2.7.16)$$

where w_i is the solution to (2.7.11) and α_i^* is between w_i and α_i. Multiplying by $\mathbf{x}_i' \phi_{ii}^{-1}$ and summing, we have

$$(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})' + \sum_{i \in A} \mathbf{x}_i' \phi_{ii}^{-1} G'''(\alpha_i^*,\ \alpha_i)\ (w_i - \alpha_i)^2 + \sum_{i \in A} \mathbf{x}_i' \phi_{ii}^{-1} \mathbf{x}_i \boldsymbol{\lambda} \quad = \quad \mathbf{0}.$$

By assumption, $G'''(\cdot,\ \alpha_i)$ is continuous and

$$\sum_{i \in A} \left| \left(\sum_{i \in A} \mathbf{x}_i' \phi_{ii}^{-1} \mathbf{x}_i \right)^{-1} \mathbf{x}_i' \phi_{ii}^{-1} \right| \quad = \quad O_p(1).$$

Therefore,

$$\boldsymbol{\lambda} \quad = \quad - \left(\sum_{i \in A} \mathbf{x}_i' \phi_{ii}^{-1} \mathbf{x}_i \right)^{-1} (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})' \ + \ O_p(n^{-2})$$

and

$$w_i \quad = \quad \alpha_i + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT}) \left(\sum_{i \in A} \mathbf{x}_i' \phi_{ii}^{-1} \mathbf{x}_i \right)^{-1} \mathbf{x}_i' \phi_{ii}^{-1} \ + \ O_p(n^{-2}).$$

■

We note that assumptions (2.7.6) and (2.7.7) will be satisfied by simple random samples from a distribution with eighth moment. See David (1981, p. 56).

It follows from the properties of the weights of (2.7.9) that the associated estimator is a regression estimator. This is made explicit in the corollary.

Corollary 2.7.1.1. Let the assumptions of Theorem 2.7.1 hold. Let y be a variable such that $E\{(\bar{y}_{HT} - \bar{y}_N)^2 \mid \mathcal{F}_N\} = O_p(n^{-1})$. Then the w_i that minimize (2.7.2) subject to (2.7.1) satisfy

$$\sum_{i \in A} w_i y_i = \bar{y}_{HT} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{HT})\hat{\boldsymbol{\beta}}_\phi + O_p(n^{-1}), \tag{2.7.17}$$

where

$$\hat{\boldsymbol{\beta}}_\phi = \left(\sum_{i \in A} \mathbf{x}_i' \phi_{ii}^{-1} \mathbf{x}_i\right)^{-1} \sum_{i \in A} \mathbf{x}_i' \phi_{ii}^{-1} y_i.$$

Proof. The w_i are given in (2.7.9). Because the remainder term of (2.7.9) has moments, the sum of products of remainder terms and y is $O_p(n^{-1})$. ∎

We have emphasized the construction of weights that minimize a variance or a quantity closely related to variance. In some cases, when $G(w_i, \alpha_i)$ is a quadratic, the minimum distance weights are the same as the minimum variance weights.

Lemma 2.7.1. Let $\mathbf{X}$ be an $n \times k$ matrix and let $\boldsymbol{\Sigma}_{ee}$ be a positive definite symmetric matrix. Assume that $\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X}$ is nonsingular and that $\boldsymbol{\Sigma}_{ee}\boldsymbol{\alpha}$ is in the column space of $\mathbf{X}$, where $\boldsymbol{\alpha}$ is a given n-dimensional vector. Then $\mathbf{w}_a = \mathbf{w}_b$, where $\mathbf{w}_b$ minimizes the Lagrangian

$$\mathbf{w}_b'\boldsymbol{\Sigma}_{ee}\mathbf{w}_b + (\mathbf{w}_b'\mathbf{X} - \bar{\mathbf{x}}_N)\boldsymbol{\lambda}_b, \tag{2.7.18}$$

$\mathbf{w}_a$ minimizes the Lagrangian

$$(\mathbf{w}_a - \boldsymbol{\alpha})'\boldsymbol{\Sigma}_{ee}(\mathbf{w}_a - \boldsymbol{\alpha}) + (\mathbf{w}_a'\mathbf{X} - \bar{\mathbf{x}}_N)\boldsymbol{\lambda}_a, \tag{2.7.19}$$

and $\boldsymbol{\lambda}_a$ and $\boldsymbol{\lambda}_b$ are vectors of Lagrangian multipliers.

Proof. By assumption, $\boldsymbol{\Sigma}_{ee}\boldsymbol{\alpha}$ is in the column space of $\mathbf{X}$. Therefore, we can find a nonsingular transformation of $\mathbf{X}$ such that the first column of the transformed matrix is $\mathbf{Z}_1 = \boldsymbol{\Sigma}_{ee}\boldsymbol{\alpha}$. Denote the transformed matrix by $\tilde{\mathbf{X}} = (\mathbf{Z}_1, \tilde{\mathbf{X}}_2)$, where $\tilde{\mathbf{X}}_2$ is an $n \times (k-1)$ matrix of full rank. Transform the matrix further by letting

$$\mathbf{Z}_2 = \tilde{\mathbf{X}}_2 - \mathbf{Z}_1\boldsymbol{\Gamma},$$

where

$$\boldsymbol{\Gamma} = (\mathbf{Z}_1'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\boldsymbol{\Sigma}_{ee}^{-1}\tilde{\mathbf{X}}_2 = (\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{ee}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}'\tilde{\mathbf{X}}_2.$$

To simplify the notation, let, with no loss of generality, $\tilde{\mathbf{X}}_2 = \mathbf{X}_2$. Then the calibration constraint becomes $\mathbf{w}'\mathbf{Z} = \bar{\mathbf{z}}_N$, where $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ and $\bar{\mathbf{z}}_N = (\bar{z}_{1,N}, \bar{\mathbf{x}}_{2,N} - \bar{z}_{1,N}\boldsymbol{\Gamma}')$. The solution to (2.7.19) is

$$\begin{aligned}
\mathbf{w}_a' &= \boldsymbol{\alpha}' + (\bar{\mathbf{z}}_N - \bar{\mathbf{z}}_\alpha)(\mathbf{Z}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{ee}^{-1} \\
&= \boldsymbol{\alpha}' + (\bar{z}_{1,N} - \bar{z}_{1,\alpha})(\mathbf{Z}_1'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z}_1)^{-1}\boldsymbol{\alpha}' \\
&\quad + (\bar{\mathbf{z}}_{2,N} - \bar{\mathbf{z}}_{2,\alpha})(\mathbf{Z}_2'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\boldsymbol{\Sigma}_{ee}^{-1} \\
&= \bar{z}_{1,N}(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{ee}\boldsymbol{\alpha})^{-1} + \bar{\mathbf{z}}_{2N}(\mathbf{Z}_2'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\boldsymbol{\Sigma}_{ee}^{-1},
\end{aligned}$$

where $\bar{\mathbf{z}}_\alpha = (\bar{z}_{1,\alpha}, \bar{\mathbf{z}}_{2,\alpha}) = \boldsymbol{\alpha}'\mathbf{Z} = (\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{ee}\boldsymbol{\alpha}, \bar{\mathbf{z}}_{2,\alpha})$ and, by construction, $\bar{\mathbf{z}}_{2,\alpha} = \mathbf{0}$. The solution to (2.7.18) is

$$\begin{aligned}
\mathbf{w}_b' &= \bar{\mathbf{z}}_N(\mathbf{Z}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{ee}^{-1} \\
&= \bar{z}_{1,N}(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{ee}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}' + \bar{\mathbf{z}}_{2,N}(\mathbf{Z}_2'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\boldsymbol{\Sigma}_{ee}^{-1} \\
&= \mathbf{w}_a'
\end{aligned}$$

and we have the conclusion. ■

2.8 WEIGHT BOUNDS

We have described the regression estimator as a function that is linear in y and have demonstrated procedures that can be used to construct weights. For most of these procedures there are no bounds on possible values for the weights. If the weights are to be used for estimating totals for a wide range of characteristics of the finite population, every weight should be greater than 1. Also, a very large weight on an observation can result in large variances for some estimates. Therefore, in the application of regression estimation, procedures are used to control the magnitude of the weights.

If the weights are constructed by minimizing a quadratic function, the bounds can be added as constraints in a quadratic programming problem. For example, the weights might be constructed by minimizing

$$\mathbf{w}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{w} \tag{2.8.1}$$

subject to the constraints

$$\mathbf{w}'\mathbf{X} - \bar{\mathbf{x}}_N = \mathbf{0}, \tag{2.8.2}$$

$$L_1 \leq w_i \leq L_2, \qquad i \in A, \tag{2.8.3}$$

where L_1 and L_2 are arbitrary constants. A popular version of constraint (2.8.3) restricts w_i to be greater than 1 for an estimated total or greater than N^{-1} for an estimated mean. There is no guarantee that a solution exists for this problem. For example, if the population mean of x is zero and all observations in the sample are positive, there is no set of positive weights that sum to 1 and such that $\Sigma_{i\in A} w_i x_i = 0$. At least one element with a negative x value and at least one with a positive x value are required to construct a set of weights satisfying the two restrictions.

Deville and Särndal (1992) discuss ways to define objective functions that contain bounds on the weights. Huang and Fuller (1978) defined an iterative procedure that restricted the range of the weights by adding a function of $\mathbf{x}_i$ to the minimization problem. Increasing the value of the function, equivalent to a larger σ_{eeii} in (2.8.1), will result in a weight for i that is closer to the mean weight.

A procedure that produces positive weights in a large fraction of cases is that discussed in Section 2.5. The weights b_i of (2.5.8) are always positive and, by construction, sum to 1. The weights w_i of (2.5.10) will be positive much more often than will the ordinary regression weights.

If it is impossible to find weights satisfying (2.8.2) and (2.8.3), or if the calibration property is not important, weights can be constructed to meet (2.8.3) by relaxing the calibration constraint (2.8.2). Husain (1969) used quadratic programming to obtain positive weights. He also considered, for $\mathbf{\Sigma}_{ee} = \mathbf{I}$ and simple random sampling, an alternative to (2.8.1)–(2.8.3). He minimized

$$\mathbf{w}'\mathbf{w} + \gamma(\mathbf{w}'\mathbf{X} - \bar{\mathbf{x}}_N)\mathbf{\Sigma}_{\bar{x}\bar{x}}^{-1}(\mathbf{w}'\mathbf{X} - \bar{\mathbf{x}}_N)' \tag{2.8.4}$$

subject to

$$L_1 \leq w_i \leq L_2. \tag{2.8.5}$$

The problem (2.8.4)–(2.8.5) always has a solution. Husain showed that for simple random sampling the value of γ that minimizes the variance under the linear normal model is

$$\gamma_{opt} = [k(1 - R^2)]^{-1}(n - k - 2)R^2,$$

where k is the dimension of $\mathbf{x}_1$, the vector without the 1, and R^2 is the squared multiple correlation between $\mathbf{x}_1$ and y.

Rao and Singh (1997) proposed a method of ridge shrinkage to satisfy bounds on the weights and to satisfy the calibration constraints within specified tolerances. Another method used to obtain weights meeting bounds requirements is to drop x variables from the model.

2.9 MAXIMUM LIKELIHOOD AND RAKING RATIO

In many situations the number of elements in certain categories are known or can be treated as known. For example, it is common practice in surveys of the human population to use updated census numbers as auxiliary information. To study estimation for such situations, consider a simple random sample from a multinomial distribution defined by the entries in a two-way table. The logarithm of the likelihood, except for a constant, is

$$\sum_{k=1}^{r}\sum_{m=1}^{c} a_{km}\log p_{km}, \tag{2.9.1}$$

where a_{km} is the estimated fraction in cell km, p_{km} is the population fraction in cell km, r is the number of rows, and c is the number of columns. If (2.9.1) is maximized subject to the restriction $\Sigma\Sigma\, p_{km} = 1$, one obtains the maximum likelihood estimators $\hat{p}_{km} = a_{km}$. Now assume that the marginal row fractions $p_{k\cdot,N}$, $k = 1, 2, \ldots, r$, and the marginal column fractions $p_{\cdot m,N}$, $m = 1, 2, \ldots, c$, are known. By analogy to (2.2.21) we define the estimators of p_{im} to be the p_{km} that maximize the likelihood subject to the constraints by using the Lagrangian

$$\sum_{k=1}^{r}\sum_{m=1}^{c} a_{km}\log p_{km} + \sum_{k=1}^{r}\lambda_k\left(\sum_{m=1}^{c} p_{km} - p_{k\cdot,N}\right) + \sum_{m=1}^{c}\lambda_{r+m}\left(\sum_{k=1}^{r} p_{km} - p_{\cdot m,N}\right), \tag{2.9.2}$$

where λ_k, $k = 1, 2, \ldots, r$, are for the row restrictions and λ_{r+m}, $m = 1, 2, \ldots, c$, are for the column restrictions. There is no explicit expression for the solution to (2.9.2), and iterative methods are required.

If we expand (2.9.1) about the a_{km}, we have

$$\sum_{k=1}^{r}\sum_{m=1}^{c} a_{km}\log p_{km} \doteq \sum_{k=1}^{r}\sum_{m=1}^{c}\left[a_{km}\log a_{km} + (p_{km} - a_{km}) - a_{km}^{-1}(p_{km} - a_{km})^2\right]. \tag{2.9.3}$$

Minimizing the right side of (2.9.3) subject to the constraints gives a first approximation to the solution. The procedure can be iterated replacing a_{km} with the estimates from the previous round, but convergence is not guaranteed if there are too many empty cells. See Ireland and Kullback (1968) and Bishop, Fienberg, and Holland (1975, Chapter 3).

A procedure that produces estimates close to the maximum likelihood solution is called *raking ratio* or *iterative proportional fitting*. The procedure iterates, first making ratio adjustments for the row restrictions, then making ratio adjustments for the column restrictions, then making ratio adjustments for the row restrictions, and so on. The method was given by Deming and Stephan (1940); see also Stephan (1942). As with the likelihood procedure, convergence is not guaranteed. Raking has been used heavily in practice. Often, only a few iterations are used, with the final ratio adjustment made on the category deemed to be most important.

Deville and Särndal (1992) suggested the objective function

$$\sum_{i \in A} \left[w_i \log(\alpha_i^{-1} w_i) + \alpha_i - w_i \right], \tag{2.9.4}$$

where α_i are initial weights, to construct the raking ratio weights and the function

$$\sum_{i \in A} \left[w_i - \alpha_i - \alpha_i \log(\alpha_i^{-1} w_i) \right] \tag{2.9.5}$$

to construct the maximum likelihood weights. By Theorem 2.7.1, the estimators obtained by minimizing (2.9.4) and (2.9.5) are asymptotically equivalent to regression estimators with weights defined by the second derivatives. The second derivative of (2.9.4) is w_i^{-1} and that of (2.9.5) is $\alpha_i w_i^{-2}$. Because both derivatives evaluated at $w_i = \alpha_i$ are α_i^{-1}, both estimators are close to the regression estimator with weights that minimize

$$\sum_{i \in A} \alpha_i^{-1} (w_i - \alpha_i)^2 .$$

2.10 REFERENCES

Section 2.1. Beale (1962), Cochran (1977), Mussa (1999).

Section 2.2. Cochran (1942, 1977), Fuller (1975, 2002), Fuller and Isaki (1981), Hidiroglou, Fuller, and Hickman (1980), Kauermann and Carroll (2001), Mickey (1959), Montanari (1987, 1998, 1999), Rao (1994), Särndal (1982), Särndal, Swensson, and Wretman (1989,1992).

Section 2.3. Brewer (1963b), Brewer, Hanif, and Tam (1988), Cassel, Särndal, and Wretman (1976, 1979, 1983), Firth and Bennett (1998), Fuller (1975, 2002), Fuller and Hidiroglou (1978), Fuller and Isaki (1981), Hidiroglou (1974), Isaki (1970), Isaki and Fuller (1982), Little (2004), Montanari (1999), Pfeffermann (1984), Royall (1970, 1976, 1992), Särndal (1980), Scott and Smith (1974), Tam (1986,1988), Wright (1983), Wu and Sitter (2001), Zyskind (1976).

Section 2.4. Fuller (1996, 2002), Holt and Smith (1979), Rao (1994).

Section 2.5. Fuller (2002), Park (2002), Tillé (1998, 1999).

Section 2.7. Deville and Särndal (1992), Särndal (2007).

Section 2.8. Deville and Särndal (1992), Huang and Fuller (1978), Husain (1969), Park (2002), Park and Fuller (2005), Rao and Singh (1997).

Section 2.9. Bishop, Fienberg, and Holland (1975), Deville and Särndal (1992), Fuller (2002), Ireland and Kullback (1968), Park (2002).

2.11 EXERCISES

1. (Section 2.2.1) Show that the estimator

$$\sum_{i \in A} w_i y_i$$

can be viewed as the regression coefficient for the regression of y_i on x_i, where $x_i = (\sum_{j \in A} w_j^2)^{-1} w_i$.

2. (Section 2.3) Compute the entries of Table 2.2 with the order of the original weights reversed. That is, the original weight for element 1 is 0.15 and the original weight for the last element is 0.07. Compare the estimators as estimators of the total of y under model (2.3.26).

3. (Section 2.3.2) Assume that $\hat{\theta}$ of (2.3.35) is the θ that minimizes the objective function

$$\sum_{i \in A} [y_i - \alpha(\mathbf{x}_i, \theta)]^2 .$$

Assume that $\alpha(\mathbf{x}_i, \theta)$ is continuous with continuous first and second derivatives. What condition for the first derivative is sufficient for the design consistency criterion (2.3.37)?

4. (Section 2.3) Compute the weights for the regression estimator $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}_\pi$, where $\mathbf{x}_i = (1, x_{1,i})$ and $\hat{\boldsymbol{\beta}}_\pi$ is defined in (2.2.56) for the sample of

Table 2.2. Compare the conditional model variances of Example 2.3.1 with the conditional model variance of the regression estimator using (2.2.56).

5. (Section 2.3) Show that minimizing the Lagrangian

$$\sum_{i\in A}(w_i - n^{-1})^2 n \; + \; \lambda_1 \left(\sum_{i\in A} w_i - 1 \right) \; + \; \lambda_2 \left(\sum_{i\in A} w_i x_i - \bar{x}_N \right)$$

produces the same w_i as minimizing

$$\sum_{i\in A} w_i^2 + \lambda_1 \left(\sum_{i\in A} w_i - 1 \right) + \lambda_2 \left(\sum_{i\in A} w_i x_i - \bar{x}_N \right).$$

6. (Section 2.1, 2.2) Let x be an auxiliary variable and let

$$\begin{array}{rcll} y & = & x & \text{if } x < C \\ & = & 0 & \text{otherwise.} \end{array}$$

Assume that x is distributed as a uniform random variable on $(0,\ 1)$ and that the mean of x is known. For a random sample of size n, let the mean of y be estimated with the ratio estimator (2.1.1). Plot y against x for $C = 0.5$. Plot the least squares regression line for $C = 0.5$. For what values of C is the ratio estimator (2.1.1) superior to the simple mean?

7. (Section 2.3) Let a population be composed of two equal-sized strata. Assume that

$$\begin{pmatrix} y_{1i} \\ x_{1i} \end{pmatrix} \sim NI\left(\begin{pmatrix} \mu_{y1} \\ \mu_{x1} \end{pmatrix}, \begin{pmatrix} 1.25 & 0.50 \\ 0.50 & 1.00 \end{pmatrix} \right)$$

in stratum 1, and

$$\begin{pmatrix} y_{2i} \\ x_{2i} \end{pmatrix} \sim NI\left(\begin{pmatrix} \mu_{y2} \\ \mu_{x2} \end{pmatrix}, \begin{pmatrix} 3.25 & 1.50 \\ 1.50 & 1.00 \end{pmatrix} \right)$$

in stratum 2. Assume that a stratified sample is selected with two elements in stratum 1 and nine elements in stratum 2.

(a) Give the design optimal regression weights for the case in which only $\bar{x}_N$ is known. What is the expected value of the design variance through the order n^{-1} terms?

(b) Assume that the survey was designed under the assumption that

$$\begin{aligned} y_{hi} &= \mu_{hi} + (x_{hi} - \bar{x}_{hN})\beta + e_{hi} \\ e_{hi} &\sim ii(0,\ \sigma_h^2). \end{aligned}$$

Construct the best unbiased estimator of the mean of y conditional on the realized x_{hi} and known $\bar{x}_N$.

8. (Section 2.2, 2.3) Regression can be used to convert the Horvitz–Thompson estimator into a location and scale-invariant estimator. Construct the regression estimator that minimizes the Lagrangian

$$\sum_{i\in A} w_i^2\pi_i^2 + (\lambda_1,\ \lambda_2)\left[\left(\sum_{i\in A} w_i - 1\right),\ \left(\sum_{i\in A} w_i\pi_i - N^{-1}\sum_{i\in U}\pi_i\right)\right]'.$$

Assume that there are different values of π_i to avoid singularities. Show that the estimator, denoted by $\bar{y}_{\pi,reg}$, is scale and location invariant. Let the finite population be a sample from a superpopulation satisfying

$$\begin{aligned} y_i &= (1,\ \pi_i)\boldsymbol{\beta} + e_i, \\ e_i &\sim ind(0,\ \pi_i^2\sigma^2). \end{aligned}$$

Give $V\{\bar{y}_{\pi,reg}|\mathbf{X}, \bar{\mathbf{x}}_N\}$. Compare the variance to the conditional model variances of

$$\bar{y}_{HT} = N^{-1}\sum_{i\in A}\pi_i^{-1}y_i$$

and

$$\bar{y}_{\pi} = \left(\sum_{i\in A}\pi_i^{-1}\right)^{-1}\sum_{i\in A}\pi_i^{-1}y.$$

Compare $E\{V_\infty(\cdot|\mathcal{F})\}$ for the three estimators as a function of (β_0, β_1) under the assumption of Poisson sampling, where $V_\infty\{\cdot|\mathcal{F}\}$ is the variance of the approximate distribution.

9. (Section 2.3) Assume that the vector $\mathbf{y}$ associated with Table 2.2 is

$$\mathbf{y}' = (0.5,\ 1.1,\ 1.7,\ 1.7,\ 2.9,\ 3.2,\ 3.9,\ 4.8,\ 5.9,\ 6.5).$$

Test the hypothesis that the reduced model (2.3.26) is adequate against the alternative model (2.3.33). Use the estimated design variance and assume that the sample is a Poisson sample. Construct the regression

estimator (2.3.29) assuming that $\bar{x}_N = 2.283$. Estimate the design variance of the estimator. Ignore the finite population correction.

10. (Section 2.2.3) Let a simple random sample of size 21 be selected and classified into three poststrata as displayed in Table 2.7. Assume that the population sizes for the three poststrata are $N_1 = 20$, $N_2 = 30$, $N_3 = 50$. Estimate the mean using poststratification. Estimate the variance using equation (2.2.74). Estimate the variance using equation (2.2.31).

Table 2.7 Poststratified Observations

Post-stratum 1	Post-stratum 2	Post-stratum 3
40	99	108
50	109	113
15	104	93
10	49	83
	53	89
		56
		73
		95
		74
		115
		64
		86

11. (Section 2.2, 2.3) Consider the model

$$\begin{aligned} y_i &= \beta_0 + e_i, \\ e_i &\sim ind(0, \sigma^2). \end{aligned}$$

Assume that a sample of size n has been selected with probabilities π_i. Construct the model optimal estimator for β_0. Then construct the model optimal regression estimator of $\bar{y}_N$ for the model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \pi_i^{-1} + e_i, \\ e_i &\sim ind(0, \sigma^2), \end{aligned}$$

assuming that the population mean of π_i^{-1} is known. Is the second estimator design consistent? Explain. Compare the model variances of the two estimators.

12. (Section 2.2) For a simple random sample of size n, let a sample 2×2 table have entries n_{ij}, where

$$n = \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij}.$$

(a) Assume that the row marginals $N_{i\cdot}$, $i = 1, 2$, are known. Find estimates of the cell proportions, denoted by p_{ij}, by finding quantities m_{ij} that minimize

$$\sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij}^{-1}(m_{ij} - n_{ij})^2$$

subject to the restrictions

$$\sum_{j=1}^{2} m_{ij} = N^{-1} N_{i\cdot} n, \qquad i = 1, 2.$$

(b) Compare the estimators of (a) with estimators of p_{ij} obtained with the ordinary regression estimator, where the regression estimator is constructed using individual sample elements.

(c) Assume that the row marginals $N_{i\cdot}$, $i = 1, 2$, and the column marginals, $N_{\cdot j}$, $j = 1, 2$, are known. Construct the ordinary regression estimator of the cell proportions.

13. (Section 2.2.2) Assume that a sequence of finite populations is created as realizations of iid random variables generated from a distribution function with finite fifth moments. Let a Poisson sample be selected with probabilities π_i, where the π_i are not all equal. Let $\mathbf{x}_i = (1, \pi_i^{-1})$ and define $\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\Sigma_{i \in A} \mathbf{x}_i' \mathbf{x}_i)^{-1} \Sigma_{i \in A} \mathbf{x}_i' y_i$. Assume that $\{\pi_i\}$ is a fixed sequence, that $lim_{N \to \infty} n_B^{-1} \Sigma_{i \in U} \pi_i = \mu_\pi$, and that $K_L < N n_B^{-1} \pi_i < K_U$ for positive K_L and K_U, where n_B is the expected sample size.

(a) Show that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N \mid \mathcal{F}_N$ converges to zero in probability almost surely, where

$$\boldsymbol{\beta}_N = \begin{pmatrix} n_B & N \\ N & \sum_{i \in U} \pi_i^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i \in U} \pi_i y_i \\ \sum_{i \in U} y_i \end{pmatrix}.$$

(b) Show that $\Sigma_{i\in U} e_i = 0$ almost surely, where $e_i = y_i - \mathbf{x}_i\boldsymbol{\beta}_N$.

(c) Show that

$$\bar{\mathbf{x}}_N \left(N^{-1} \sum_{i\in A} \mathbf{x}_i'\mathbf{x}_i \right)^{-1} - (0,\ 1) \ = \ O_p\left(n_B^{-1/2}\right).$$

(d) Show that

$$\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}} - \bar{y}_N \ = \ N^{-1} \sum_{i\in A} \pi_i^{-1} e_i + O_p\left(n_B^{-1}\right).$$

14. (Section 2.3) The regression estimator of (2.3.3) can be viewed as the fitted regression model

$$\hat{y}_i = \bar{y}_\pi + (\mathbf{x}_i - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}$$

evaluated at $\mathbf{x}_i = \bar{\mathbf{x}}_N$. For a regression estimator of the mean defined in terms of a fitted regression, the regression estimator of the mean will be design consistent if the fitted regression passes through $(\bar{y}_c, \bar{\mathbf{x}}_c)$, where $(\bar{y}_c, \bar{\mathbf{x}}_c)$ is a vector of design-consistent estimators. Let a regression estimator of the mean be defined by

$$\bar{y}_{reg,b} = \sum_{i\in A} (\alpha_i + b_i) y_i,$$

where $\alpha_i = (\Sigma_{j\in A}\pi_j^{-1})^{-1}\pi_i^{-1}$. Let $\mathbf{b} = (b_1,\ b_2,\ \ldots,\ b_n)'$, $\mathbf{z}_{ci} = (1,\ \mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})$, $\mathbf{Z}_c = (\mathbf{z}_{c1}', \mathbf{z}_{c2}',\ \ldots,\ \mathbf{z}_{cn}')'$, and $\bar{\mathbf{z}}_c = (0,\ \bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})$. Show that

$$\mathbf{b}' \ = \ \bar{\mathbf{z}}_c(\mathbf{Z}_c'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\boldsymbol{\Sigma}_{ee}^{-1}$$

minimizes the Lagrangian,

$$\mathbf{b}'\boldsymbol{\Sigma}_{ee}\mathbf{b} + \sum_{j=1}^{k+1} \lambda_j \left(\sum_{i=1}^{n} b_i z_{cji} - \bar{z}_{cj} \right).$$

Is the regression estimator constructed with $\boldsymbol{\alpha}$ and $\mathbf{b}$ location invariant? Is the estimator with $\alpha_{HT,i} = N^{-1}\pi_i^{-1}$ and $\mathbf{z}_{ci,HT} = (1, \mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,HT})$, where $(\bar{y}_{HT}, \bar{\mathbf{x}}_{1,HT}) = \Sigma_{i\in A}\alpha_{HT,i}(y_i, \mathbf{x}_{1,i})$, location invariant?

15. (Section 2.7) Prove the following:

Result. Let $\mathcal{F}_N = \{y_1,\ y_2,\ \ldots,\ y_N\}$, where $\{y_i\}$ is a fixed sequence. Assume that a sequence of Poisson samples is selected from the sequence of populations. Assume that the selection probabilities satisfy

$$K_L \ \leq \ \pi_i \ \leq \ K_U$$

for all i, where K_L and K_U are positive constants. Assume that

$$\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} (x_i,\ x_i^2) \quad = \quad (\mu,\ M_2),$$

where $M_2 - \mu^2 > 0$. Then

$$E\left\{\max_{i\in A} x_i\right\} \quad = \quad O(N^{1/2}).$$

16. (Section 2.3) Consider the model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i,$$
$$e_i \sim ind(0, x_{1,i}\sigma^2),$$

where $E\{|\ x_{1,i}e_i\ |^{2+\delta}\}$ is finite, $\mathbf{x}_i = (1, x_{1,i}) = \left(1, Nn_B^{-1}\pi_i\right)$, n_B is the expected sample size and $x_{1,i} > 0$. Define a regression estimator by

$$\bar{y}_{reg} = (1, \bar{\mathbf{x}}_{1,N})\,\hat{\boldsymbol{\beta}},$$

where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A} \mathbf{x}_i'\pi_i^{-1}\mathbf{x}_i\right)^{-1} \sum_{i\in A} \mathbf{x}_i'\pi_i^{-1}y_i.$$

Is $\bar{y}_{reg}$ location and scale invariant? Let the sequence $\{\pi_i\}$ satisfy

$$\lim_{N\to\infty} N^2 n_B^{-2} \sum_{i\in U_N} (\pi_i - \bar{\pi}_N)^2 = \sigma_\pi^2,$$

where $\sigma_\pi^2 > 0$ is finite. Is $\bar{y}_{reg}$ design consistent under Poisson sampling? Compare $V\{\cdot \mid \mathbf{X}, \bar{\mathbf{x}}_N\}$ for $\bar{y}_{reg}, \bar{y}_\pi$ and $\bar{y}_{HT}$, where

$$\bar{y}_\pi = \left(\sum_{i\in A} \pi_i^{-1}\right)^{-1} \sum_{i\in A} \pi_i^{-1}y_i$$

and $\bar{y}_{HT}$ is as defined in (1.2.24).

17. (Section 2.3) Consider the model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i,$$

where $\mathbf{x}_i = \left(1, \pi_i, \pi_i^{-1}\right)$ and e_i are independent $\left(0, \sigma_{e,i}^2\right)$ random variables independent of $\mathbf{x}_i$. Let samples be selected with probabilities

$\pi_i, i = 1, 2, \ldots, N$, using a design such that $\bar{y}_\pi$ is design consistent. Define a regression estimator by

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}},$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Let $\mathbf{D}_\pi = \text{diag}(\pi_1, \pi_2, \ldots, \pi_n)$. Show that $\bar{y}_{reg}$ is design consistent for $\mathbf{V} = \mathbf{I}$, $\mathbf{V} = \mathbf{D}_\pi$, and $\mathbf{V} = \mathbf{D}_\pi^2$. Define a situation for which each $\mathbf{V}$ is the preferred $\mathbf{V}$. Assume that $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ is nonsingular.

18. (Section 2.1) Let model (2.1.10) hold and let a simple random sample of size n be available. What is the best predictor of $\Sigma_{i \in A^c} y_i$, the sum of the elements not sampled, given that $\bar{x}_N$ is known? Show that the best predictor of the population total is $\hat{\beta}$ of (2.1.11) multiplied by $N\bar{x}_N$.

19. (Section 2.2) Construct a regression estimator of the form $\bar{y}_{reg} = \Sigma_{i \in A} w_i y_i$ by minimizing

$$\sum_{i \in A} \pi_i^2 (1 - \pi_i)^{-1} w_i^2$$

subject to the restriction

$$\sum_{i \in A} w_i \mathbf{x}_i = \bar{\mathbf{x}}_N,$$

where $\mathbf{x}_i = (1, \pi_i(1 - \pi_i)^{-1}, \mathbf{x}_{2i})$. Is the regression estimator design consistent given that the probability that element i is in the sample is π_i? Assume that the sequence of designs is such that the Horvitz–Thompson estimator of the first three moments of $(y_i, \mathbf{x}_i)$ are design consistent. If the sample is a Poisson sample, is the estimator the design-optimal estimator of Theorem 2.2.4? Is there an $\mathbf{x}_{2i}$ such that the estimator for a stratified sample is the optimal estimator of Theorem 2.2.4?

20. (Section 2.2) Assume that we select a Poisson sample with selection probabilities π_i. Give an expression for the covariance between $\bar{x}_{HT}$ and $\bar{y}_{HT}$. Give an expression for the approximate covariance between $\bar{x}_\pi$ and $\bar{y}_\pi$. Consider the regression estimator

$$\bar{y}_{HT} + (\bar{x}_N - \bar{x}_{HT})[\hat{V}\{\bar{x}_{HT} \mid \mathcal{F}\}]^{-1}\hat{C}\{\bar{x}_{HT}, \bar{y}_{HT} \mid \mathcal{F}\}$$

and the regression estimator

$$\bar{y}_\pi + (\bar{x}_N - \bar{x}_\pi)[\hat{V}\{\bar{x}_\pi \mid \mathcal{F}\}]^{-1}\hat{C}\{\bar{x}_\pi, \bar{y}_\pi \mid \mathcal{F}\}.$$

Given the large-sample approximations, is one estimator always superior to the other?

21. (Section 2.7) Let A be a simple random sample of size n, and let

$$\bar{y}_{reg} = \bar{y}_n + (\bar{x}_N - \bar{x}_n)\hat{\beta}_{IV},$$

where

$$\hat{\beta}_{IV} = \left(\sum_{i\in A} z_i(x_i - \bar{x}_N)\right)^{-1} \sum_{i\in A} z_i y_i$$

and

$$\begin{aligned} z_i &= 1 \qquad \text{if } x_i > \bar{x}_N \\ &= 0 \qquad \text{if } x_i = \bar{x}_N \\ &= -1 \qquad \text{if } x_i < \bar{x}_N. \end{aligned}$$

Is $\bar{y}_{reg}$ calibrated in the sense of (2.7.1)? Is the estimator location invariant? Assume a sequence of populations such that estimators of moments have errors that are $O_p(n^{-1/2})$. Show that $\hat{y}_{reg}$ is design consistent. Give the large-sample variance of $\bar{y}_{reg}$. Assume that the distribution of x is symmetric with $P\{x_i = \bar{x}_N\} = 0$. What is the probability that all regression weights are positive?

22. (Section 2.2) For a sequence of samples and populations, let $w_i = N^{-1}\pi_i^{-1}$,

$$\begin{pmatrix} \hat{\sigma}_{yy} & \hat{\sigma}_{yx} \\ \hat{\sigma}_{xy} & \hat{\sigma}_{xx} \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{i\in A} w_i(y_i - \bar{y}_{HT})^2 & \sum_{i\in A} w_i(x_i - \bar{x}_{HT})(y_i - \bar{y}_{HT}) \\ \sum_{i\in A} w_i(x_i - \bar{x}_{HT})(y_i - \bar{y}_{HT}) & \sum_{i\in A} w_i(x_i - \bar{x}_{HT})^2 \end{pmatrix},$$

$$(\bar{y}_{HT}, \bar{x}_{HT}) = \sum_{i\in A} w_i(y_i, x_i),$$

$$\bar{y}_{reg} = \bar{y}_{HT} + (\bar{x}_N - \bar{x}_{HT})\hat{\beta},$$

and $\hat{\beta} = \hat{\sigma}_{xx}^{-1}\hat{\sigma}_{xy}$. Assume that the sequence is such that

$$V\{(\bar{y}_{HT}, \bar{x}_{HT}, \hat{\sigma}_{yy}, \hat{\sigma}_{xy}, \hat{\sigma}_{xx})' \mid \mathcal{F}_N\} = O(n^{-1}).$$

Show that

$$\begin{aligned}\bar{y}_{reg} &= \bar{y}_{HT} + (\bar{x}_N - \bar{x}_{HT})\boldsymbol{\beta}_N \\ &\quad + S_{xN}^{-2}(\bar{x}_N - \bar{x}_{HT})\sum_{i\in A} w_i(x_i - \bar{x}_N)e_i + O_p(n^{-1.5}),\end{aligned}$$

where $e_i = y_i - \bar{y}_N - (x_i - \bar{x}_N)\beta_N$,

$$\beta_N = \left(\sum_{i\in U}(x_i - \bar{x}_N)^2\right)^{-1}\sum_{i\in U}(x_i - \bar{x}_N)(y_i - \bar{y}_N),$$

and

$$S_{xN}^2 = (N-1)^{-1}\sum_{i\in U}(x_i - \bar{x}_N)^2.$$

Evaluate $E\{(\bar{x}_N - \bar{x}_{HT})\sum_{i\in A} w_i(x_i - \bar{x}_N)e_i \mid \mathcal{F}_N\}$.

23. (Section 2.2) Use expression (2.2.29) and the assumptions of Theorem 2.2.1 to show, for a simple random sample, that

$$\begin{aligned}\bar{y}_{reg} - \bar{y}_N &= (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})\mathbf{S}_{xx,N}^{-1}(n-1)^{-1}\sum_{i\in A_N}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,n})'a_i \\ &\quad + \bar{a}_n + O_p(n^{-1.5}) \\ &= \bar{a}_n - (1-f_N)n^{-1}(N-1)^{-1}\sum_{i=1}^{N}\mathbf{z}_{1,i}\mathbf{z}_{1,i}'a_i \\ &\quad + O_p(n^{-1.5}),\end{aligned}$$

where $f_N = N^{-1}n$, $\mathbf{z}_{1,i}' = \mathbf{S}_{xx,N}^{-0.5}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'$,

$$\mathbf{S}_{xx,N} = (N-1)^{-1}\sum_{i=1}^{N}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N}),$$

$$\bar{y}_{reg} = \bar{y}_n + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})\hat{\boldsymbol{\beta}}_1$$

and $\hat{\boldsymbol{\beta}}_1$ is defined by (2.2.10).

24. (Section 2.2.1) Prove result (2.2.13).

25. (Section 2.1) The variance of the approximate distribution of the separate ratio estimator given in (2.1.12) is

$$V\{\bar{y}_{st,s} \mid \mathcal{F}\} = \sum_{h=1}^{H} W_h^2(1-f_h)n_h^{-1}(S_{y,h}^2 - 2R_hS_{xy,h} + R_h^2S_{x,h}^2)$$

and the corresponding variance of the combined ratio estimator is

$$V\{\bar{y}_{st,c} \mid \mathcal{F}\} = \sum_{h=1}^{H} W_h^2(1 - f_h)n_h^{-1}(S_{y,h}^2 - 2RS_{xy,h} + R^2S_{xy,h}),$$

where $f_h = N_h^{-1}n_h$, $R_h = \bar{y}_{h,N}\bar{x}_{h,N}^{-1}$, and $R = \bar{y}_N\bar{x}_N^{-1}$. Let $H = 10$, $(\bar{x}_{N,h}, R_h, W_h) = (1, R, 0.1)$ for all h, and $(S_{y,h}^2, S_{xy,h}, S_{x,h}^2)n_h^{-1} = (1, 0.7, 1)$ for all h. Assume that the finite population correction can be ignored. Give a value for R such that the squared bias of the separate ratio estimator based on (2.1.7) equals the variance of the separate ratio estimator. What is the bias of the combined ratio estimator for that R?

2.12 APPENDIX 2A: MISSOURI DATA

Table 2.8: **NRI Data from Missouri**

Stratum	Obs. Number	Weight	Segment Size	Cult. Crop	Forest	Federal
1	1	33	165	55	110	0
	2	33	162	54	0	0
	3	33	162	108	0	0
	4	33	168	56	56	0
	5	33	168	56	0	0
	6	33	162	54	0	0
	7	33	162	54	108	0
	8	33	162	54	0	108
	9	33	162	0	0	0
	10	33	162	54	54	0
	11	33	234	234	0	0
	12	33	162	108	0	0
	13	33	165	0	110	0
	14	33	162	108	54	0
	15	33	339	0	0	0
	16	33	162	54	0	0
	17	33	162	54	0	0
	18	33	162	162	0	0
	19	33	204	0	68	68
	20	33	165	0	165	0
	21	33	162	0	0	162
	22	33	162	54	108	0
	23	33	147	49	98	0
	24	33	162	0	162	0
	25	33	165	110	0	0
	26	33	162	0	0	108
	27	33	165	110	0	0
	28	33	162	54	0	0
	29	33	324	108	108	0
	30	33	168	0	168	0
2	31	35	159	0	53	0
	32	35	171	57	0	0
	33	35	159	0	53	0
	34	35	159	106	53	0

Continued

Stratum	Obs. Number	Weight	Segment Size	Cult. Crop	Forest	Federal
	35	35	165	0	0	0
	36	35	165	0	0	0
	37	35	165	0	0	0
	38	35	159	106	0	0
	39	35	159	0	0	0
	40	35	345	0	230	0
	41	35	165	55	0	0
	42	35	156	52	0	0
	43	35	168	112	56	0
	44	35	168	56	0	0
	45	35	309	309	0	0
	46	35	159	159	0	0
	47	35	159	159	0	0
	48	35	333	0	0	222
	49	35	162	108	0	0
	50	35	162	108	0	0
	51	35	165	55	0	0
	52	35	159	159	0	0
	53	35	165	165	0	0
	54	35	159	0	0	159
	55	35	165	0	0	0
	56	35	159	0	0	106
	57	35	159	0	0	0
	58	35	162	0	0	108
	59	35	165	110	0	0
	60	35	153	102	0	0
	61	35	162	162	0	0
	62	35	159	53	0	0
	63	35	162	108	0	0
3	64	26	162	54	0	0
	65	26	162	54	0	0
	66	26	162	0	108	0
	67	26	165	55	0	0
	68	26	168	56	56	0
	69	26	162	108	0	0
	70	26	162	54	108	0
	71	26	162	0	0	0
	72	26	162	108	0	0
	73	26	159	0	0	0

Continued

Stratum	Obs. Number	Weight	Segment Size	Cult. Crop	Forest	Federal
	74	26	162	0	108	0
	75	26	165	55	0	0
	76	26	165	0	0	0
	77	26	156	52	0	0
	78	26	165	0	55	110
	79	26	100	0	50	50
	80	26	177	0	118	0

CHAPTER 3

USE OF AUXILIARY INFORMATION IN DESIGN

3.1 INTRODUCTION

We have studied estimation procedures before discussing design because it is necessary to specify the estimation method in order to evaluate the strategy composed of a design–estimator pair. As we discussed in Section 1.1, there are many characteristics of interest in the typical survey. Also, there is generally considerable information about the population that can be used at the design stage. In our discussion of estimation, such things as the sampling frame, stratum boundaries, and cluster sizes were treated as fixed. These are some of the properties of the design that are determined at the design specification stage.

Designing a survey is a particularly challenging activity. Because there are many possible characteristics of interest, a realistic objective function will be multiple valued. Furthermore, few analysts can specify all characteristics of interest at the design stage. Even a carefully specified list of objectives will undergo changes as a survey progresses. This is completely natural and will

Sampling Statistics. By Wayne A. Fuller

lead the wise person to build a survey design capable of meeting unanticipated requirements. One should remember that the reason for conducting a survey is to "find out things." Very often, the most interesting results are only mildly associated with the primary design variables.

The determination of sample size and design depends on objectives and available resources. Designing a sample is an iterative process where the initial questions are often of the form, "I want to study banks; how big a sample do I need?" or "I have $200,000. Can I find out something useful about the debt status of families in Iowa?" Only after extended discussions will the objectives be refined enough to be useful for sample design. Similarly, considerable effort will be required to determine the possible methods of data collection, the restrictions on data collecting, and the sources of information that can be used in design. Finally, it is possible that it will be determined that available resources are not sufficient to support a study that can hope to meet the objectives.

In this chapter we study some simplified design problems. The simplest problem specifies a single y-characteristic, a class of designs, a cost function, and an estimator. The class of designs is then searched for the design that minimizes variance, or mean square error, for a given cost. The determination of the best design typically requires specification of the characteristics of the population, which are unknown. Thus, one must specify a model, complete with parameters, to solve the design problem. This model, introduced in Section 1.2, is called the *design model*. The variance of the estimator calculated under the design model is called the *anticipated variance*.

It is difficult to include in a formal cost function the attractiveness of the design to the user. Designs that produce samples that can be understood intuitively to "represent" the study population and that are relatively simple are preferred. Also important is the ease with which the design can be explained and defended given the possibility of a user unhappy with the quantitative results.

There are almost always constraints on data collection. Certain times of the day and certain days will generally be excluded from personal data collection. There is always a conflict between the desire for more data and what can reasonably be requested of a personal respondent. Although such considerations can theoretically be included in a cost function, they are generally treated as side constraints on the designs.

There is a component of the estimation problem that is often given little discussion and is sometimes neglected entirely. This is the requirement of variance estimation. If the properties of the estimator are to be conveyed correctly to the user, it is necessary to estimate the distributional properties.

This means that the design should be such that it is possible to construct good estimates of the variance of the estimator.

We restrict consideration to probability designs in which the probability of selection is known at the time of selection, or can be calculated after selection for items selected in the sample. We restrict the probabilities to be positive for all elements of the population. Generally, we consider estimators that are design consistent.

3.1.1 Selection probabilities

We have given some design results in Theorems 1.2.3 and 1.2.6. As an extension of those results, consider the problem of determining the optimal selection probabilities when the design model is the regression model. Assume that

$$\begin{aligned} y_i &= \mathbf{x}_i\boldsymbol{\beta} + e_i, \\ e_i &\sim ind(0,\ \gamma_{ii}\sigma^2), \end{aligned} \tag{3.1.1}$$

where e_i is independent of $\mathbf{x}_j$ for all i and j. Assume that the finite population is a realized sample of size N from (3.1.1), where $\mathbf{x}_i$ has finite fourth moments. The finite population mean of $\mathbf{x}$, denoted by $\bar{\mathbf{x}}_N$, and the γ_{ii}, $i = 1, 2, \ldots, N$, are known.

Given a sample of n observations satisfying (3.1.1), conditional on $\mathbf{X}$, the best linear unbiased estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{D}_\gamma^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\gamma^{-1}\mathbf{y},$$

where $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')'$, $\mathbf{D}_\gamma = \text{diag}(\gamma_{11}, \gamma_{22}, \ldots, \gamma_{nn})$, and $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$. Similarly, the best predictor of $\bar{y}_N$ is

$$\hat{\theta} = N^{-1}\left(\sum_{i\in A} y_i + (N-n)\bar{\mathbf{x}}_{N-n}\hat{\boldsymbol{\beta}}\right), \tag{3.1.2}$$

where $\bar{\mathbf{x}}_{N-n} = (N-n)^{-1}(N\bar{\mathbf{x}}_N - \sum_{i\in A}\mathbf{x}_i)$. See (2.3.6). We give a result, due to Isaki and Fuller (1982), that defines the large-sample best strategy as a combination of predictor and selection probabilities that minimize the anticipated variance for the design model (3.1.1). Recall from Section 1.2.3 that the anticipated variance of an estimator $\hat{\theta}$ as an estimator of $\bar{y}_N$ is

$$\begin{aligned} AV\{\hat{\theta} - \bar{y}_N\} &= E\{E[(\hat{\theta} - \bar{y}_N)^2 \mid \mathcal{F}_N)]\} \\ &\quad - [E\{E[(\hat{\theta} - \bar{y}_N) \mid \mathcal{F}_N]\}]^2. \end{aligned}$$

If the superpopulation has moments and the inclusion indicator variables are independent of y, we may reverse the order of expectations, taking the

expectation of the conditional expectation, conditional on the sample inclusion indicators and the population matrix of explanatory variables, to obtain

$$\begin{aligned} AV\{\hat{\theta} - \bar{y}_N\} &= E\{E[(\hat{\theta} - \bar{y}_N)^2 \mid (\mathbf{d}', \mathbf{X}_N)]\} \\ &\quad - [E\{E[(\hat{\theta} - \bar{y}_N) \mid (\mathbf{d}', \mathbf{X}_N)]\}]^2, \end{aligned} \tag{3.1.3}$$

where $\mathbf{X}_N$ is the $N \times k$ population matrix of auxiliary variables and $\mathbf{d} = (I_1, I_2, \ldots, I_N)$, as defined in Section 1.2.1. The conditional expectation conditional on $(\mathbf{d}', \mathbf{X}_N)$ is sometimes called the *model expectation*. The matrix $\mathbf{X}_N$ is fixed in Theorem 3.1.1, but we retain the notation of (3.1.3).

Theorem 3.1.1. Let $\{\mathbf{x}_i, \gamma_{ii}\}$ be a sequence of fixed vectors. Let $\mathcal{F}_N = [(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_N, \mathbf{x}_N)]$, $N > k + 1$, where the y_i are realizations from the model (3.1.1). Let $\mathcal{P}_c$ be the class of fixed-sample-size nonreplacement designs with fixed probabilities admitting design-consistent estimators of the population mean. For designs in $\mathcal{P}_c$, assume that

$$\lim_{N\to\infty} (\bar{y}_N, \bar{\mathbf{x}}_N) = (\mu_y, \boldsymbol{\mu}_x) \quad \text{a.s.}, \tag{3.1.4}$$

$$\lim_{N\to\infty} \mathbf{S}_{zz,N} = \boldsymbol{\Sigma}_{zz} \quad \text{a.s.}, \tag{3.1.5}$$

$$\lim_{N\to\infty} nE\{\bar{\mathbf{x}}_N(\mathbf{X}'\mathbf{D}_\gamma^{-1}\mathbf{X})^{-1}\bar{\mathbf{x}}_N\} = \boldsymbol{\mu}_x\mathbf{H}^{-1}\boldsymbol{\mu}_x', \tag{3.1.6}$$

and

$$\lim_{N\to\infty} n^{-1}\mathbf{X}_N'\mathbf{D}_{\pi,N}\mathbf{D}_{\gamma,N}^{-1}\mathbf{X}_N =: \lim_{N\to\infty} \mathbf{H}_N = \mathbf{H}, \tag{3.1.7}$$

where $\mathbf{z}_i = (y_i, \mathbf{x}_i)$, $\mathbf{X}_N' = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_N')$, $\mathbf{D}_{\pi,N} = \text{diag}(\pi_{1,N}, \pi_{2,N}, \ldots, \pi_{N,N})$, $\pi_{i,N}$ is the inclusion probability for element i in population $\mathcal{F}_N$, $\mathbf{D}_{\gamma,N} = \text{diag}(\gamma_{11}, \gamma_{22}, \ldots, \gamma_{NN})$, and

$$\mathbf{S}_{zz,N} = (N-1)^{-1}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}}_N)'(\mathbf{z}_i - \bar{\mathbf{z}}_N). \tag{3.1.8}$$

Assume that

$$0 < K_s < n\left(\sum_{j=1}^{N}\gamma_{jj}^{0.5}\right)^{-1}\gamma_{ii}^{0.5} < K_m < 1 \tag{3.1.9}$$

for all i. Let $\mathcal{D}_\ell$ be the class of linear-in-y predictors of the form

$$\psi_\ell = \sum_{i\in A}\alpha_{i,A}y_i, \tag{3.1.10}$$

satisfying

$$E\{(\psi_\ell - \bar{y}_N) \mid (\mathbf{d}', \mathbf{X}_N)\} = 0, \tag{3.1.11}$$

where $\alpha_{i,A}$ are permitted to be functions of sample characteristics other than y. Let there be $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ such that $\mathbf{x}_i\boldsymbol{\tau}_1 = \gamma_{ii}^{0.5}$ and $\mathbf{x}_i\boldsymbol{\tau}_2 = \gamma_{ii}$ for all i. Let $\hat{\theta}$ be the predictor (3.1.2), and let the inclusion probabilities for the design be proportional to $\gamma_{ii}^{0.5}$. Then

$$\lim_{N\to\infty} nAV\{\hat{\theta}-\bar{y}_N\} = \lim_{N\to\infty} \left((N^{-1}\sum_{i=1}^{N}\gamma_{ii}^{0.5})^2 - N^{-2}\sum_{i=1}^{N}\gamma_{ii} \right) \sigma^2 \quad (3.1.12)$$

and

$$\lim_{N\to\infty} n[AV\{\hat{\theta}-\bar{y}_N\} - AV\{\psi_\ell - \bar{y}_N\}] \leq 0 \quad (3.1.13)$$

for all $\psi_\ell \in \mathcal{D}_\ell$ and all $p \in \mathcal{P}_c$.

Proof. The predictor (3.1.2) has the minimum conditional model variance, conditional on $(\mathbf{d}', \mathbf{X}_N)$, for predictors in the class of linear conditionally unbiased predictors for model (3.1.1). The predictor can be written as $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ because $\gamma_{ii} = \mathbf{x}_i\boldsymbol{\tau}_2$. See (ii) of Theorem 2.3.1. Thus, we can write

$$\hat{\theta} = \sum_{i\in A} \alpha_{i,A} y_i =: \boldsymbol{\alpha}\mathbf{y},$$

where

$$\boldsymbol{\alpha} = \bar{\mathbf{x}}_N(\mathbf{X}'\mathbf{D}_\gamma^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\gamma^{-1}.$$

The conditional expectation $E[(\hat{\theta}-\hat{y}_N) \mid (\mathbf{d}', \mathbf{X}_N)] = 0$ because the predictor (3.1.2) is conditionally unbiased. Therefore, the anticipated variance is given by the first term on the right of the equality of (3.1.3). The conditional variance conditional on $(\mathbf{d}', \mathbf{X}_N)$, is

$$V\{\hat{\theta} - \bar{y}_N \mid (\mathbf{d}', \mathbf{X}_N)\} = (\boldsymbol{\alpha}\mathbf{D}_\gamma\boldsymbol{\alpha}' - 2N^{-1}\boldsymbol{\alpha}\mathbf{D}_\gamma\mathbf{J}_n + N^{-2}\mathbf{J}_N'\mathbf{D}_{\gamma,N}\mathbf{J}_N)\sigma^2,$$

where $\mathbf{J}_N$ is an N-dimensional column vector of 1's and $\mathbf{J}_n$ is an n-dimensional column vector of 1's. Now

$$\boldsymbol{\alpha}\mathbf{D}_\gamma\mathbf{J}_n = \boldsymbol{\alpha}\mathbf{X}\boldsymbol{\tau}_2 = \bar{\mathbf{x}}_N\boldsymbol{\tau}_2 = N^{-1}\mathbf{J}_N'\mathbf{D}_{\gamma,N}\mathbf{J}_N$$

and

$$V\{\hat{\theta} - \bar{y}_N \mid (\mathbf{d}', \mathbf{X}_N)\} = (\boldsymbol{\alpha}\mathbf{D}_\gamma\boldsymbol{\alpha}' - N^{-2}\mathbf{J}_N'\mathbf{D}_{\gamma,N}\mathbf{J}_N)\sigma^2. \quad (3.1.14)$$

Because $\mathbf{J}_N'\mathbf{D}_{\gamma,N}\mathbf{J}_N$ is a population quantity, we need only consider $E\{\boldsymbol{\alpha}\mathbf{D}_\gamma\boldsymbol{\alpha}'\}$ in determining the anticipated variance.

To evaluate $E\{\boldsymbol{\alpha}\mathbf{D}_\gamma\boldsymbol{\alpha}'\}$, note that

$$\lim_{N\to\infty} nE\{\boldsymbol{\alpha}\mathbf{D}_\gamma\boldsymbol{\alpha}'\} = \lim_{N\to\infty} nE\{\bar{\mathbf{x}}_N(\mathbf{X}'\mathbf{D}_\gamma^{-1}\mathbf{X})^{-1}\bar{\mathbf{x}}_N'\}$$

with limit given in (3.1.6). With no loss of generality, let $\gamma_{ii}^{0.5}$ be the first element of $\mathbf{x}_i$, and let $\mathbf{T}$ denote the upper triangular Gramm–Schmidt matrix with 1's on the diagonal such that $\mathbf{T}'\mathbf{H}\mathbf{T}$ is a diagonal matrix. Note that if the first element of $\mathbf{x}_i$ is $\gamma_{ii}^{0.5}$, the 1–1 element of $\mathbf{H}$ is

$$h_{11} = \lim_{N\to\infty} n^{-1} \sum_{i=1}^{N} \gamma_{ii}^{0.5} \pi_{i,N} \gamma_{ii}^{-1} \gamma_{ii}^{0.5} = 1.$$

Then, by (3.1.7),

$$\begin{aligned} \boldsymbol{\mu}_x \mathbf{H}^{-1} \boldsymbol{\mu}_x' &= \boldsymbol{\mu}_x \mathbf{T} \mathbf{T}^{-1} \mathbf{H}^{-1} \mathbf{T}^{-1\prime} \mathbf{T}' \boldsymbol{\mu}_x' \\ &= \mu_{x1}^2 + \sum_{j=2}^{k} \mu_{rj}^2 \left[\lim_{N\to\infty} n^{-1} \mathbf{r}_{\cdot j}' \mathbf{D}_{\pi,N} \mathbf{D}_{\gamma,N}^{-1} \mathbf{r}_{\cdot j} \right]^{-1}, \end{aligned} \tag{3.1.15}$$

where $\mathbf{r}_{\cdot j}$ is the jth column of $\mathbf{R} = (\mathbf{r}_1', \mathbf{r}_2', \ldots, \mathbf{r}_N')'$, $\mathbf{r}_i = \mathbf{x}_i \mathbf{T}$, and $\boldsymbol{\mu}_r = \boldsymbol{\mu}_x \mathbf{T}$. The second element of $\mathbf{r}_i$ is

$$r_{2,i} = x_{2,i} - \delta_{2,1} \gamma_{ii}^{0.5},$$

where

$$\begin{aligned} \delta_{2,1} &= \lim_{N\to\infty} \left(\sum_{i=1}^{N} \pi_{i,N} \gamma_{ii}^{-1} \gamma_{ii} \right)^{-1} \sum_{i=1}^{N} \pi_{i,N} \gamma_{ii}^{-1} \gamma_{ii}^{0.5} x_{2,i} \\ &= \lim_{N\to\infty} n^{-1} \sum_{i=1}^{N} \pi_{i,N} \gamma_{ii}^{-0.5} x_{2,i}. \end{aligned} \tag{3.1.16}$$

From (3.1.16), the mean μ_{r2} will equal zero if $\{\mathbf{x}_i\}$ satisfies (3.1.4) and (3.1.5), and

$$\pi_{i,N} = n \left(\sum_{j=1}^{N} \gamma_{jj}^{0.5} \right)^{-1} \gamma_{ii}^{0.5}. \tag{3.1.17}$$

The third element of $\mathbf{r}_i$ is

$$r_{3,i} = x_{3,i} - \delta_{3,1} \gamma_{ii}^{0.5} - \delta_{3,2} (x_{2,i} - \delta_{2,1} \gamma_{ii}^{0.5}),$$

where

$$\delta_{3,1} = \lim_{N\to\infty} n^{-1} \sum_{i=1}^{N} \pi_{i,N} \gamma_{ii}^{-0.5} x_{3,i}$$

and

$$\delta_{3,2} = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} \pi_{i,N}\gamma_{ii}^{-1}(x_{2,i} - \delta_{2,1}\gamma_{ii}^{0.5})x_{3,i}}{\sum_{i=1}^{N} \pi_{i,N}\gamma_{ii}^{-1}(x_{2,i} - \delta_{2,1}\gamma_{ii}^{0.5})^2}.$$

It follows that $\mu_{r3} = 0$ if (3.1.17) holds. A similar argument applies for the remaining elements of $\mathbf{r}_i$. Therefore, (3.1.15) attains the minimum value of

$$\mu_{x1}^2 = \left(N^{-1}\sum_{i=1}^{N}\gamma_{ii}^{0.5}\right)^2$$

for the $\pi_{i,N}$ of (3.1.17). The minimum holds for all designs satisfying (3.1.4) and (3.1.5) and hence for designs admitting design-consistent estimators with probabilities (3.1.17).

Given (3.1.4), (3.1.5), and probabilities satisfying (3.1.9), it can be shown that there exists a sequence of designs with inclusion probabilities (3.1.17) such that

$$E\{(\bar{y}_{HT} - \bar{y}_N)^2 \mid \mathcal{F}\} = O(n^{-1}).$$

See Lemma 2 of Isaki and Fuller (1982). Therefore, result (3.1.13) is established. Substituting μ_{x1}^2 into (3.1.14), we obtain result (3.1.12). ■

The important result of Theorem 3.1.1 is that selection probabilities should be proportional to the square root of the design variances. It is the variances, not the values of the design variables, that are of primary importance in determining optimal selection probabilities, given that the design variables can be used in a regression estimator.

For the regression estimator (3.1.2) to be design consistent, $\gamma_{ii}^{0.5}$ must be a linear combination of the elements of $\mathbf{x}_i$. In the theorem it is also assumed that γ_{ii} is expressible as $\gamma_{ii} = \mathbf{x}_i\boldsymbol{\tau}_2$.

To extend the discussion of optimal selection probabilities to the situation with unequal costs, assume that it costs c_i to observe element i and that a total of C is available for the survey. We desire selection probabilities that minimize the large-sample anticipated variance of the regression estimator. Assume that the model (3.1.1) is a full model as defined in Section 2.3. We write a regression estimator of the mean as

$$\bar{y}_{reg} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}, \tag{3.1.18}$$

and the estimator of the total as

$$\hat{T}_{y,reg} = N\bar{y}_{reg},$$

where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\sigma_i^{-2}\right)^{-1}\sum_{i\in A}\mathbf{x}_i'y_i\sigma_i^{-2}$$

and $\sigma_i^2 = \gamma_{ii}\sigma^2$. Then the approximate anticipated variance of the estimated total is $V\{\hat{T}_{y,reg}\} = V\{\hat{T}_e\}$ and

$$AV\{\hat{T}_e - T_e\} = \sum_{i\in U}\sigma_i^2\pi_i^{-1}(1-\pi_i), \tag{3.1.19}$$

where π_i is the selection probability. See (1.2.41) and note that $E\{e_i\} = 0$. The Lagrangian formed to minimize the approximate anticipated variance subject to the constraint that the expected cost is C is

$$\sum_{i\in U}\sigma_i^2\pi_i^{-1}(1-\pi_i) + \lambda\left(\sum_{i\in U}\pi_i c_i - C\right). \tag{3.1.20}$$

The optimal selection probabilities are

$$\pi_i = \lambda^{-1/2}c_i^{-1/2}\sigma_i,$$

where λ is the Lagrangian multiplier and

$$\lambda^{1/2} = C^{-1}\sum_{i\in U}c_i^{-1/2}\sigma_i.$$

It is possible for some π_i to exceed 1. If so, one sets those π_i equal to 1, reduces the cost accordingly, and solves the reduced problem. If the cost per element is a constant, c, the minimum value for the variance of $\hat{T}_e$ is

$$AV_{min}\{\hat{T}_e - T_e\} = Cc^{-1}\left(\sum_{i\in U}\sigma_i\right)^2 - \sum_{i\in U}\sigma_i^2. \tag{3.1.21}$$

Expression (3.1.21) is known as the *Godambe–Joshi lower bound.* See Godambe and Joshi (1965).

In most design problems there are many variables of interest. If one is willing to specify a variance model and a maximum acceptable variance for each of K variables, mathematical programming can be used to determine optimal selection probabilities. Let the cost of observing element i be c_i. Then one chooses probabilities to minimize expected cost

$$C = \sum_{i\in U}c_i w_i^{-1}$$

subject to the variance constraints

$$\sum_{i \in U} w_i \sigma_{ki}^2 \leq V_k, \ k = 1, 2, \ldots, K,$$

where σ_{ki}^2 is the model variance for characteristic k for element i, V_k is the maximum acceptable variance for characteristic k, and $w_i = \pi_i^{-1}$.

3.1.2 Strata formation

The stratification model of Theorem 1.2.6 is a special case of model (3.1.1) in which $\mathbf{x}_i$ is an H-dimensional vector of indicator functions identifying the strata, and the σ_i^2 are constant within a stratum. If element i is in stratum h, the hth element of $\mathbf{x}_i$ is 1 and the other $H - 1$ elements of $\mathbf{x}_i$ are zero. Then the regression estimator reduces to the simple stratified estimator and the selection probabilities are those defined in (1.2.59). Note that for the n_h of (1.2.59) and constant costs, the individual selection probabilities are proportional to σ_i. For stratification, the variance expression (3.1.19) is exact, except for rounding.

Joint probabilities of selection do not appear in the anticipated variance of Theorem 3.1.1 because the anticipated variance is functionally independent of the joint probabilities under model (3.1.1). If all the x-variables are indicator variables for strata, the strata and the selection probabilities define the design. For other x-variables it is natural to impose some type of control in sample selection, stratification based on the $\mathbf{x}_i$ being the most common tool.

Stratification on the x-variables provides protection against model misspecification and will reduce the $O(n^{-2})$ term in the variance of the regression estimator if the stratum indicators are not included in the regression defining the estimator. If the true relationship is linear and stratum indicators are included in the defining regression, it is possible for the $O(n^{-2})$ term in the variance to increase because of the increase in the number of parameters estimated. See Exercise 11. Order n^{-1} gains in efficiency are possible from stratification if the true relationship between y and $\mathbf{x}$ is nonlinear.

The way that strata are formed depends on the dimension of $\mathbf{x}$, on the type of estimator that one plans to use, and on the degree to which simple estimators are desirable. Given a single x-variable and a σ_i that is monotonically related to x_i, a reasonable way to form strata is to order the data on σ_i and use the cumulative sum of the π_i to form strata. The cumulative sums can be normalized so that the total is the number of units to be selected and stratum boundaries chosen so that the sum of normalized sizes in a stratum is nearly an integer greater than 1. If there are no extreme sizes, one can choose stratum boundaries so that the sum of σ_i is approximately the same in each

stratum. Very often, one will then choose to select elements within the strata by simple random sampling. Typically, there is only a very small increase in the anticipated variance from using a single stratum rate in place of the individual σ_i as selection probabilities. See Godfrey, Roshwalb, and Wright (1984) and Kott (1985).

Defining strata becomes a complex operation when there are many variables of interest and many variables that can be used to form strata. If $\mathbf{x}$ is of small dimension, a simple procedure for forming strata is to split the population into two (or more) groups on the basis of x_1, then split each of those groups into two (or more) groups on the basis of x_2, and so on, until the desired number of strata are formed. Because this procedure is realistic only for a small number of x-variables, the set of variables used in design is often restricted to a small number deemed most important. The number of design variables can also be reduced by using the first few principal components of the original set. See Pla (1991). Other methods of forming strata include cluster analysis and mathematical programming. See Hartley (1965), Jarque (1981), Golder and Yeomans (1973), and McCallion (1992).

Example 3.1.1. We consider sample design for a population of workplaces, where the data are patterned after data collected in the Canadian Workplace and Employee Survey conducted by Statistics Canada. The data represent a population of workplaces in the retail sector in a province. The data are not those collected by Statistics Canada, but display characteristics similar to the original data. Two characteristics are available for each workplace: payroll, denoted by z_2, and total employment, denoted by z_1. These data are for time t and our task is to design a sample for time $t+1$. There are 2029 workplaces on the list. Payroll and employment are highly correlated, and if we regress log payroll on log employment, the residuals appear to have nearly constant variance.

Assume that resources are available for a study composed of 100 workplaces and assume that the cost of observing a workplace is the same for all workplaces. Assume that our objective is to estimate total payroll for next year. The two components that are primary determinants of the efficiency of a design are the probabilities of selection and the control variables, where control is most often exercised through stratification.

Our design model for the population is

$$z_{2,t+1,i} = \beta_0 + z_{2,t,i}\beta_1 + e_{2,t+1,i}, \tag{3.1.22}$$
$$e_{2,t+1,i} \sim ind(0, z_{2,t,i}^{2\gamma}\sigma_{e2}^2),$$

where $z_{2,t,i}$ is the payroll of workplace i in year t in thousands of dollars and γ is a parameter. By the results of Section 3.1.1, the optimal selection probabilities should be proportional to the square roots of the variances and

hence to $z_{2,t,i}^{\gamma}$. A variable such as $z_{2,t,i}^{\gamma}$ is often called the *size* of element t. One model used for economic variables assumes $\gamma = 1$.

If model (3.1.22) holds exactly with $\gamma = 1$, the strategy composed of a design with probabilities proportional to $z_{2,t,i}$ and the regression estimator approximately attains the Godambe–Joshi lower bound. In a large-scale survey with many characteristics, one is guaranteed that the design model will not hold for all characteristics. Therefore, it is wise to stratify to protect against model failure and to provide the potential for gains in efficiency. Our method of forming strata is to order the population on size, cumulate the sizes, and form stratum boundaries using the cumulated sizes. Because the mean and variance are monotonically related, the strata will furnish control on both. We require a sample size of at least two in each stratum to permit construction of a design-unbiased estimator of variance. Also, the two-per-stratum design will give good efficiency if there is a nonlinear relationship between the y-variable and size. One can select an unequal probability sample in each stratum, but we will see that little efficiency is lost by selecting simple random samples within strata.

The sum of the $z_{2,t,i}$ for the 2029 workplaces is 320,117, and with $\gamma = 1$, the probabilities proposed for a sample of 100 are

$$\tilde{\pi}_i = 100(320,117)^{-1} z_{2,t,i}.$$

Therefore, any workplace with a size greater than 3201 should be included in the sample with certainty. The nine largest workplaces have a size greater than 3201, the sum of the sizes for these nine is 50,000, and the sum of the sizes for the 2020 remaining workplaces is 270,117. Because 91 units remain to be selected, any unit with a size greater than 2968 should be included with certainty. The tenth and eleventh sizes are 3000 and 2800. The size of the tenth workplace exceeds the bound, and the size of the eleventh is 94% of the bound. Therefore, we add these two workplaces to the certainty class. This type of very skewed population where the size measure leads to a sample with some certainty units is common for economic variables.

To form the remaining 2018 elements, with a total size of 264,317, into strata for the remaining 89 sample units, we form cumulative sums and divide the sums by 2969.85 so that the total is 89. We form strata for the largest units to obtain good approximations to the desired probabilities by keeping the population number in the stratum small and the normalized sum of sizes for the stratum close to an integer. We place the five largest of the remaining workplaces in a stratum, from which four will be selected. The sum of the optimal selection probabilities is 4.16 for this stratum. Four units are placed in the next stratum with a sum of optimal probabilities equal to 2.96, from which three are to be selected.

The remaining workplaces are placed in 41 strata of approximately equal size, where two units are to be selected from each stratum. To form the 41 strata, an iterative procedure is used. The cumulative sum of the ordered sizes is formed, where the order is from largest to smallest. The cumulative sums are normalized so that the final sum is 82. Let k be the index j such that $\mid C_{(j)} - 2 \mid$ is a minimum, where $C_{(j)}$ is the cumulative sum associated with the jth element in the ordered set. Then element k and all elements with smaller order indices form stratum 1. The elements in stratum 1 are removed and cumulative sums of sizes formed for the remaining elements. These new cumulative sums are normalized so that the final sum is 80 and the $\mid C_{(j)} - 2 \mid$ criterion is applied to the new sums to form the second stratum. The elements for the second stratum are removed, the cumulative sums for the remaining elements formed and normalized so that the final sum is 78, and so on. With this procedure of forming strata, strata 3, 4, and 5 have 3, 3, and 4 population elements, respectively. The sum of the sizes for these three strata are 1.97, 1.79, and 2.24, respectively. The last three strata contain 129, 159, and 299 workplaces, with sizes of 1.99, 1.99, and 1.98, respectively.

To calculate the anticipated variance for the design, we assume that β_1 of (3.1.22) is equal to 1. Then the anticipated variance of the stratified mean under the design model is

$$AV\{\bar{z}_{2,t+1,st}\} = \sum_{h=1}^{43}(1 - f_h)W_h^2 n_h^{-1}(S_{1h}^2 + S_{2h}^2), \tag{3.1.23}$$

where $f_h = N_h^{-1}n_h$, $W_h = N^{-1}N_h$,

$$S_{1h}^2 = N_h^{-1} \sum_{i \in U_h} z_{2,t,i}^2 \sigma_{e2}^2,$$

$$S_{2h}^2 = (N_h - 1)^{-1} \sum_{i \in U_h} (z_{2,t,i} - \bar{z}_{2,t,h,N})^2,$$

and $\bar{z}_{2,t,h,N}$ is the stratum mean for stratum h. The second part of the stratum variance, S_{2h}^2, is the contribution of the variability among the $z_{2,t,i}$ within the strata. That term is the difference between the anticipated variance of the simple stratified estimator and the large-sample anticipated variance of the regression estimator for the stratified design,

$$V\{\bar{y}_{reg,st}\} = \sum_{h=1}^{43}(1 - f_h)W_h^2 n_h^{-1} S_{1h}^2, \tag{3.1.24}$$

where

$$\bar{y}_{reg,st} = \bar{y}_{st} + (\bar{z}_{2,t,N} - \bar{z}_{2,t,st})\hat{\beta}$$

and

$$\hat{\beta} = [\hat{V}\{\bar{z}_{2,t,st}\}]^{-1}\hat{C}\{\bar{z}_{2,t,st}, \bar{z}_{2,t+1,st}\}.$$

We must approximate σ^2_{e2} to calculate the anticipated variance. In applications, there may be outside information on the relationship between variables, such as a previous census or survey, or one may be forced to use an estimate based on general experience. We have no additional information, but suppose that there is a strong correlation between $z_{2,t,i}$ and $z_{2,t+1,i}$. If we assume an R^2 of 0.80 and assume that $V\{z_{2,t}\} = V\{z_{2,t+1}\} =: \sigma^2_{z2}$, then

$$N^{-1}\sum_{i=1}^{N} z^2_{2,t,i}\sigma^2_{e2} = (1 - R^2)V\{z_{2,t+1}\}$$

and

$$\begin{aligned}\sigma^2_{e2} &= \left(N^{-1}\sum_{i=1}^{N} z^2_{2,t,i}\right)^{-1}(1 - R^2)\sigma^2_{z2} = 0.2(\sigma^2_{z2} + \bar{z}^2_{2,N})^{-1}\sigma^2_{z2} \\ &= 0.2(2.4225)^{-1}2.1736 = 0.1794,\end{aligned}$$

where $\sigma^2_{z2} = 2.1736 \times 10^5$ is the finite population variance of $z_{2,t}$ and $\bar{z}_{2,N} = 157.77$.

For the stratified design outlined, the anticipated variance (3.1.23) of the simple stratified estimator of the mean of $z_{2,t+1}$ is

$$\begin{aligned}AV\{\bar{z}_{2,t+1,st}\} &= \sum_{h=1}^{43}(1 - f_h)W^2_h n^{-1}_h S^2_{1h} + \sum_{h=1}^{43}(1 - f_h)W^2_h n^{-1}_h S^2_{2h} \\ &= 27.31 + 0.65 = 27.96. \qquad (3.1.25)\end{aligned}$$

Note the small contribution of the S^2_{2h} to the sum. Stratification removes most of the correlation between $z_{2,t,i}$ and $z_{2,t+1,i}$.

The minimum anticipated variance for a sample of 100 workplaces with 11 certainty workplaces is obtained with the optimum selection probabilities

$$\begin{aligned}\pi_{i,s} &= \left(n_s^{-1}\sum_{i\in U_s} z_{2,t,i}\right)^{-1} z_{2,t,i} \\ &= (2969.85)^{-1} z_{2,t,i},\end{aligned}$$

where $n_s = 89$ and U_s is the set of indices for the 2018 noncertainty workplaces. Thus,

$$AV_{min}\{\bar{z}_{2,t+1}\} = N^{-2}\sum_{i\in U}\pi^{-1}_{i,s}(1 - \pi_{i,s})z^2_{2,t,i}\sigma^2_{e2}$$

$$
\begin{aligned}
&= N^{-2}\left(n^{-1}(\sum_{i\in U_s} z_{2,t,i})^2 - \sum_{i\in U_s} z_{2,t,i}^2\right)\sigma_{e2}^2 \\
&= N^{-2}\left(2969.85\sum_{i\in U_s} z_{2,t,i} - \sum_{i\in U_s} z_{2,t,i}^2\right)\sigma_{e2}^2 \\
&= 27.15, \qquad (3.1.26)
\end{aligned}
$$

and the stratified design with the simple stratified estimator has an anticipated variance that is only 3% larger than the theoretical minimum. Use of the regression estimator would reduce the large-sample anticipated variance to 27.31. Hence, the theoretical loss relative to the lower bound from using stratification probabilities in place of optional probabilities is less than 1%.

Assume now that employment is also of interest and that the design model for employment is model (3.1.22) with $(z_{1,t+1,i}, z_{1,t,i})$ replacing $(z_{2,t+1,i}, z_{2,t,i})$. If one specifies that payroll and employment are "equally important," there are a number of alternative designs, depending on the definition of "equally important." If one is willing to specify maximum variances for each, mathematical programming can be used to determine a minimum cost design. We develop a design using both $z_{1,t,i}$ and $z_{2,t,i}$, but without specifying maximum variances.

There are two workplaces that have payroll less than 2800 and that have employment greater than 201.23, a size that would lead to certainty inclusion if employment was used as size to determine probabilities. Adding these two workplaces to the set of 11 workplaces with large payroll gives a certainty group of 13 workplaces. The average of the payrolls is 157.77, and the average number of employees is 9.9177. To keep the magnitudes similar, we work with $16z_{1,t,i}$. Among the possible sizes to use for selection of the remaining 87 sample elements are the weighted sum of the two sizes, the square root of the sum of the two squared sizes, and the maximum of the two sizes. See Kott and Baily (2000) for use of the maximum of the sizes. We placed the 2016 workplaces in strata on the basis of $16z_{1,t,i} + z_{2,t,i}$.

Using the ordered list, ordered on $16z_{1,t,i} + z_{2,t,i}$, workplaces 14 through 18 are placed in a stratum from which four will be selected. The next four workplaces are placed in a stratum from which three are to be selected. Using the described procedure that minimizes $\mid C_{(j)} - 2 \mid$, the remaining workplaces are placed in strata of approximately equal sum size so that two workplaces are selected in each stratum. There are 131, 156, and 302 workplaces in the last three strata.

To complete the set of design assumptions, we assume that the correlation between $z_{1,t+1,i}$ and $z_{1,t,i}$ is 0.80, which gives $\sigma_{e1}^2 = 0.1747$ for $16z_{1t}$. The

anticipated variances for the stratified means of $16z_{1,t+1}$ and $z_{2,t+1}$ are

$$AV\{16\bar{z}_{1,t+1,st}\} = 30.65 + 4.27 = 34.92$$

and

$$AV\{\bar{z}_{2,t+1,st}\} = 28.45 + 4.19 = 32.64,$$

where the two parts of each anticipated variance correspond to the two parts of (3.1.25). The anticipated variance for $16\bar{z}_{1,t+1,st}$ under the design for $\bar{z}_{2,t+1,st}$ is $35.23 + 23.99 = 59.22$. Thus, moving to the second design decreases the variance of $\bar{z}_{1,t+1,st}$ by 41% and increases the variance of $\bar{z}_{2,t+1,st}$ by 17%. If one uses the regression estimator, the decrease for $\bar{z}_{1,t+1,st}$ is 13% and the increase for $\bar{z}_{2,t+1,st}$ is 4%.

Consider now an alternative objective for the survey design. Assume that one is interested in estimating the regression of log payroll on log employment. In particular, we are interested in estimating β_1 of the regression model

$$\begin{aligned} y_{t+1,i} &= \beta_0 + (x_{t+1,i} - \mu_{x,t+1})\beta_1 + e_{t+1,i}, \\ e_{t+1,i} &\sim iid(0, \sigma^2), \end{aligned} \tag{3.1.27}$$

where $y_{t+1,i} = \log z_{2,t+1,i}$, $x_{t+1,i} = \log z_{1,t+1,i}$, and $e_{t+1,i}$, is assumed to be independent of $x_{1,t+1,j}$ for all i and j. Because we are interested in β_1 as a superpopulation parameter, we use notation appropriate for an infinite superpopulation. For design purposes, we assume that the estimator of (β_0, β_1) will be

$$(\hat{\beta}_{0\pi}, \hat{\beta}_{1\pi})' = (\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{y},$$

where the ith row of $\mathbf{X}$ is $(1, x_{t+1,i} - \bar{x}_{t+1,N})$. Thus,

$$\hat{\beta}_{1\pi} = \frac{\sum_{i\in A}\pi_i^{-1}(x_{t+1,i} - \bar{x}_{t+1,\pi})(y_{t+1,i} - \bar{y}_{t+1,\pi})}{\sum_{i\in A}\pi_i^{-1}(x_{t+1,i} - \bar{x}_{t+1,\pi})^2} \tag{3.1.28}$$

and the design objective is to minimize the approximate variance,

$$V\{\hat{\beta}_{1\pi}\} \doteq \left(\sum_{i\in U}(x_{t+1,i} - \bar{x}_{t+1,N})^2\right)^{-2} \sum_{i\in U}\pi_i^{-1}(x_{t+1,i} - \bar{x}_{t+1,N})^2\sigma_e^2. \tag{3.1.29}$$

Absent any other information, the π_i that minimize (3.1.29) are proportional to $\mid x_{t+1,i} - \bar{x}_{t+1,N} \mid$. We do not know $x_{t+1,i} - \bar{x}_{t+1,N}$, but it is reasonable to use $x_{t,i} - \bar{x}_{t,N}$ as a proxy variable. If we do so, we must recognize that a very small value of $x_{t,i} - \bar{x}_{t,N}$ could be associated with a relatively large

$x_{t+1,i} - \bar{x}_{t+1,N}$. Therefore, we impose a lower bound on our measure of size. We define our size measure by

$$\begin{aligned} m_i &= 0.675\sigma_x && \text{if } \mid x_{t,i} - \bar{x}_{t,N} \mid < 0.675\sigma_x \\ &= \mid x_{t,i} - \bar{x}_{t,N} \mid && \text{otherwise.} \end{aligned} \tag{3.1.30}$$

This measure of size does not minimize the anticipated variance of $\hat{\beta}_{0\pi}$, but using the lower bound on the size will give us a good compromise design for both $\hat{\beta}_{0\pi}$ and $\hat{\beta}_{1\pi}$.

Next we ask if any control variables can be used to reduce the variance of the estimator of β_1. The model (3.1.27) for time t is

$$y_{t,i} = \beta_0 + (x_{t,i} - \mu_{x,t})\beta_1 + e_{t,i} \tag{3.1.31}$$

and it is reasonable to assume that $e_{t+1,i}$ and $e_{t,i}$ are correlated. That is, a firm that has a high payroll relative to employment at time t will probably have relatively a high payroll at time $t+1$. Thus, we fit the model for the time t data and use the residuals $\hat{e}_{t,i}$ for a stratification variable. If we stratify, we need to ask ourselves if it is reasonable to suppose that the conditional distribution of the errors given the stratification is also zero mean with constant variance. Proceeding as if this is the case, we desire selection probabilities close to m_i in strata with similar time t residuals. We also want the sample to be "representative" with respect to $x_{t,i}$ as well as with respect to m_i.

We could order the population on the residuals, form 50 equal-sized strata, where size is m_i of (3.1.30), and select two elements in each stratum using the Brewer scheme of Section 1.4. An alternative procedure, and the one we adopt, is to use the three variables $x_t, \hat{e}_t$, and m_i as stratification variables. Let the m_i be normalized so that the sum of the m_i is 100.

We first divide the population into two parts on the basis of m_i. There are 1363 workplaces with m_i equal to $0.675\sigma_x$. The sum of the m_i for the 2029 workplaces is 1810.79, and the sum for the 1363 workplaces is 872.03. The workplaces with $m_i = 0.675\sigma_x$ are ordered on $\hat{e}_{t,i}$ and divided into six nearly equal-sized groups. Then each of these groups is ordered on $x_{t,i}$ and divided into four nearly equal-sized groups, creating a total of 24 strata for the set of workplaces with $m_i = 0.675\sigma_x$. Each of these strata contains 56 or 57 workplaces.

The second set of workplaces is ordered on $x_{t,i}$ and divided into seven subsets. Each of the first six subsets has a sum size of approximately 8 and the last subset has a sum size of approximately 4. The subsets are ordered on $\hat{e}_{t,i}$ and subdivided. Each of the six subsets is divided into four subgroups, and the seventh is divided into two subgroups. The subgroups are formed to have approximately the same size in terms of m_i. The strata range in size from 10 to 44 workplaces.

Given these strata, the sum sizes within a stratum are fairly similar and the simple stratified sample with two per stratum will have selection probabilities that are approximately proportional to m_i.

To produce an anticipated variance, we require values for σ_e^2 and for the correlation between $e_{t+1,i}$ and $e_{t,i}$. The estimated value of σ_e^2 for the regression using time t data is 0.06155, and we use this as the anticipated value. In the absence of information on the correlation between $e_{t+1,i}$ and $e_{t,i}$, but feeling it should be fairly high, we assume a squared correlation of 0.36. Then the anticipated variance of $\hat{\beta}_1$ is given by expression (3.1.29), where $\sigma_e^2 = 0.64(0.06155) = 0.03939$ and $x_{t+1,i}$ is replaced with $x_{t,i}$. The design using m_i as the size variable is nearly five times as efficient for $\hat{\beta}_1$ as the design based on $(z_{1,t,i}, z_{2,t,i})$. Conversely, estimates of $(\bar{z}_{1,t+1,N}, \bar{z}_{2,t+1,N})$ under the design for those parameters are six to nine times more efficient than estimates from the design for the regression coefficient. See Table 3.1.

Under the design for β_1, the anticipated variance of the regression estimator for $\bar{z}_{2,t+1,N}$ using $z_{2,t,i}$ as the auxiliary variable is 121.55 and the anticipated variance of the regression estimator for $\bar{z}_{1,t+1,N}$ using $z_{1,t,i}$ as an auxiliary variable in 96.84. Regression estimation removes the portion of the variance due to within-stratum variation in mean values. There is no way to compensate for the fact that the selection probabilities have a poor correlation with the standard deviations of the $z_{1,t+1,i}$ and $y_{2,t+1,i}$.

Table 3.1 Anticipated Variances of Stratified Estimator Under Alternative Designs

Designed to Estimate:	Statistic				
	$16\bar{z}_{1,t+1,st}$	$\bar{z}_{2,t+1,st}$	Median $z_{2,t+1}$	$100\hat{\beta}_0$	$100\hat{\beta}_1$
$\bar{z}_{2,t+1,N}$	59.22	27.96	53.37	10.46	12.97
$(\bar{z}_{1,t+1,N}, \bar{z}_{2,t+1,N})$	33.06	30.69	57.63	10.61	13.19
β_1	208.33	279.47	39.70	4.67	2.68
Median $z_{2,t+1}$	1762.95	2296.06	23.86	4.53	7.12
Compromise	36.69	34.67	45.87	7.26	7.27

As a fourth objective, assume that we wish to estimate the median of z_{2t}. From the discussion in Section 1.3.5, we know that the approximate variance of the median is the variance of the cumulative distribution function (CDF) at the 50% point divided by the square of the slope of the CDF at that point. The variance of the CDF at point a_o is the variance of the mean of the indicator function

$$I_y(a_o) = 1 \qquad \text{if } y \leq a_o$$
$$= 0 \qquad \text{otherwise.}$$

By specifying a distribution for e_t in model (3.1.22), we can compute the variance of $I_y(a_o)$ for each i. The normal distribution is a natural choice but will lead to many exceedingly small probabilities. Therefore, we set a lower bound of 0.04 on the probabilities and define the design size to be $b_i = [p_i(1-p_i)]^{0.5}$, where

$$p_i = \max\{\Phi\left[(0.4243 z_{2,t,i})^{-1} \mid z_{2,t,i} - 78\ |\right], 0.04\},$$

$\Phi[\cdot]$ is the normal CDF, and 78 is the median of z_{2t}. With the lower bound of 0.04, the smallest probabilities are about 0.4 of the largest probabilities, a fairly conservative design. We order the population on z_{2ti} and form strata that are of equal size with respect to b_i. The stratum containing the smallest workplaces contains 70 workplaces, the stratum with the largest workplaces contains 70 or 71 workplaces, and the strata close to the median contain 27 or 28 workplaces.

The anticipated variance of the median under the median design is about 0.4 of the variance under the design for $(\bar{z}_{1,t+1,N}, \bar{z}_{2,t+1,N})$. However, the anticipated variance of the stratified estimator of the mean of $z_{2,t+1,N}$ under the design for the median is more than 80 times that under the design for the mean of z_2. Under the design for the median, the anticipated variances of the regression estimators of $16\bar{z}_{1,t+1,N}$ and $\bar{z}_{2,t+1,N}$ are 717.23 and 571.12, respectively. The large differences between these variances and those of the second line of Table 3.1 reflect the large differences between the set of selection probabilities that is optimum for the mean of z_2 and the set that is optimum for the median. The differences in probabilities are greatest for the large workplaces. The large workplaces are assigned very large probabilities for estimating the mean but assigned small probabilities for estimating the median.

Table 3.1 contains anticipated variances for the four designs we have outlined and for a compromise design. The size for the compromise design is $z_{2,t,i} + 16z_{1,t,i} + 64$. The constant was chosen so that the probabilities of selection for the smallest workplaces are about twice those for the design using $z_{2,t,i} + 16z_{1,t,i}$ as the size measure. Eight workplaces are included in the sample with certainty and the three strata with the smallest workplaces have 108, 128, and 155 workplaces. Increasing the smallest probabilities relative to those optimum for z_2 produce improvements for the median and regression coefficients with modest loss for the estimated totals of z_1 and z_2.

The variances for the compromise design demonstrate the nature of a compromise. The variances of estimators are never the best and never the

worst. The difference between the compromise and the specific design is greatest for the estimator of slope. This is because the design for slope has higher probabilities for small workplaces than do the other designs.

In constructing designs for specific objectives, we never attempted to achieve the absolute minimum variance. Also, by using the classic stratified design with $n_h \geq 2$, simple unbiased estimators of variance are available for the estimators. One must remember that the design model is imperfect and that the survey will be used for purposes other than those specified at the design stage. It is wise to create a design and then consider increasing the smallest selection probabilities. Models such as (3.1.22) often specify variances that are too small for small units. ■■

3.1.3 Stratification and variance estimation

A topic often overlooked in discussion of survey design is variance estimation. We give an illustration of the effect of stratification on the efficiency of variance estimation.

Example 3.1.2. Assume that the auxiliary information on a population of 1600 elements is that each element is associated with one and only one of the integers from one to 1600. Let the design model for the characteristic y be

$$y_i = \beta_0 + x_i\beta_1 + e_i, \tag{3.1.32}$$

where $e_i \sim NI(0, \sigma_e^2)$ and

$$x_i = i, \qquad i = 1, 2, \ldots, 1600.$$

Assume that a sample of size 64 is to be selected. Table 3.2 contains the variance of the stratified sample mean for different numbers of strata and different values of the correlation between y and x. We assume that the population is divided into equal-sized strata on the basis of the x values, and that the same number of elements is selected in each stratum. We ignore the finite population correction term throughout. We standardize all table entries so that the variance of the mean of a simple random sample is 100. We set $\sigma_e^2 = (1 - \rho^2)\sigma_y^2$, where ρ is the correlation between x and y, and note that $\rho^2\sigma_y^2$ is a multiple of σ_x^2.

The approximate variance of the stratified mean for equal-sized samples selected from equal-sized strata is

$$V\{\bar{y}_{st}\} = n^{-1}\sigma_{y,w}^2,$$

Table 3.2 Variance of Stratified Sample Mean Under Alternative Designs

Number of Strata	Population Squared Correlation					
	0	0.25	0.50	0.75	0.90	0.99
1	100.00	100.00	100.00	100.00	100.00	100.00
2	100.00	81.25	62.50	43.75	32.50	25.75
4	100.00	76.56	53.12	29.69	15.62	7.19
8	100.00	75.39	50.78	26.17	11.41	2.55
16	100.00	75.10	50.20	25.29	10.35	1.39
32	100.00	75.02	50.05	25.07	10.09	1.10
64	100.00	75.01	50.01	25.02	10.02	1.02

where $\sigma^2_{y,w}$ is the pooled within-stratum variance,

$$\begin{aligned}\sigma^2_{y,w} &= (N-H)^{-1}\sum_{h=1}^{H}\sum_{j=1}^{M}(y_{hj}-\bar{y}_{Mh})^2 \\ &= H^{-1}\sum_{h=1}^{H}\sigma^2_{yh},\end{aligned}$$

H is the number of strata, M is the population number of elements in each stratum, $\bar{y}_{Mh}$ is the population mean of the hth stratum, and σ^2_{yh} is the population variance of the hth stratum. In our example

$$\sigma^2_{yh} \doteq (12)^{-1}(M^2-1)\rho^2+\sigma^2_e.$$

If the variances are standardized so that the variance of the mean of a simple random sample is 100, then σ^2_{yh} is

$$\left[H^{-2}\rho^2+(1-\rho^2)\right]\times 100.$$

Two things are clear from Table 3.2. The anticipated variance of the stratified sample mean decreases monotonically as the number of strata increases for any positive ρ^2. Second, the decrease is modest from 16 to 64 strata, except for very large correlations. Also see Cochran (1977, Section 5A.8).

The estimated variance for stratified designs with equal sized strata and equal sized samples is

$$\hat{V}\{\bar{y}_{st}\} = n^{-1}(n-H)^{-1}\sum_{h=1}^{H}\sum_{j=1}^{M}(y_{hj}-\bar{y}_h)^2, \qquad (3.1.33)$$

Table 3.3 Variance of Estimated Variance of Stratified Sample Mean Under Alternative Designs

Number of Strata	Population Squared Correlation					
	0	0.25	0.50	0.75	0.90	0.99
1	100.00	96.25	85.00	66.25	51.40	41.19
2	101.61	66.84	38.74	17.31	7.65	3.00
4	105.00	61.53	29.57	9.12	2.36	0.30
8	112.50	63.94	29.00	7.70	1.45	0.06
16	131.25	74.02	33.07	8.40	1.41	0.02
32	196.87	110.81	49.31	12.38	2.00	0.02

where M is the number of sample elements in each stratum. An approximate variance of the estimated variance for our illustration population is

$$V[\hat{V}\{\bar{y}_{st}\}] \doteq \frac{\rho^4(720)^{-1}(4M^4 + 10M^2 - 5)\lambda^4 + 4\rho^2\sigma_{xh}^2\sigma_e^2 + 2\sigma_e^4}{n^2(n-H)},$$

where $\sigma_{xh}^2 = (12)^{-1}(M^2-1)\lambda^2$ is the within-stratum variance of x. Table 3.3 contains the variances of the estimated variances standardized so that the variance of the estimated variance for the simple random sample design with $\rho^2 = 0$ is 100. The distribution of x is uniform. Therefore, the variance of x^2 is about 0.8 of the square of the variance of x. Because e is normally distributed, the variance of e^2 is twice the square of the variance of e. Therefore, the variance of the estimated variance is smaller for populations with a large ρ^2.

The degrees of freedom for the variance estimator are 63, 62, 60, 56, 48, and 32 for the designs with 1, 2, 4, 8, 16, and 32 strata, respectively. If $\rho^2 = 0$, the two-per-stratum design has a variance of the estimated variance that is approximately twice that of the simple random sample. On the other hand, the variance of the estimated variance for populations with $\rho^2 > 0$ first declines and then increases as the number of strata increase. This is because, initially, the decrease in the variance is more important than the decrease in the degrees of freedom.

This example demonstrates the limited gains in efficiency that are generally obtained from heavy stratification on a single variable, and the cost in terms of variance of the estimated variance associated with heavy stratification. In practice, several variables are often available for use in stratification. The gain from using additional variables in stratification can be substantial if the additional variables have low correlation with the initial stratification variables and high correlation with some of the y-variables.

We have considered simple estimators to illustrate the relationship between efficiency in variance estimation and efficiency in mean estimation. In our illustration, given a linear relationship between the characteristic of interest and the design variable, the regression estimator could be used. The approximate variance of the variance estimator for the regression estimator, with $\mathbf{x}_i$ containing the design variable and stratum indicators, would be $2(n - H - 1)^{-1}\sigma_e^4$.

■■

A common procedure is to select one unit per stratum and to combine or "collapse" two adjacent strata to form a variance estimation stratum. The two-per-stratum variance is calculated using the collapsed strata. For an equal probability design with equal population sizes, the estimated variance for the mean of the two strata calculated from the collapsed stratum is

$$\hat{V}_{col}\{\bar{y}_{col}\} = 0.25(y_1 - y_2)^2,$$

where y_1 is the observation in original stratum 1 and y_2 is the observation in original stratum 2. Assuming large population sizes so that finite correction terms can be ignored,

$$E\{\hat{V}_{col}(\bar{y}_{col})\} = 0.25(\mu_1 - \mu_2)^2 + 0.25(\sigma_1^2 + \sigma_2^2),$$

where $\bar{y}_{col} = 0.5(y_1 + y_2)$, μ_1 and μ_2 are the stratum means in strata 1 and 2, respectively, and σ_1^2 and σ_2^2 are the two variances. If $\mu_1 = \mu_2$, the variance of the one-per-stratum design is the same as the variance of the two-per-stratum design. If $\mu_1 \neq \mu_2$, the collapsed strata variance estimator is positively biased for the variance of the one-per-stratum design. In fact, the estimator is a biased estimator of the variance of a two-per-stratum design because the variance of a two-per-stratum design is

$$V\{\bar{y}_{2,st}\} = 0.125(\mu_1 - \mu_2)^2 + 0.25(\sigma_1^2 + \sigma_2^2).$$

A person using a one-per-stratum design and the collapsed variance estimator will, on average, produce wider confidence limits for estimates than a person using a two-per-stratum design.

There is sometimes a need for an unbiased estimator of variance beyond that of a measure of reliability of estimates. Examples include small area estimation and design of future surveys. If such use is anticipated, the one-per-stratum design with the collapsed strata variance estimator is a questionable choice.

3.1.4 Controlled two-per-stratum design

Two-per-stratum designs are popular because unbiased design variance estimation is possible, it is relatively easy to implement unequal probability sampling for the designs, and the designs give good efficiency for many populations. Also, replication variance estimation, particularly balanced half-sample variance estimation, can be used. See Chapter 4. We discuss a way to impose additional control on a two-per-stratum design. See Park and Fuller (2002).

Consider selection of a two-per-stratum sample with four sets of strata. Let each stratum be divided into two equal-sized groups. The division can be on the basis of the original ordering or on the basis of a second variable. Let the group with smaller values of the division variable be called group 1 and the group with larger values be called group 2. Given the partition, there are three types of samples of size 2 in a stratum: two elements from group 1, two elements from group 2, and one element from each group. For groups of size m and ordinary two-per-stratum sampling, the probabilities for the three types of samples are $m(m-1)\left[2m(2m-1)\right]^{-1}$, $m(m-1)\left[2m(2m-1)\right]^{-1}$, and $m(2m-1)^{-1}$, respectively. As m increases, these probabilities approach 0.25, 0.25, and 0.50, respectively.

Consider a design in which for each set of four strata, we impose the restriction that one stratum has two elements in group 1, one stratum has two elements in group 2, and two strata have one element in group 1 and one element in group 2. Let there be m elements in each group of stratum $h, h = 1, 2, 3, 4$. To select the sample, one of the 24 possible arrangements is chosen at random. Then random samples of one or two elements are chosen in the designated strata. Using an analysis-of-variance decomposition, we write the characteristic for element k in group j of stratum h as

$$y_{hjk} = \bar{y}_N + \alpha_{Nh} + \beta_{Nj} + \gamma_{N,hj} + e_{hjk}, \tag{3.1.34}$$

where

$$\begin{aligned}
\alpha_{Nh} &= 0.5(\bar{y}_{N,h1} + \bar{y}_{N,h2}) - \bar{y}_N \qquad h = 1, 2, 3, 4, \\
\beta_{Nj} &= 0.25 \sum_{h=1}^{4} \bar{y}_{N,hj} - \bar{y}_N \qquad j = 1, 2, \\
\gamma_{N,hj} &= \bar{y}_{N,hj} - \alpha_{Nh} - \beta_{Nj} - \bar{y}_N, \\
\bar{y}_{N,hj} &= m^{-1} \sum_{k=1}^{m} y_{hjk}
\end{aligned}$$

and $e_{hjk} = y_{hjk} - \bar{y}_{N,hj}$. The probability that element k in group j of stratum h is selected is m^{-1}, an unbiased estimator of the finite population mean is

$$\hat{\theta} = N^{-1} \sum_{hjk \in A} m y_{hjk} = 8^{-1} \sum_{hjk \in A} y_{hjk},$$

and $\hat{\theta}$ is the simple mean.

To evaluate the variance of $\hat{\theta}$, consider the sample composed of two elements from group 1 of stratum 1, two elements from group 2 of stratum 2, one element from each group of stratum 3, and one element from each group of stratum 4. For this arrangement, the error in $\hat{\theta}$ as an estimator of $\bar{y}_N$ is

$$\begin{aligned} \hat{\theta} - \bar{y}_N &= N^{-1}[m(\gamma_{11} - \gamma_{12}) - m(\gamma_{21} - \gamma_{22}) \\ &\quad + m(e_{111} + e_{112}) + m(e_{221} + e_{222}) \\ &\quad + \sum_{h=3}^{4} \sum_{j=1}^{2} m e_{hj1}], \end{aligned} \tag{3.1.35}$$

where the k identification is chosen for convenience and we suppress the N subscript on $\gamma_{N,hj}$. It follows that

$$\begin{aligned} E\{(\hat{\theta} - \bar{y}_N)^2 | \mathcal{F}, A_{[1]}\} &= N^{-2}[m(\gamma_{11} - \gamma_{12}) - m(\gamma_{21} - \gamma_{22})]^2 \\ &\quad + N^{-2}[2m^2(1 - 2m^{-1})S_{11}^2 \\ &\quad + 2m^2(1 - 2m^{-1})S_{22}^2 \\ &\quad + m^2(1 - m^{-1})(S_{31}^2 + S_{32}^2) \\ &\quad + m^2(1 - m^{-1})(S_{41}^2 + S_{42}^2)], \end{aligned} \tag{3.1.36}$$

where $A_{[1]}$ denotes the described sample arrangement and

$$S_{hj}^2 = (m-1)^{-1} \sum_{k=1}^{m} (y_{hjk} - \bar{y}_{N,hj})^2.$$

The estimator $\hat{\theta}$ is unbiased under the design, and the variance of $\hat{\theta}$ is

$$\begin{aligned} &E[E\{(\hat{\theta} - \bar{y}_N)^2 | \mathcal{F}, A_{[r]}\} \mid \mathcal{F}] \\ &= (64)^{-1} \left((4/3) \sum_{h=1}^{4} \sum_{j=1}^{2} \gamma_{hj}^2 + \sum_{r=1}^{4} \sum_{j=1}^{2} (1 - 1.5m^{-1}) S_{rj}^2 \right). \end{aligned} \tag{3.1.37}$$

In expression (3.1.37) the $\gamma_{N,hj}$, defined in (3.1.34), are treated as fixed quantities.

To compute an anticipated variance, assume that

$$y_{ijk} = \mu + \alpha_h + \zeta_j + b_{hj} + e_{hjk}, \tag{3.1.38}$$

where α_h are fixed stratum effects, ζ_j is the fixed group effect with $\zeta_1 = -\zeta_2$, the e_{hjk} are $iid(0, \sigma_e^2)$ random variables, the b_{hj} are $iid(0, \sigma_b^2)$ random variables, and the e_{hjk} are independent of b_{st} for all hjk and st. Then

$$E\{V(\hat{\theta} - \bar{y}_N \mid \mathcal{F})\} = (16)^{-1}\sigma_b^2 + 8^{-1}(1 - m^{-1})\sigma_e^2. \tag{3.1.39}$$

The analogous anticipated variance of the estimated mean for the two-per-stratum design is

$$E\{V(\hat{\theta}_{2st} - \bar{y}_N \mid \mathcal{F})\} = 8^{-1}\left[C\zeta_1^2 + 0.5C\sigma_b^2 + (1 - m^{-1})\sigma_e^2\right], \tag{3.1.40}$$

where $\hat{\theta}_{2st}$ is the estimated mean and $C = 2(2m-1)^{-1}(m-1)$. If $\zeta_1^2 > 0.25(m-1)^{-1}\sigma_b^2$, the anticipated variance for the controlled design is less than that of the two-per-stratum design.

To construct an estimator of the variance, we note that for a given h and t,

$$\begin{aligned} &E\{[m(y_{h1k} - y_{h2r}) - m(y_{t1k} - y_{t2r})]^2 \mid \mathcal{F}\} \\ &\quad = E\{[m(\gamma_{h1} - \gamma_{h2}) - m(\gamma_{t1} - \gamma_{t2})]^2 \mid \mathcal{F}\} \\ &\qquad + m^2(1 - m^{-1})(S_{h1}^2 + S_{h2}^2) + m^2(1 - m^{-1})(S_{t1}^2 + S_{t2}^2) \end{aligned}$$

and

$$E\{(y_{hrk} - y_{hrj})^2 \mid \mathcal{F}\} = 2S_{hr}^2 \quad \text{for } j \neq k.$$

Because every arrangement of strata is equally likely,

$$\begin{aligned} \hat{V}\{\hat{\theta} \mid \mathcal{F}\} &= N^{-2}\{[m(y_{q11} - y_{q21}) - m(y_{r11} - y_{r21})]^2\} \\ &\quad + N^{-2}\{m^2(1 - 2m^{-1})(y_{h11} - y_{h12})^2 \\ &\qquad + m^2(1 - 2m^{-1})(y_{t21} - y_{t22})^2\}, \end{aligned} \tag{3.1.41}$$

where strata h and t have two elements in one group, stratum q has one element in each group, and stratum r has one element in each group, is a design-unbiased estimator of the variance of $\hat{\theta}$. If the finite population correction is ignored, $\hat{V}\{\hat{\theta} \mid \mathcal{F}\}$ is the residual mean square from the regression of y_{hjk} on stratum indicator variables and an indicator variable for group.

3.1.5 Design for subpopulations

A common set of competing design objectives is the desire for good estimates at the, say, national level and a desire for good estimates at the subnational, say, state level. If the within-state variances are the same, the optimum allocation to states for national estimates is proportional to the state population. Conversely, the optimal allocation for individual state estimates, assuming that all are equally important, is an equal sample size for each state. Because it is difficult for users to give precise measures of relative importance, a common design allocates the sample proportional to the square roots of the subpopulation sizes. If some subpopulations are very small, a lower bound may be placed on the subpopulation sample size.

Table 3.4 Alternative Subpopulation Sample Allocations

ID	Propor-tional	Square Root	Bound	ID	Propor-tional	Square Root	Bound
CA	1206	551	528	SC	143	190	182
TX	742	433	415	OK	123	176	169
NY	676	413	396	OR	122	175	168
FL	569	379	363	CT	121	175	167
IL	442	334	320	IA	104	162	155
PA	437	332	318	MS	101	160	153
OH	404	319	306	KS	96	155	149
MI	354	299	286	AR	95	155	148
NJ	299	275	263	UT	79	142	136
GA	292	271	260	NV	71	134	128
NC	287	269	258	NM	65	128	122
VA	252	252	242	WV	64	127	122
MA	226	239	229	NE	61	124	120
IN	217	234	224	ID	46	108	120
WA	210	230	220	ME	45	107	120
TN	203	226	217	NH	44	106	120
MD	199	224	215	HI	43	104	120
WI	191	219	210	RI	37	97	120
MD	189	218	209	MT	32	90	120
AZ	183	215	206	DE	28	84	120
MN	175	210	201	SD	27	82	120
LA	159	200	192	ND	23	76	120
AL	158	200	192	AK	22	75	120
CO	153	196	188	VT	22	74	120
KY	144	190	183	WY	17	66	120

Example 3.1.3. Table 3.4 was constructed using the populations of the 50 states of the United States as given by the 2000 U.S. Census. The numbers in the column "proportional" are approximately proportional to the populations, and the numbers in the column "square root" are approximately proportional to the square roots of the populations.

The variances of estimates for three designs are compared in Table 3.5 under the assumption of common within-state variances. The variances are standardized so that the variance for the national estimate under proportional sampling is 1. The numbers used in variance calculations were not rounded. With proportional sampling the variance for the largest state (California) is 8.3 and the variance for the smallest state (Wyoming) is 573.1. The average of the state variances is 132. If a fixed number is taken in each state, the variance of the estimator for each state is 50, but the variance of the national estimator is more than twice that for proportional allocation.

Allocating the sample proportional to the square roots of the population leads to a large reduction in the largest state variance and in the average of the state variances relative to proportional allocation. The cost of this reduction is an increase of 22% in the variance of the national estimator relative to proportional allocation. The variance of the state estimator for Wyoming is still three times that for equal allocation.

Table 3.5 Variances Under Alternative Designs

	Design			
Variance	Proportional	Square Root	Square Root Lower Bound	Fixed
National	1.00	1.22	1.27	2.19
Min. state	8.29	18.13	18.92	50.00
Max. state	573.08	150.76	83.33	50.00
Ave. state	131.96	64.31	58.23	50.00

The third column of the table is for an allocation with a lower bound on the number allocated to a state. If one allocates proportional to population, Wyoming would receive 17 units in a sample of 10,000. Under allocation proportional to the square roots, Wyoming would receive 66 of the 10,000. Under the allocation of the third column, the 13 smallest states each receive 120 units. This allocation leads to a largest state variance that is nearly half that of the square root allocation. The variance of the national estimator is about 5% larger for the design with a lower bound than for the proportional-to-square-root design. In designing such a sample, it is relatively easy to prepare tables similar to Table 3.5 for alternative designs for the clients.

Preparing such tables is often preferable to asking the client to specify design requirements. ■■

3.2 MULTIPLE-STAGE SAMPLES

3.2.1 Cluster sampling

Cluster sampling was introduced in Section 1.2.7. Although cluster sampling requires no new theory for estimation, design for cluster samples warrants discussion for two reasons. First, it may be possible to form clusters of different sizes. For example, if the clusters are area clusters of households, it is possible to design clusters of different sizes using materials such as census block statistics. Second, known cluster size is an auxiliary variable that can be used at both the design and estimation stages. However, the cluster size is not always known at the design stage. For example, a sample of residence addresses provides a cluster sample of persons, but the number of persons per resident is often not known at the design stage.

Clusters must be formed so that the data collector finds it relatively easy to identify the unit correctly. For example, if the land area is being sampled for a survey of agricultural practices, units based on land ownership or farm operators would be practical, while units with boundaries defined by latitude and longitude would be less desirable. On the other hand, a segment based on latitude and longitude would be practical for a survey of forests given that geopositioning units can be used to determine location.

If determinations are to be made on field plots, such as measurements on plants, the plots must be of a reasonable size. It is well known that "edge effects" can bias the mean for small plots. Early studies of plots are those of Mahalanobis (1946) and Sukhatme (1947).

Practical considerations often limit the possible cluster sizes to a set of discrete possibilities. For example, much of the sampling in the NRI, the survey introduced in Example 1.2.2, is based on the Public Land Survey System. That system is based on units called *sections* which are 1 square mile. A section contains 640 acres and can be further divided into quarter sections of 160 acres and quarter-quarter sections of 40 acres. Thus, sampling units whose nominal sizes are multiples of 40 acres or of 160 acres were considered as possible sampling units for the NRI.

Given a range of practical sampling units, the classical problem is to form clusters to minimize the variance for a fixed cost or, equivalently, minimize the cost for a fixed variance. Typically, the cost per observation unit (interview, acre of land, meter of line segment) decreases as the size of the sampling unit

increases. Conversely, the variance of the mean per observation unit based on a fixed number of observation units increases as the size of the sampling unit increases. A size equal to a one-day workload typically reduces travel costs per observation. In surveys being conducted for the first time, both the cost and the variance as functions of cluster size can be approximated only crudely. In other cases, a previous survey, a pretest, or a census can furnish information on the correlation structure of the population. See Jessen (1978, Chapter 4) for an excellent discussion of the choice of sampling unit.

The cluster size is a natural stratification variable because we expect the cluster total to be related to cluster size. Often, the size used in design is not the actual number of elements in the cluster. For example, the design size of the segments studied in Example 2.1.1 was 160 acres, but the realized sizes varied from 100 to 345 acres. Similarly, the number of addresses in a block is not always the same as the number of households, but the number of addresses can be used as a measure of size because it will be strongly correlated with the number of households.

Given that the cluster sizes are known, we study alternative selection and estimation strategies. Consider the design model

$$y_{ij} = \mu_y + b_i + e_{ij}, \tag{3.2.1}$$

where b_i are the primary unit effects, e_{ij} are secondary unit effects, the b_i are independent $(0, \sigma_b^2)$ random variables, the e_{ij} are independent $(0, \sigma_e^2)$ random variables, and the e_{ij} are independent of the b_k for all i, j, and k. Note that μ_y is the mean per element, not the mean per cluster. Assume a population of cluster sizes with mean μ_M and variance σ_M^2. Also, assume that e_{ij}, b_i, and M_i are mutually independent. Then the expected total for a cluster of size M_i is $M_i\mu_y$ and the variance of the cluster total for a cluster of size M_i, denoted by γ_{ii}, is

$$\gamma_{ii} = E\{[y_i - E(y_i \mid M_i)]^2 \mid M_i\} = M_i^2\sigma_b^2 + M_i\sigma_e^2.$$

The mean of cluster totals is

$$E\{y_i\} = E\left\{\sum_{j=1}^{M_i} y_{ij}\right\} = \mu_M\mu_y$$

and the population variance of the cluster totals is

$$\begin{aligned} V\{y_i\} &= V\left\{M_i\mu_y + M_ib_i + \sum_{j=1}^{Mi} e_{ij}\right\} \\ &= \sigma_M^2\mu_y^2 + (\mu_M^2 + \sigma_M^2)\sigma_b^2 + \mu_M\sigma_e^2. \end{aligned} \tag{3.2.2}$$

Because the expected value for a cluster total is proportional to M_i and because the conditional variance of a cluster total is positively related to cluster size, the ratio estimator is a natural estimator to consider. The ratio estimator of the total of y is

$$\hat{T}_{y,rat} = T_M \hat{\theta}_n, \tag{3.2.3}$$

where

$$\hat{\theta}_n = \hat{T}_{y,HT} \hat{T}_{M,HT}^{-1}, \tag{3.2.4}$$

$$(\hat{T}_{y,HT}, \hat{T}_{M,HT}) = \sum_{i \in A} \pi_i^{-1}(y_i, M_i), \tag{3.2.5}$$

M_i is the number of elements in cluster i, and T_M is the total of M_i for the population. Some large-sample properties of the estimator of a ratio are given in Theorem 1.3.7 and ratio estimation is discussed in Section 2.1.

Given known σ_b^2 and σ_e^2, and the regression estimator, the best large-sample strategy is to select clusters with probabilities proportional to $(M_i^2\sigma_b^2 + M_i\sigma_e^2)^{1/2}$ and use the regression estimator with, say, $x_i = M_i$ or $\mathbf{x}_i = (M_i, M_i^2)$. See Theorem 3.1.1. Stratified random sampling with strata formed on the basis of cluster size and sampling with probability proportional to cluster size are common in practice. Selection with probabilities proportional to M_i is often called selection with *probability proportional to size* (PPS). Simple random sampling and PPS are popular partly because of simplicity and partly because practitioners may be unwilling to specify the ratio of σ_b^2 to σ_e^2.

We compare equal probability sampling, sampling with probability proportional to M_i, and sampling with probability proportional to $\gamma_{ii}^{0.5} = (M_i^2\sigma_b^2 + M_i\sigma_e^2)^{1/2}$. In all three cases we use the ratio estimator and ignore the finite population correction.

Under the model, the approximation for the variance of the ratio estimator of the mean per element for a simple random sample of size n is

$$\begin{aligned} V\{\bar{M}_n^{-1}\bar{y}_n\} &\doteq n^{-2}\mu_M^{-2} E\left\{ V\left(\sum_{i \in A} (y_i - \mu_y M_i) \mid \mathcal{F} \right) \right\} \\ &= n^{-1}\mu_M^{-2} N^{-1} E\left\{ \sum_{i=1}^{N} \gamma_{ii} \right\}, \end{aligned} \tag{3.2.6}$$

where $(\bar{M}_n, \bar{y}_n)$ is the simple sample mean of (M_i, y_i).

Given the model, the approximate variance of the ratio estimator for a sample of size n selected with probabilities equal to $nT_M^{-1}M_i$ is

$$V\left\{T_M^{-1}\sum_{i\in A} n^{-1}T_M M_i^{-1}(y_i-\mu_y M_i)\right\}$$
$$= \mu_M^{-2}N^{-2}E\left\{n^{-1}T_M\sum_{i=1}^{N} M_i^{-1}(y_i-\mu_y M_i)^2\right\}$$
$$= n^{-1}\mu_M^{-2}N^{-1}E\left\{\bar{M}_N\sum_{i=1}^{N} M_i^{-1}\gamma_{ii}\right\}, \tag{3.2.7}$$

where we used the variance for replacement sampling.

The approximation for the variance of the ratio estimator for a sample of size n selected with probabilities equal to $nT_\kappa^{-1}\gamma_{ii}^{0.5}$, where T_κ is the population total of $\kappa_i = \gamma_{ii}^{0.5}$, is

$$V\left\{\hat{T}_{M,HT}^{-1}n^{-1}\sum_{i\in A} T_\kappa\gamma_{ii}^{-0.5}(y_i-\mu_y M_i)\right\}$$
$$\doteq n^{-2}\mu_M^{-2}N^{-2}E\left\{V\left(T_\kappa\sum_{i\in A}\gamma_{ii}^{-0.5}(y_i-\mu_y M_i)\mid\mathcal{F}\right)\right\}$$
$$= n^{-1}\mu_M^{-2}N^{-2}E\left\{\left(\sum_{i=1}^{N}\gamma_{ii}^{0.5}\right)^2\right\}. \tag{3.2.8}$$

By Theorem 3.1.1, the minimum large-sample variance is that associated with probabilities proportional to $\gamma_{ii}^{0.5}$ and we have

$$N^{-1}\left(\sum_{i=1}^{N}\gamma_{ii}^{0.5}\right)^2 \le \bar{M}_N\sum_{i=1}^{N} M_i^{-1}\gamma_{ii} \le \sum_{i=1}^{N}\gamma_{ii}. \tag{3.2.9}$$

If all $M_i = \mu_M$, the procedures are equivalent. If $\sigma_e^2 = 0$, the two unequal probability procedures are equivalent.

The comparison of (3.2.6) and (3.2.8) neglects the fact that the cost for the PPS designs might be greater because the number of elements expected to be observed is larger for unequal probability selection than for equal probability selection. Also, stratification on γ_{ii} will reduce the differences among the procedures.

3.2.2 Two-stage sampling

The design options expand considerably when we permit subsampling of the primary sampling units. A classic design problem is the sample allocation between primary and secondary units for a population of equal-sized PSUs. We determine the optimal allocation under the design model (3.2.1). Consider a design in which m secondary units are to be selected from each of n_1 PSUs, with simple random sampling used at both stages. Assume that all PSUs are of size M. By the estimation theory for two-stage samples introduced in Section 1.2.8, the anticipated variance of the mean per element for a sample of n_1 primary units with m secondary units per primary unit is

$$\begin{aligned} V\{\hat{\theta}\} &= (1 - N^{-1}n_1)n_1^{-1}\sigma_b^2 \\ &\quad + \left[1 - (NM)^{-1}n_1 m\right](n_1 m)^{-1}\sigma_e^2, \end{aligned} \tag{3.2.10}$$

where

$$\hat{\theta} = (n_1 m)^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{m} y_{ij}.$$

A reasonable approximation to the cost function for many surveys is

$$C = c_1 n_1 + c_2 n_1 m, \tag{3.2.11}$$

where c_1 is the cost associated with the PSUs and c_2 is the cost associated with the SSUs. In an area sample for personal interview, c_1 will include travel cost and costs such as sketching the sampling unit and listing dwellings in the unit. The cost c_2 is the cost of conducting the interview. If one minimizes (3.2.10) subject to (3.2.11), one obtains

$$m_{opt} = \left[c_1 c_2^{-1} \sigma_b^{-2} \sigma_e^2\right]^{1/2}. \tag{3.2.12}$$

Of course, m_{opt} must be rounded to an integer between 1 and M. The number of PSUs is obtained from (3.2.11), or from (3.2.10) if a certain variance is desired. The result (3.2.12) is reasonable because m increases as c_2 decreases and as σ_e^2 increases.

Example 3.2.1. A survey of land cover for the northern part of the state of Alaska is being designed. Aerial photography will be used to obtain square images that are 3 miles on a side. Because of the large distances between units, each image costs about $2000. The secondary sampling unit is a square 1/2 mile on a side, called a *segment*. Thus, each image contains 36 segments. It is estimated that the cost of data collection for a segment is about $200.

The ratio of σ_e^2 to σ_b^2 will depend on the characteristic being observed, but a 6:1 ratio is judged reasonable. Using this variance ratio, $m = 7.7$ is optimal. In the study, $m = 9$ was used because of the simplicity associated with one-in-four subsampling. ■■

Designs in which the weights for all secondary units are the same are called *self-weighting*. Self-weighting designs are appealing for surveys of human populations conducted for general-purpose use because in such surveys there is little reason to postulate unequal variances for individuals.

Given PSUs selected with equal probability, the sample will be self-weighting when the same fraction of secondary units is selected in each PSU selected. Another popular design for two-stage samples is one in which the PSUs are selected with PPS, where size is the number of SSUs in the PSU. If the same number of secondary units is selected in each PSU, called a *fixed take*, the design is self-weighting. Often, m is chosen to meet practical restrictions such as the number of interviews that can be completed in a fixed time period.

We compare the PPS fixed take with the equal-probability proportional-take procedure. Recall that given a known total number of elements in the population and an equal-probability cluster sample, the ratio estimator is inferior to PPS sampling with the Horvitz–Thompson estimator under the population model (3.2.1). See (3.2.6) and (3.2.7). A ratio estimator of the total of y for a two-stage sample is

$$\hat{T}_{2s,rat} = N\bar{M}_N \left(\sum_{i \in A_1} \pi_i^{-1} M_i \right)^{-1} \hat{T}_{2s}, \tag{3.2.13}$$

where

$$\hat{T}_{2s} = \sum_{i \in A_1} \pi_i^{-1} \hat{y}_i,$$

$$\hat{y}_i = \sum_{j \in B_i} \pi_{(ij)|i}^{-1} y_{ij},$$

and $T_M = N\bar{M}_N$ is the total number of elements in the population. See (1.2.77).

Recall from (1.2.78) that the design variance of a two-stage sample can be written as

$$V_1\{\hat{T}_{1s} \mid \mathcal{F}\} + E\{V[\hat{T}_{2s} \mid (A_1, \mathcal{F})] \mid \mathcal{F}\},$$

where $\hat{T}_{1s}$ is the estimated total with all $m_i = M_i$, $V_1\{\hat{T}_{1s} \mid \mathcal{F}\}$ is the variance of the estimated total with $m_i = M_i$ for all i, and $V[\hat{T}_{2s} \mid (A_1, \mathcal{F})]$ is

the conditional variance, conditional on the selected first-stage units. Under model (3.2.1), omitting the first-stage finite population correction and simple random sampling at the second stage, the anticipated variance is

$$\begin{aligned} E[V\{\hat{T}_{2s,rat} \mid \mathcal{F}\}] &\doteq E[V_1\{\hat{T}_{rat} \mid \mathcal{F}\}] \\ &\quad + N^2 n^{-2} E\left\{ E\left(\sum_{i \in A_1} M_i^2 \left(1 - M_i^{-1} m_i\right) m_i^{-1} S_{e,i}^2 \mid \mathcal{F} \right) \right\} \\ &= E[V_1\{\hat{T}_{rat} \mid \mathcal{F}\}] + E\{N^2 n^{-1} (\bar{M}_N^2 m^{-1} - \bar{M}_N)\} \sigma_e^2 \\ &\doteq N^2 [\, n^{-1} (\mu_M^2 + \sigma_M^2) \sigma_b^2 + n^{-1} m^{-1} \mu_M^2 \sigma_e^2 \,], \end{aligned} \tag{3.2.14}$$

where $S_{e,i}^2$ is the mean square of e_{ij} for PSU i, the fixed subsampling rate is $\bar{M}_N^{-1} m$,

$$\begin{aligned} E[V_1\{\hat{T}_{rat} \mid \mathcal{F}\}] &= E\left\{ n^{-1} N \sum_{i \in U} (y_i - \mu_y M_i)^2 \right\} \\ &= N n^{-1} E\left\{ \sum_{i \in U} \gamma_{ii} \right\}, \end{aligned}$$

and $\gamma_{ii} = M_i^2 \sigma_b^2 + M_i \sigma_e^2$.

Under model (3.2.1) and PPS fixed-take sampling,

$$\begin{aligned} E[V\{\hat{T}_{2s,PPS} \mid \mathcal{F}\}] &= E[V_1\{\hat{T}_{PPS} \mid \mathcal{F}\}] \\ &\quad + E\left\{ E\left(T_M^2 \sum_{i \in A_1} \left(1 - M_i^{-1} m\right) m^{-1} S_{e,i}^2 \mid \mathcal{F} \right) \right\} \\ &= E\{V_1(\hat{T}_{PPS} \mid \mathcal{F})\} + E\{N^2 n^{-1} (\bar{M}_N^2 m^{-1} - \bar{M}_N)\} \sigma_e^2 \\ &\doteq N^2 (n^{-1} \mu_M^2 \sigma_b^2 + n^{-1} m^{-1} \mu_M^2 \sigma_e^2), \end{aligned} \tag{3.2.15}$$

where

$$\begin{aligned} E\{V_1(\hat{T}_{PPS} \mid \mathcal{F})\} &= E\left\{ N n^{-1} \bar{M}_N \sum_{i \in U} M_i^{-1} (y_i - \mu_y M_i)^2 \right\} \\ &= N n^{-1} E\left\{ \bar{M}_N \sum_{i \in U} M_i^{-1} \gamma_{ii} \right\}. \end{aligned}$$

Therefore, the PPS scheme is preferred to equal-probability selection for populations generated by model (3.2.1) under the cost function (3.2.11).

The cost structure (3.2.11) is often a reasonable approximation for the PPS fixed-take design. It follows that (3.2.12) furnishes an approximation to the optimum number of secondary units per PSU.

3.3 MULTIPLE-PHASE SAMPLES

3.3.1 Two-phase samples

The procedure called *double sampling* or *two-phase* sampling is typically employed in the following situation. There exists a procedure, relatively cheap to implement, that produces a vector of observations denoted by $\mathbf{x}$. The vector $\mathbf{x}$ is correlated with the characteristics of interest, where the vector of interest is denoted by $\mathbf{y}$. In some situations, the $\mathbf{x}$-variables are of interest in themselves and may be a part of $\mathbf{y}$. It is very expensive to make determinations on $\mathbf{y}$.

In the most popular form of two-phase sampling, a relatively large sample is selected and $\mathbf{x}$ determined on this sample. This sample is called the *first-phase sample* or *phase 1 sample*. Determinations for the vector $\mathbf{y}$ are made on a subsample of the original sample. The subsample is called the *second-phase sample* or *phase 2 sample*. In the form originally suggested by Neyman (1938), the original sample was stratified on the basis of $\mathbf{x}$ and the stratified estimator for $\mathbf{y}$ constructed using the estimated stratum sizes estimated with the phase 1 sample. We first describe this particular, and important, case of two-phase sampling. We simplify the discussion by considering scalar y. We also begin with a simple sampling procedure.

Assume that a simple random nonreplacement sample of n_1 elements is selected and assume that on the basis of the observations, the sample of n_1 elements is divided into G subgroups, also called *phase 2 strata*. We use the terms *group* and *phase 2 stratum* interchangeably. A sample of n_{2g}, $n_{2g} > 0$, elements is selected in the gth group for $g = 1, 2, \ldots, G$. The estimated number of elements in the gth phase 2 stratum is

$$\hat{N}_{1,g} = N n_1^{-1} \sum_{i \in A_1} x_{ig}, \tag{3.3.1}$$

where A_1 is the set of elements in the phase 1 sample,

$$\begin{aligned} x_{ig} &= 1 \quad \text{if element } i \text{ is in group } g \\ &= 0 \quad \text{otherwise} \end{aligned}$$

and

$$\bar{x}_{1g} = n_1^{-1} \sum_{i \in A_1} x_{ig}.$$

The vector of estimated population sizes is

$$\left(\hat{N}_{1,1}, \hat{N}_{1,2}, \ldots, \hat{N}_{1,G}\right) = N n_1^{-1} \sum_{i \in A_1} \mathbf{x}_i, \tag{3.3.2}$$

where the gth element of $\mathbf{x}_i$ is as defined in (3.3.1).

A reasonable estimator of the population total of y is obtained by weighting the stratum sample means of y with the estimated group sizes. This estimated total is

$$\hat{T}_{2p,st} = \sum_{g=1}^{G} \hat{N}_{1,g} \bar{y}_{2g}, \tag{3.3.3}$$

where

$$\bar{y}_{2g} = n_{2g}^{-1} \sum_{j \in A_{2g}} y_j,$$

and A_{2g} is the set of indexes of elements in group g of the phase 2 sample. The corresponding estimator of the population mean is

$$\bar{y}_{2p,st} = \sum_{g=1}^{G} \bar{x}_{1g} \bar{y}_{2g}, \tag{3.3.4}$$

where $\bar{x}_{1g}$, defined in (3.3.1), is the estimated fraction of the population that is in group g.

The estimator (3.3.4) is, conditionally on the phase 1 sample, unbiased for the phase 1 sample mean. That is,

$$E\left\{\bar{y}_{2p,st} \mid (A_1, \mathcal{F})\right\} = \bar{y}_1,$$

where

$$\bar{y}_1 = n_1^{-1} \sum_{i \in A_1} y_i.$$

It follows that the estimator (3.3.3) is unbiased for the finite population total,

$$\begin{aligned} E\{\hat{T}_{2p,st} \mid \mathcal{F}\} &= E\{E[\hat{T}_{2p,st} \mid (A_1, \mathcal{F})] \mid \mathcal{F}\} \\ &= E\{\hat{T}_1 \mid \mathcal{F}\} \\ &= T, \end{aligned} \tag{3.3.5}$$

and again using conditional expectations,

$$\begin{aligned} V\{(\hat{T}_{2p,st} - T) \mid \mathcal{F}\} &= V\{E[(\hat{T}_{2p,st} - T) \mid (A_1, \mathcal{F})] \mid \mathcal{F}\} \\ &\quad + E\{V[\hat{T}_{2p,st} \mid (A_1, \mathcal{F})] \mid \mathcal{F}\} \\ &= V\{(\hat{T}_1 - T) \mid \mathcal{F}\} \\ &\quad + E\{V[\hat{T}_{2p,st} \mid (A_1, \mathcal{F})] \mid \mathcal{F}\}, \end{aligned} \tag{3.3.6}$$

where

$$\hat{T}_1 = N n_1^{-1} \sum_{i \in A_1} y_i. \tag{3.3.7}$$

No assumption about the procedure used to form the groups for phase 2 selection is required in showing that $\hat{T}_{2p,st}$ is unbiased for T. Similarly, no assumption about the phase 1 design, beyond the existence of the variance, is used in deriving (3.3.6). By (3.3.6), the variance of $\hat{T}_{2p,st}$ for a two-phase sample is always larger than the variance of the estimated total computed from a phase 1 sample in which y is directly observed. Thus, two-phase sampling is used when the determinations on y are much more expensive than determinations on x, or when other operational conditions restrict the possible number of y determinations.

Given a cost function and a design model, the optimal design can be determined. Let σ_y^2 be the population variance of y and σ_w^2 be the common within phase 2 strata variance of y. Let $\sigma_b^2 = \sigma_y^2 - \sigma_w^2$. Assume that each phase 1 observation costs c_1 and each phase 2 observation costs c_2. Ignoring the phase 1 finite population correction factor,

$$V\{N^{-1}\hat{T}_{2p,st} \mid \mathcal{F}_N\} = n_1^{-1}\sigma_y^2 + (n_2^{-1} - n_1^{-1})\sigma_w^2 = n_1^{-1}\sigma_b^2 + n_2^{-1}\sigma_w^2, \tag{3.3.8}$$

and for $n_2 > 0$, the optimal sample sizes satisfy

$$n_2 = \left[(\sigma_b^2 c_2)^{-1}\sigma_w^2 c_1\right]^{1/2} n_1. \tag{3.3.9}$$

Result (3.3.9) can be compared to result (3.2.12). Variance expression (3.2.10) is analogous to expression (3.3.8), but the boundary conditions differ. In two-stage sampling, $1 \leq m \leq M$, whereas for two-phase sampling, $0 < n_2 \leq n_1$. Because $n_2 = n_1$ is an acceptable design, the variance for two-phase sampling must be checked against the variance for single-phase sampling to determine the optimal procedure.

Example 3.3.1. Suppose that we have 10,000 units to spend, $c_1 = 1$, $c_2 = 3$, $\sigma_y^2 = 100$, and $\sigma_w^2 = 40$. Then the n_1 from (3.3.9) is 4141 and the n_2 is 1953. The two-phase variance is 0.035. As an alternative design,

consider selecting a simple random sample of size 3333 and observing y on the selected sample. The variance of the alternative design is 0.030. Therefore, the described two-phase design is less efficient than using only the phase 1 design as a single phase. Suppose that the cost of observing y is 100. Then the optimal sample sizes are $n_1 = 1100$ and $n_2 = 89$. The two-phase variance of the mean is 0.504. If we use the simple random sample design, the variance of the mean is 1. ■■

It is relatively easy to estimate $V\{\hat{T}_{2p,st}|(A_1, \mathcal{F})\}$ because it is the variance of a stratified sample. It is more difficult to estimate the first term of (3.3.6), and variance estimation is very difficult when the phase 1 sample is selected by a complex design.

If the phase 1 sample is a simple random nonreplacement sample, the first term of (3.3.6) is

$$V\{\hat{T}_1 - T \mid \mathcal{F}\} = N(N - n_1)n_1^{-1}S_y^2.$$

For the stratified phase 2 sample with a simple random nonreplacement phase 1 sample,

$$\hat{S}_y^2 = \sum_{g=1}^{G} N^{-1}\hat{N}_{1g}n_{2g}^{-1}\sum_{i\in A_{2g}}(y_i - \bar{y}_{2p,st})^2, \tag{3.3.10}$$

where $\bar{y}_{2p,st}$ is defined as in (3.3.4), is a consistent estimator of S_y^2, under mild assumptions. Then a consistent estimator of the variance (3.3.6) is

$$\hat{V}\{\hat{T}_{2p,st} - T \mid \mathcal{F}\} = N(N - n_1)n_1^{-1}\hat{S}_y^2 + \sum_{g=1}^{G}\hat{N}_{1g}(\hat{N}_{1g} - n_{2g})n_{2g}^{-1}s_{y2g}^2, \tag{3.3.11}$$

where

$$s_{y2g}^2 = (n_{2g} - 1)^{-1}\sum_{i\in A_{2g}}(y_i - \bar{y}_{2g})^2.$$

We now consider two-phase samples with more complex phase 1 designs. Let the probability that element i is included in the phase 1 sample be π_{1i}, and let $\pi_{2i|A1}$ be the probability that element i is included in the sample given the specific phase 1 sample containing i. Thus,

$$\hat{T}_{2p} = \sum_{i\in A_2}\pi_{1i}^{-1}\pi_{2i|A1}^{-1}y_i \tag{3.3.12}$$

is unbiased for T, by the arguments of (3.3.5). Let π_{2i} be the unconditional probability that element i is included in the phase 2 sample. Then

$$\pi_{2i} = \pi_{1i}\pi_{2i|1i},$$

where $\pi_{2i|1i}$ is the conditional probability that element i is included in the phase 2 sample given that i is in the phase 1 sample.

We assume that the population is divided into G groups that serve as phase 2 strata. We assume simple random sampling within strata and that the sampling rate for group g for the phase 2 sample, $\pi_{2i|1i} = f_{2g}$, is fixed. The number of phase 2 sample elements in group g is n_{2g}, where n_{2g} is the integer closest to $n_{1g}f_{2g}$, and n_{1g} is the number of phase 1 sample elements in group g. The rounding error is ignored in the subsequent discussion. Let $\mathbf{x}_i$ be a G-dimensional vector with a 1 in position g when element i is in group g and zeros in the remaining locations, as defined in (3.3.1). The phase 1 vector of means is

$$(\bar{y}_{1\pi},\ \bar{\mathbf{x}}_{1\pi}) = \sum_{i\in A_1} w_{1i}(y_i,\ \mathbf{x}_i),$$

where $w_{1i} = (\Sigma_{j\in A_1}\pi_{1j}^{-1})^{-1}\pi_{1i}^{-1}$ and $\bar{\mathbf{x}}_{1\pi}$, but not $\bar{y}_{1\pi}$, is observed. An estimator of the mean of y is

$$\bar{y}_{2pr,st} = \sum_{g=1}^{G} \bar{x}_{1\pi,g}\bar{y}_{2\pi,g}, \tag{3.3.13}$$

where $\bar{x}_{1\pi,g} = \Sigma_{j\in A_1} w_{1i}x_{ig}$, is the phase 1 estimated fraction of the population in group g,

$$\begin{aligned}\bar{y}_{2\pi,g} &= \left(\sum_{i\in A_{2g}} \pi_{2i}^{-1}\right)^{-1} \sum_{i\in A_{2g}} \pi_{2i}^{-1}y_i \\ &= \left(\sum_{i\in A_{2g}} \pi_{1i}^{-1}\right)^{-1} \sum_{i\in A_{2g}} \pi_{1i}^{-1}y_i \end{aligned} \tag{3.3.14}$$

is the phase 2 mean of y in group g, and A_{2g} is the set of indexes of the elements in group g of the phase 2 sample.

Estimator (3.3.13) uses a separate ratio estimator as the phase 2 estimator. An alternative estimator of the mean is

$$\bar{y}_{2p,st} = \sum_{g=1}^{G} n_{1g}\bar{u}_{2g}, \tag{3.3.15}$$

where

$$\bar{u}_{2g} = n_{2g}^{-1} \sum_{i \in A_{2g}} u_i$$

and $u_i = w_{1i} y_i$. The estimator of total associated with (3.3.15) is

$$\hat{T}_{y,DE} = \sum_{i \in A_2} \pi_{1i}^{-1} \pi_{2i|1i}^{-1} y_i.$$

The estimator $\hat{T}_{y,DE}$ is a Horvitz–Thompson estimator and is called a *double expansion estimator* by Kott and Stukel (1997). Note that if $\pi_{2i|A1}$ is used in place of $\pi_{2i|1i}$, and if $\pi_{2i|A1}$ is not fixed, the estimator is not a Horvitz–Thompson estimator. If the phase 1 probabilities have a wide range and the phase 2 stratum sample sizes are small, (3.3.15) may be preferred to estimator (3.3.13).

We give the limiting properties of $\bar{y}_{2p,st}$ in Theorem 3.3.1. Because of the complex nature of the sequence of populations and samples, we break with our usual practice and index elements in the sequence with k instead of N.

Theorem 3.3.1. Let $\{(y_i, \mathbf{x}_i)\}$ be a sequence of iid random variables with fifth moment, where the G elements of $\mathbf{x}_i$ are indicators for group membership and every element is a member of one and only one of the G groups. Let $\{\mathcal{F}_k, A_{1k}\}$ be a sequence of populations and samples where A_{1k} is a sample of size n_{1k} from $\mathcal{F}_k, \mathcal{F}_k \subset \mathcal{F}_{k+1}, A_{1k} \subset A_{1,k+1}$, and $\mathcal{F}_k$ contains the first N_k elements of $\{(y_i, \mathbf{x}_i)\}$. Assume that $\{\mathcal{F}_k, A_{1k}\}$ is such that the phase 1 estimator of a mean vector, denoted by $\hat{\boldsymbol{\theta}}_k$, satisfies

$$[V\{\hat{\boldsymbol{\theta}}_k \mid \mathcal{F}_k\}]^{-1/2} (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{Nk})' \mid \mathcal{F}_k \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \quad \text{a.s.}, \tag{3.3.16}$$

where $V\{\hat{\boldsymbol{\theta}}_k \mid \mathcal{F}_k\}$ is $O_p(n_{1k}^{-1})$, $\hat{\boldsymbol{\theta}}_k = (\bar{y}_{1\pi k}, \bar{\mathbf{x}}_{1\pi k})$, $\boldsymbol{\theta}_{Nk} = (\bar{y}_{Nk}, \bar{\mathbf{x}}_{Nk})$, and $(\bar{y}_{1\pi k}, \bar{\mathbf{x}}_{1\pi k})$ is defined in (3.3.13). Assume that the $\pi_{2i|1i}$ are fixed and constant within groups and assume that the sequence of phase 1 selection probabilities satisfy

$$K_L < n_1^{-1} N_k \pi_{1i} < K_U \tag{3.3.17}$$

for all i, for some positive K_L and K_U. Assume that the design is such that

$$\lim_{k \to \infty} N_k^{-1} \sum_{i \in A_{1k}} \pi_{1i}^{-1} (1, \mathbf{x}_i, y_i, y_i^2)' (1, \mathbf{x}_i, y_i, y_i^2) = \mathbf{M} \quad \text{a.s.}, \tag{3.3.18}$$

where $\mathbf{M}$ is a matrix of constants.

Then, as $k \to \infty$,

$$[V\{\bar{y}_{2p,st,k} \mid \mathcal{F}_k\}]^{-1/2} (\bar{y}_{2p,st} - \bar{y}_{Nk}) \mid \mathcal{F}_k \xrightarrow{\mathcal{L}} N(0, 1) \quad \text{a.s.}, \tag{3.3.19}$$

where $\bar{y}_{2p,st,k}$ is as defined in (3.3.15),

$$V\{\bar{y}_{2p,st,k} \mid \mathcal{F}_k\} = V\{\bar{y}_{1\pi k} \mid \mathcal{F}_k\} + E\left\{\sum_{g=1}^{G} n_{1gk}^2 \left(n_{2gk}^{-1} - n_{1gk}^{-1}\right) \tilde{S}_{1gk}^2 \mid \mathcal{F}_k\right\}, \quad (3.3.20)$$

$$\tilde{S}_{1gk}^2 = (n_{1gk} - 1)^{-1} \sum_{i \in A_{1gk}} (u_i - \bar{u}_{1gk})^2,$$

$$\bar{u}_{1gk} = n_{1gk}^{-1} \sum_{i \in A_{1gk}} w_{1i} y_i = n_{1gk}^{-1} \sum_{i \in A_{1gk}} u_i,$$

$u_i = w_{1i} y_i$, and A_{1gk} is the set of indices in group g of the phase 1 sample.

Proof. Variance expression (3.3.20) follows from expression (3.3.6) because the phase 2 sample is a stratified sample of the phase 1 values $w_{1i} y_i$. The term inside the expectation in (3.3.20) is the variance of the stratified sample estimator of $\sum_{i \in A_1} w_{1i} y_i$, conditional on the elements of A_{1k}.

The error in the phase 2 estimator of the phase 1 mean is

$$\bar{y}_{2p,st,k} - \bar{y}_{1\pi k} = \sum_{g=1}^{G} n_{1gk} (\bar{u}_{2gk} - \bar{u}_{1gk}),$$

where $u_i = w_{1i} y_i$ and $\bar{u}_{2gk}$ is as defined in (3.3.15). The conditional variance, conditional on the phase 1 sample, is

$$V\{\bar{y}_{2p,st,k} - \bar{y}_{1\pi k} \mid A_{1k}\} = \sum_{g=1}^{G} n_{1gk}^2 (n_{2gk}^{-1} - n_{1gk}^{-1}) \tilde{S}_{1gk}^2,$$

where $n_{1gk}^2 \tilde{S}_{1gk}^2$ is of order 1. By assumption (3.3.18),

$$\lim_{k \to \infty} N_k^{-1} \sum_{i \in A_{1k}} \pi_{1i}^{-1} = 1 \quad \text{a.s.} \quad (3.3.21)$$

and

$$\lim_{k \to \infty} n_{1k}^{-1} n_{1gk} = \bar{x}_{g,\infty} \quad \text{a.s.},$$

where $\bar{x}_{g,\infty}$ is the fraction of the population in group g. Let $\bar{y}_{2p,st,k} = \bar{y}_{1\pi k} + \bar{y}_{2p,st,k} - \bar{y}_{1\pi k}$ and consider the sequence

$$[V_\infty\{(\bar{y}_{2p,st,k} - \bar{y}_{1\pi k}) \mid \mathcal{F}_k\}]^{-1/2} (\bar{y}_{2p,st,k} - \bar{y}_{1\pi k}), \quad (3.3.22)$$

where

$$n_{1k}V_\infty\{(\bar{y}_{2p,st,k} - \bar{y}_{1\pi k}) \mid \mathcal{F}_k\}$$
$$= \lim_{k\to\infty} n_{1k}\left(\sum_{g=1}^{G} n_{1gk}^2\left(n_{2gk}^{-1} - n_{1gk}^{-1}\right)\tilde{S}_{1gk}^2 \mid \mathcal{F}_k\right) \quad (3.3.23)$$

is a well-defined a.s. limit by (3.3.18). By assumptions (3.3.17) and (3.3.18), $n_{1gk}N_k^{-1}\pi_{1i}^{-1}y_i$ has finite third moment. Therefore,

$$[V_\infty\{(\bar{y}_{2p,st,k} - \bar{y}_{1\pi k}) \mid \mathcal{F}_k\}]^{-1/2}(\bar{y}_{2p,st,k} - \bar{y}_{1\pi k}) \mid (A_{1k}, \mathcal{F}_k) \xrightarrow{\mathcal{L}} N(0,1), \quad (3.3.24)$$

for almost all sequences of phase 1 samples, by a modification of Corollary 1.3.5.1.

By assumption, the standardized phase 1 mean converges in distribution to a normal random variable. By Corollary 1.3.6.1 it follows from the limiting normality of the phase 1 mean and from (3.3.24) that the normalized sum

$$n_{1k}^{1/2}(\bar{y}_{1\pi k} - \bar{y}_{Nk}) + n_{1k}^{1/2}(\bar{y}_{2p,st,k} - \bar{y}_{1\pi k})$$

also converges to a normal random variable, and

$$[V\{\bar{y}_{2p,st,k} \mid \mathcal{F}_k\}]^{-1/2}(\bar{y}_{2p,st,k} - \bar{y}_{Nk}) \mid \mathcal{F}_k \xrightarrow{\mathcal{L}} N(0,1) \quad \text{a.s.}, \quad (3.3.25)$$

where

$$V\{\bar{y}_{2p,st,k} \mid \mathcal{F}_k\} = V\{\bar{y}_{1\pi k} \mid \mathcal{F}_k\} + V_\infty\{(\bar{y}_{2p,st,k} - \bar{y}_{1\pi k}) \mid \mathcal{F}_k\}.$$

■

The limiting distribution of estimator (3.3.13) follows from Theorem 3.3.1.

Corollary 3.3.1.1. Let the assumptions of Theorem 3.3.1 hold and let the estimator of the mean of y be $\bar{y}_{2pr,st}$ defined in (3.3.13). Then

$$[V\{\bar{y}_{2pr,st,k} \mid \mathcal{F}_k\}]^{-1/2}(\bar{y}_{2pr,st,k} - \bar{y}_{Nk}) \mid \mathcal{F}_k \xrightarrow{\mathcal{L}} N(0,1) \quad \text{a.s.}, \quad (3.3.26)$$

where

$$V\{\bar{y}_{2pr,st,k} \mid \mathcal{F}_k\} = V\{\bar{y}_{1\pi,k} \mid \mathcal{F}_k\}$$
$$+ E\left\{\sum_{g=1}^{G} n_{1gk}^2\left(n_{2gk}^{-1} - n_{1gk}^{-1}\right)\tilde{S}_{1,e,g,k}^2 \mid \mathcal{F}_k\right\},$$

$$\tilde{S}^2_{1,e,g,k} = (n_{1gk} - 1)^{-1} \sum_{i \in A_{1gk}} e^2_{1ik},$$

and $e_{1ik} = w_{1i}(y_i - \bar{y}_{2\pi,gk})$.

Proof. Let

$$(\bar{u}_{2gk}, \bar{w}_{1,2gk}) = n^{-1}_{2gk} \sum_{i \in A_{2g}} (u_i, w_{1i}).$$

By the vector extension of Theorem 3.3.1, the vector $(\bar{u}_{2gk}, \bar{w}_{1,2gk})$, properly normalized, has a limiting normal distribution for each g because the number of groups is fixed and the within-group sampling rates are constant. Hence, the ratio $\bar{y}_{2\pi,gk} = \bar{u}_{2gk}\bar{w}^{-1}_{1,2gk}$ has a limiting normal distribution for each g. The result analogous to (3.3.24) then holds for estimator (3.3.13), and (3.3.26) follows. ∎

The two-phase estimator with a stratified phase 2 sample given in (3.3.13) is a special case of a regression estimator. See Section 2.2.3. To consider two-phase regression estimation with an extended vector of observations on the phase 1 sample, let $\mathbf{x}_i$ be a k-dimensional vector of observations made on the phase 1 sample, where the first G elements of $\mathbf{x}_i$ are the x_{ig}, $g = 1, 2, \ldots, G$, of (3.3.13). Then a two-phase regression estimator of the mean of y is

$$\begin{aligned} \bar{y}_{2p,reg} &= \bar{\mathbf{x}}_{1\pi}\hat{\boldsymbol{\beta}}_{2\pi,y\cdot x} \\ &= \bar{y}_{2\pi} + (\bar{\mathbf{x}}_{1\pi} - \bar{\mathbf{x}}_{2\pi})\hat{\boldsymbol{\beta}}_{2\pi,y\cdot x}, \end{aligned} \tag{3.3.27}$$

where

$$(\bar{y}_{2\pi}, \bar{\mathbf{x}}_{2\pi}) = \left(\sum_{i \in A_2} \pi^{-1}_{2i} \right)^{-1} \sum_{i \in A_2} \pi^{-1}_{2i}(y_i, \mathbf{x}_i)$$

and

$$\hat{\boldsymbol{\beta}}_{2\pi,y\cdot x} = \left(\sum_{i \in A_2} \pi^{-1}_{2i}\mathbf{x}'_i\mathbf{x}_i \right)^{-1} \sum_{i \in A_2} \pi^{-1}_{2i}\mathbf{x}'_i y_i.$$

Alternative estimators of $\boldsymbol{\beta}$ can be used. See Chapter 2.

The extension of Theorem 3.3.1 to the more general $\mathbf{x}_i$ vector is given in Corollary 3.3.1.2.

Corollary 3.3.1.2. Let the assumptions of Theorem 3.3.1 hold, and let $\mathbf{x}_i$ be an $(r-1)$-dimensional vector with the first G elements defined by x_{ig} of (3.3.1). Assume that the covariance matrix of $(y,\ \mathbf{x})$ is positive definite. Let the vectors $\hat{\boldsymbol{\theta}}_{1k} = (\bar{y}_{1\pi k},\ \bar{\mathbf{x}}_{1\pi k})$ and $\hat{\boldsymbol{\theta}}_{2k} = vech\ \hat{\mathbf{M}}_k$ satisfy (3.3.16), where

$$\hat{\mathbf{M}}_k = \left(\sum_{i\in A_{2k}} \pi_{2i}^{-1}\right)^{-1} \sum_{i\in A_{2k}} \pi_{2i}^{-1}(y_i,\ \mathbf{x}_i)'(y_i,\ \mathbf{x}_i),$$

$vech\ \mathbf{M} = (m_{11},\ m_{21},\ \ldots,\ m_{r1},\ m_{22},\ \ldots,\ m_{r2},\ \ldots,\ m_{rr})'$, and m_{ij} is the ijth element of the $r \times r$ matrix $\mathbf{M}$.

Then

$$[V_\infty\{\bar{y}_{2p,reg,k} \mid \mathcal{F}_k\}]^{-1/2}(\bar{y}_{2p,reg,k} - \bar{y}_{Nk}) \mid \mathcal{F}_k \xrightarrow{\mathcal{L}} N(0,\ 1) \quad \text{a.s.,} \quad (3.3.28)$$

where $\bar{y}_{2p,reg,k}$ is as defined in (3.3.27),

$$\begin{aligned} V_\infty\{\bar{y}_{2p,reg,k} \mid \mathcal{F}_k\} &= V\{\bar{y}_{1\pi k} \mid \mathcal{F}_k\} \\ &\quad + E\left\{\sum_{g=1}^{G} \bar{x}_{1\pi,gk}^2 (n_{2gk}^{-1} - n_{1gk}^{-1})\sigma_{eg}^2 \mid \mathcal{F}_k\right\}, \\ \sigma_{eg}^2 &= E\left\{(y_i - \mathbf{x}_i\boldsymbol{\beta}_{y\cdot x})^2 \mid x_{ig} = 1\right\}, \end{aligned}$$

and $\boldsymbol{\beta}_{y\cdot x} = [E\{\mathbf{x}'\mathbf{x}\}]^{-1} E\{\mathbf{x}'y\}$.

Proof. Consider the regression estimated total

$$\hat{T}_{2p,reg,k} = \hat{T}_{2,y,k} + (\hat{\mathbf{T}}_{1,x,k} - \hat{\mathbf{T}}_{2,x,k})\hat{\boldsymbol{\beta}}_{2\pi,y\cdot x,k},$$

where

$$\begin{aligned} (\hat{T}_{2,y,k},\ \hat{\mathbf{T}}_{2,x,k}) &= \sum_{i\in A_{2k}} \pi_{2i}^{-1}(y_i,\ \mathbf{x}_i), \\ \hat{\mathbf{T}}_{1,x,k} &= \sum_{i\in A_{1k}} \pi_{1i}^{-1}\mathbf{x}_i, \end{aligned}$$

and $\hat{\boldsymbol{\beta}}_{2\pi,y\cdot x,k}$ is as defined in (3.3.27). Now, by the moment assumptions,

$$\hat{\boldsymbol{\beta}}_{2\pi,y\cdot x,k} = \boldsymbol{\beta}_{y\cdot x,N,k} + O_p(n_k^{-1/2})$$

and it follows that

$$\hat{T}_{2p,reg,k} = \tilde{T}_{2p,reg,y} + O_p(n_k^{-1}),$$

where $\tilde{T}_{2p,reg,k} = \hat{T}_{2,y,k} + (\hat{\mathbf{T}}_{1,x,k} - \hat{\mathbf{T}}_{2,x,k})\boldsymbol{\beta}_{y\cdot x}$. The difference

$$\begin{aligned} N_k^{-1}(\tilde{T}_{2p,reg,k} - \hat{T}_{1,y,k}) &= N_k^{-1} \sum_{i \in A_{2k}} \pi_{2i}^{-1} e_i - N_k^{-1} \sum_{i \in A_{1k}} \pi_{1i}^{-1} e_i \\ &= N_k^{-1} \sum_{g=1}^{G} n_{2gk}^{-1} n_{1gk} \sum_{i \in A_{2gk}} \pi_{1i}^{-1} e_i \\ &\quad - N_k^{-1} \sum_{g=1}^{G} \sum_{i \in A_{1gk}} \pi_{1i}^{-1} e_i, \end{aligned} \tag{3.3.29}$$

where $e_i = y_i - \mathbf{x}_i \boldsymbol{\beta}_{y\cdot x}$. Therefore, by analogy between (3.3.29) and (3.3.25), we have result (3.3.28). ■

To use the results of Corollary 3.3.1.2 to set approximate confidence intervals, a consistent estimator of the variance is required. Theoretically, the procedure used to construct (3.3.11) can be extended to more complex designs if the joint selection probabilities are known. Write the estimated variance as

$$\hat{V}\{\bar{y}_{2p,reg} \mid \mathcal{F}\} = N^{-2}[\hat{V}_1\{\hat{T}_1 \mid \mathcal{F}\} + \hat{V}\{\hat{T}_{2p,st} \mid (A_1, \mathcal{F})\}].$$

To construct an estimator of $V_1\{\hat{T}_1\}$, let

$$\hat{V}_{1,HT}\{\hat{T}_1 \mid \mathcal{F}\} = \sum_{j,k \in A_1} \pi_{1jk}^{-1}(\pi_{1jk} - \pi_{1j}\pi_{2j})\pi_{1j}^{-1}\pi_{1k}^{-1} y_j y_k$$

be the full-sample estimator of the variance of the phase 1 estimated total. Then, given the phase 2 sample, an unbiased estimator of the variance is

$$\hat{V}_{2,HT}\{\hat{T}_1 \mid \mathcal{F}\} = \sum_{j,k \in A_2} \pi_{2jk|A1}^{-1} \pi_{1jk}^{-1}(\pi_{1jk} - \pi_{1j}\pi_{1k})\pi_{1j}^{-1}\pi_{1k}^{-1} y_j y_k, \tag{3.3.30}$$

where $\pi_{2jk|A1}$ is the probability that y_j and y_k are included in the phase 2 sample given sample A_1. The proof that $\hat{V}_{2,HT}\{\hat{T}_1 \mid \mathcal{F}\}$ is unbiased follows the proof of (3.3.5) and requires no assumption about the phase 2 design other than positive joint selection probabilities. The difficulty in applying (3.3.30) is the determination of the joint probabilities as well as the calculation of the double summation.

A second estimator of the variance is based on an alternative expression for the variance of $\bar{y}_{2p,reg}$. By a Taylor approximation,

$$\bar{y}_{2p,reg} - \bar{y}_N = \bar{\mathbf{x}}_{1\pi}\hat{\boldsymbol{\beta}}_{\pi,y\cdot x} - \bar{x}_N \beta_{y\cdot x} - \bar{e}_N$$

$$
\begin{aligned}
&= (\bar{\mathbf{x}}_{1\pi} - \bar{\mathbf{x}}_N)\boldsymbol{\beta}_{y\cdot x} + \bar{\mathbf{x}}_N(\hat{\boldsymbol{\beta}}_{\pi,y\cdot x} - \boldsymbol{\beta}_{y\cdot x}) \\
&\quad - \bar{e}_N + O_p(n^{-1}) \\
&= (\bar{\mathbf{x}}_{1\pi} - \bar{\mathbf{x}}_N)\boldsymbol{\beta}_{y\cdot x} + \left(\sum_{i\in A_2} \pi_{2i}^{-1}\right)^{-1} \sum_{i\in A_2} \pi_{2i}^{-1} e_i \\
&\quad - \bar{e}_N + O_p(n^{-1}),
\end{aligned} \tag{3.3.31}
$$

where $e_i = y_i - \mathbf{x}_i\beta_{y\cdot x}$, $\boldsymbol{\beta}_{y\cdot x}$ is defined in (3.3.28) and the approximation for $\bar{\mathbf{x}}_N(\hat{\boldsymbol{\beta}}_{2\pi,y\cdot x} - \boldsymbol{\beta}_{y\cdot x})$ as the ratio estimator $\bar{e}_\pi$ follows from (2.2.55). Therefore, the variance of the approximating distribution is

$$
\begin{aligned}
V_\infty\{\bar{y}_{2p,reg} - \bar{y}_N \mid \mathcal{F}\} &= \boldsymbol{\beta}'_{y\cdot x} V\{\bar{\mathbf{x}}_{1\pi} \mid \mathcal{F}\}\boldsymbol{\beta}_{y\cdot x} + V\{\bar{e}_{2\pi} \mid \mathcal{F}\} \\
&\quad + 2C\{\boldsymbol{\beta}'_{y\cdot x}\bar{\mathbf{x}}'_{1\pi},\ \bar{e}_{2\pi} \mid \mathcal{F}\}.
\end{aligned} \tag{3.3.32}
$$

By the definition of e_i as the population residual, $C\{\mathbf{x}_i, e_i\} = \mathbf{0}$ for the population, and we expect $C\{\bar{\mathbf{x}}_{1\pi},\ \bar{e}_{2\pi}\}$ to be small for most designs. If the first stage is a simple random sample and the stratified estimator (3.3.4) is used, $E\{\bar{e}_{2\pi} \mid \bar{\mathbf{x}}_1\} = 0$ because $E\{\bar{y}_g \mid n_{1g}\} = \bar{y}_{gN}$ for all g.

If $C\{\bar{\mathbf{x}}_{1\pi}\boldsymbol{\beta}_{y\cdot x},\ \bar{e}_{2\pi} \mid \mathcal{F}\}$ is zero, an estimator of the variance of $\bar{y}_{2p,reg}$ is

$$
\begin{aligned}
\hat{V}_T\{\bar{y}_{2p,reg} \mid \mathcal{F}\} &= \hat{V}_1\{\bar{\mathbf{x}}_{1\pi}\boldsymbol{\beta}_{\pi,y\cdot x} \mid \mathcal{F}\} + \hat{V}\{\bar{e}_{2\pi} \mid \mathcal{F}\} \\
&= \hat{\boldsymbol{\beta}}'_{\pi,y\cdot x}\hat{V}\{\bar{\mathbf{x}}_{1\pi} \mid \mathcal{F}\}\hat{\boldsymbol{\beta}}_{\pi,y\cdot x} \\
&\quad + \hat{V}\{\bar{e}_{2\pi} \mid \mathcal{F}\},
\end{aligned} \tag{3.3.33}
$$

where $\hat{V}_1\{\bar{\mathbf{x}}_{1\pi} \mid \mathcal{F}\}$ is a consistent estimator of the variance for the phase 1 mean of $\mathbf{x}_i$ and $\hat{V}\{\bar{e}_{2\pi} \mid \mathcal{F}\}$ is a consistent estimator of the unconditional variance of $\bar{e}_{2\pi}$. If the phase 1 finite population correction can be ignored and the phase 1 sample is a simple random sample, an estimator of $V\{\bar{e}_{2\pi} \mid \mathcal{F}\}$ is

$$
\hat{V}\{\bar{e}_{2\pi} \mid \mathcal{F}\} = \left(\sum_{i\in A_2} \pi_{2i}^{-1}\right)^{-2} \sum_{i\in A_2} \pi_{2i}^{-2}\hat{e}_i^2, \tag{3.3.34}
$$

where $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{2\pi,y\cdot x}$.

Example 3.3.2. The data of Table 3.6 were generated to illustrate some of the computations associated with two-phase sampling. The phase 1 sample is a sample of 22 elements in two strata. The elements of the phase 1 sample

are placed in three groups, also called phase 2 strata, where the groups are identified in the table. A simple random sample is selected in each group. The phase 2 sampling rate is two-in-five, four-in-nine, and three-in-eight for groups 1, 2, and 3, respectively.

Table 3.6 Data for Two-Phase Sample

Phase 1 Stratum	Element ID	Phase 1 Weight	Phase 2 Group	Phase 2 Weight	y
1	1	300	1		
	2	300	1	750	7.2
	3	300	2		
	4	300	2	675	8.6
	5	300	2		
	6	300	2	675	8.0
	7	300	3	800	6.2
	8	300	3		
2	1	200	1		
	2	200	1	500	5.2
	3	200	1		
	4	200	2		
	5	200	2	450	5.5
	6	200	2		
	7	200	2		
	8	200	2	450	6.3
	9	200	3		
	10	200	3		
	11	200	3	533	5.3
	12	200	3		
	13	200	3	533	4.9
	14	200	3		

Let

$$\begin{aligned} x_{hgi} &= 1 \quad \text{if element } i \text{ is in group } g \text{ of stratum } h \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Then the estimated total number of elements in group g is

$$\hat{T}_{xg} = \sum_{h=1}^{2} N_h \bar{x}_{hg}, \tag{3.3.35}$$

where

$$\bar{x}_{hg} = n_{h1}^{-1} \sum_{i \in A_{h1}} x_{hgi},$$

A_{h1} is the set of indices for elements in stratum h of the phase 1 sample and n_{h1} is the number of sample elements in stratum h of the phase 1 sample. The estimated numbers are

$$(\hat{T}_{x1}, \hat{T}_{x2}, \hat{T}_{x3}) = (1150, 2100, 1450).$$

The estimated covariance matrix of $(\bar{x}_{1,1}, \bar{x}_{2,1}, \bar{x}_{3,1}) = N^{-1}(\hat{T}_{x1}, \hat{T}_{x2}, \hat{T}_{x3})$, calculated by the standard stratified formula and ignoring the finite population correction is

$$\hat{V}\{\bar{x}_{1,1}, \bar{x}_{2,1}, \bar{x}_{3,1}\} = \begin{pmatrix} 1.926 & -1.137 & -0.789 \\ -1.137 & 2.589 & 1.452 \\ -0.789 & 1.452 & 2.241 \end{pmatrix} \times 10^{-2}. \tag{3.3.36}$$

The group means of y are

$$(\bar{y}_{2\pi,1}, \bar{y}_{2\pi,2}, \bar{y}_{2\pi,3}) = (6.400, 7.340, 5.572),$$

where $\bar{y}_{2\pi,g}$ is defined in (3.3.14). Although the sample sizes are small, we use estimator (3.3.13) for the mean of y because (3.3.13) is generally preferred in practice. The estimator is

$$\bar{y}_{2p,st} = N^{-1} \sum_{g=1}^{G} \hat{T}_{xg} \bar{y}_{2\pi,g} = 6.511.$$

To estimate the variance of $\bar{y}_{2p,st}$, we first estimate the phase 1 stratum variances for y by

$$\hat{S}_{yh}^2 = \left(\sum_{i \in B_{2h}} \omega_{B2hi} \right)^{-1} \sum_{i \in B_{2h}} \omega_{B2hi} (y_i - \bar{y}_{B2,h})^2,$$

where

$$\omega_{B2hi} = \left(\sum_{j \in B_{2h}} \pi_{2j}^{-1} \right)^{-1} \pi_{2i}^{-1},$$

$$\bar{y}_{B2,h} = \sum_{j \in B_{2h}} \omega_{B2hi} y_i,$$

and $B_{2h} = A_{1h} \cap A_2$ is the portion of the phase 2 sample in phase 1 stratum h. If the y_i in a stratum were iid random variables, $\hat{S}^2_{yh}$ would be an unbiased estimator of S^2_{yh}. The estimates are $(\hat{S}^2_{y1}, \hat{S}^2_{y2}) = (1.102, 0.268)$.

The estimator of the conditional variance of $\hat{y}_{2p,st}$ as an estimator of the phase 1 mean given the phase 1 sample is

$$\hat{V}\{\bar{y}_{2p,st} \mid (A_1, \mathcal{F})\} = N^{-2} \sum_{g=1}^{3} \hat{T}^2_{xg} \hat{V}\{\bar{y}_{2\pi,g} \mid (A_1, \mathcal{F})\} = 0.0959,$$

where

$$\hat{V}\{\bar{y}_{2\pi,g} \mid (A_1, \mathcal{F})\}$$
$$= \frac{(\hat{T}_{xg} - n_{2g})n_{2g}}{\hat{T}_{xg}(n_{2g} - 1)} \left(\sum_{i \in A_{2g}} \pi^{-1}_{1i} \right)^{-2} \sum_{i \in A_{2g}} \left[\pi^{-1}_{1i} (y_i - \bar{y}_{2\pi,g}) \right]^2.$$

Thus, an estimator of the variance of $\bar{y}_{2p,st}$ is

$$\begin{aligned} \hat{V}\{\bar{y}_{2p,st} - \bar{y}_N \mid \mathcal{F}\} &= \sum_{h=1}^{2} W^2_h n^{-1}_{h1} \hat{S}^2_{yh} + \hat{V}\{\bar{y}_{2p,st} \mid (A_1, \mathcal{F})\} \\ &= 0.0348 + 0.0959 = 0.1307. \end{aligned}$$

■■

The $\bar{y}_{2p,st}$ of Example 3.3.1 is not necessarily the best estimator of the mean of y because the information in the phase 1 strata is ignored in constructing $\bar{y}_{2\pi,g}$. The estimator of Corollary 3.3.1.2 can be extended to cover the situation where the population mean is known for some elements of x_i.

Example 3.3.3. We continue the analysis of the two-phase sample of Table 3.6. To incorporate the phase 1 stratum information into the estimator, we compute the regression estimator where the explanatory variables are indicators for phase 1 strata and for phase 2 groups. In our illustration there are two strata and we define the variable

$$\begin{aligned} u_{1i} &= 1 - 0.4615 \quad \text{if element } i \text{ is in stratum 1} \\ &= -0.4615 \quad \text{otherwise,} \end{aligned}$$

where 0.4615 is the fraction of the population in stratum 1. The estimated fraction of elements in phase 2 group g is the estimated mean of x_{gi}, where x_{gi} is the indicator function for group g. Let

$$a_{gi} = x_{gi} - \bar{x}_{g,st}, \qquad g = 1, 2$$

be the deviations of the indicator function from its phase 1 estimated mean. The regression estimator of the mean of y is computed using the phase 1 stratum indicator as an additional variable. By coding the regression variables with zero population mean or zero estimated population mean, the coefficient for 1, the *intercept*, in the phase 2 weighted regression of y on $(1, u_{1i}, a_{1i}, a_{2i}) =: \mathbf{z}_i$ is the regression estimator of the mean of y.

The estimated coefficient vector is

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}_2'\mathbf{W}_2\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{W}_2\mathbf{y}_2 = (6.365, 2.075, 0.512, 1.452)',$$

where $\mathbf{Z}_2$ is the $n_2 \times 4$ matrix of phase 2 observations on $\mathbf{z}_i$, $\mathbf{W}_2 = \text{diag}(w_{2i})$, and $\mathbf{y}_2$ is the vector of observations on y_i. Thus, the estimated mean of y, denoted by $\bar{y}_{2p,reg}$, is 6.365. Because the phase 1 strata are explanatory variables in the regression, the phase 1 mean of $e_i = y_i - \mathbf{z}_i\boldsymbol{\beta}_N$, where $\boldsymbol{\beta}_N = (\mathbf{Z}_N'\mathbf{Z}_N)^{-1}\mathbf{Z}_N'\mathbf{y}_N$ is uncorrelated with the phase 1 mean of x_{gi}. Therefore, we can use (3.3.33) to estimate the variance of $\bar{y}_{2p,reg}$. The variance of $\hat{\beta}_0 = \bar{y}_{2p,reg}$ as an estimator of $\bar{y}_N$ is

$$\begin{aligned} V\{\hat{\beta}_0 - \bar{y}_N \mid \mathcal{F}\} &= V\{\bar{\mathbf{z}}_{reg}\hat{\boldsymbol{\beta}} \mid \mathcal{F}\} \\ &\doteq V\{\hat{\beta}_0 \mid \mathcal{F}\} + \boldsymbol{\beta}' V\{\bar{\mathbf{z}}_{reg} - \bar{\mathbf{z}}_N \mid \mathcal{F}\}\boldsymbol{\beta}, \end{aligned}$$

where $\bar{\mathbf{z}}_{reg} - \bar{\mathbf{z}}_N = [0, 0, -(\bar{x}_{1,1} - \bar{x}_{1,N}), -(\bar{x}_{2,i} - \bar{x}_{2,N})]$. An estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ is

$$\hat{V}\{\hat{\boldsymbol{\beta}} \mid \mathcal{F}\} = n(n-r)^{-1}(\mathbf{Z}_2'\mathbf{W}_2\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{W}_2\hat{\mathbf{D}}_{ee}\mathbf{W}_2\mathbf{Z}_2(\mathbf{Z}_2'\mathbf{W}_2\mathbf{Z}_2)^{-1}$$

where $\mathbf{D}_{ee}' = \text{diag}\{\hat{e}_i^2\}$, $\hat{e}_i = y_i - \mathbf{z}_i\hat{\boldsymbol{\beta}}$, and r is the dimension of $\mathbf{z}_i$. For our illustration,

$$\hat{V}\{\hat{\boldsymbol{\beta}} \mid \mathcal{F}\} = \begin{pmatrix} 2.761 & -0.601 & -3.162 & 0.337 \\ -0.601 & 10.850 & -2.434 & -2.104 \\ -3.162 & -2.434 & 11.932 & 11.612 \\ 0.337 & -2.104 & 11.612 & 20.124 \end{pmatrix} \times 10^{-2}.$$

The estimated covariance matrix of $(\bar{x}_{1,1}, \bar{x}_{2,1})$ is the upper left 2×2 matrix of (3.3.36) of Example 3.3.2, and $(\hat{\beta}_3, \hat{\beta}_4)\hat{V}\{(\bar{x}_{1,1}, \bar{x}_{2,1}) \mid \mathcal{F}\}(\hat{\beta}_3, \hat{\beta}_4)' = 0.04271$. Thus,

$$\hat{V}\{\bar{y}_{2p,reg} - \bar{y}_N \mid \mathcal{F}\} = 0.02761 + 0.04271 = 0.07032.$$

In this example the phase 1 variance of $\bar{\mathbf{x}}_i$ contributes a sizable fraction to the total variance. The estimated variance of the regression estimator of this example is much smaller than that of the estimator of Example 3.3.2 because

y is correlated with the phase 1 strata even after adjusting for groups. Often, groups will be formed so that this partial correlation is small, but if not and if there are adequate degrees of freedom, variables for phase 1 strata can be included in the estimator. If there is a large number of phase 1 strata, a compromise procedure is to include indicators for groups of strata. If phase 1 strata are included in the regression or if there is no partial correlation between phase 1 strata and y, variance estimator (3.3.33), which assumes that $C\{\bar{\mathbf{x}}_{1\pi}, \bar{e}_{2\pi}\} = \mathbf{0}$, is appropriate. ■■

Replication methods for variance estimation for two-phase samples are studied in Section 4.4.

3.3.2 Three-phase samples

To extend the discussion to three-phase estimation, assume that a phase 3 sample of size n_3 is selected from a phase 2 sample of size n_2, which is itself a sample of a phase 1 sample of size n_1 selected from the finite population. Let A_1, A_2, and A_3 be the sets of indices for the phase 1, 2, and 3 samples, respectively.

Let the vector $(1, \mathbf{u})$ be observed on the phase 1 sample, the vector $(1, \mathbf{u}, \mathbf{x})$ observed on the phase 2 sample, and the vector $(1, \mathbf{u}, \mathbf{x}, y)$ observed on the phase 3 sample. We construct a phase 3 estimator by proceeding sequentially. Given the phase 1 and 2 samples, the regression estimator of the mean of $\mathbf{x}$ is

$$\bar{\mathbf{x}}_{2p,reg} = \bar{\mathbf{x}}_{2\pi} + (\bar{\mathbf{u}}_{1\pi} - \bar{\mathbf{u}}_{2\pi})\hat{\boldsymbol{\beta}}_{2\pi,x\cdot u}, \tag{3.3.37}$$

where

$$(\bar{\mathbf{u}}_{2\pi}, \bar{\mathbf{x}}_{2\pi}) = \left(\sum_{i\in A_2} \pi_{2i}^{-1}\right)^{-1} \sum_{i\in A_2} \pi_{2i}^{-1}(\mathbf{u}_i, \mathbf{x}_i),$$

$$\bar{\mathbf{u}}_{1\pi} = \left(\sum_{i\in A_1} \pi_{1i}^{-1}\right)^{-1} \sum_{i\in A_1} \pi_{1i}^{-1}\mathbf{u}_i,$$

and $\hat{\boldsymbol{\beta}}_{2\pi,x\cdot u}$ is the estimator

$$\left(\sum_{i\in A_2} (\mathbf{u}_i - \bar{\mathbf{u}}_{2\pi})'\pi_{2i}^{-1}(\mathbf{u}_i - \bar{\mathbf{u}}_{2\pi})\right)^{-1} \sum_{i\in A_2} (\mathbf{u}_i - \bar{\mathbf{u}}_{2\pi})'\pi_{2i}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{2\pi}).$$

Then the improved estimator of the mean of $\mathbf{x}$ together with the phase 1 mean of $\mathbf{u}$ can be used in a regression estimator constructed from phase 3. The

result is the three-phase regression estimator,

$$\bar{y}_{3p,reg} = \bar{y}_{3\pi} + (\bar{\mathbf{u}}_{1\pi} - \bar{\mathbf{u}}_{3\pi},\ \bar{\mathbf{x}}_{2p,reg} - \bar{\mathbf{x}}_{3\pi})\hat{\boldsymbol{\beta}}_{3\pi,y\cdot(u,x)}, \quad (3.3.38)$$

where

$$\hat{\boldsymbol{\beta}}_{3\pi,y\cdot(u,x)} = \left(\sum_{i\in A_3} \mathbf{r}'_{3,ux,i}\pi_{3i}^{-1}\mathbf{r}_{3,ux,i}\right)^{-1} \sum_{i\in A_3} \mathbf{r}'_{3,ux,i}\pi_{3i}^{-1}(y_i - \bar{y}_{3\pi}),$$

$$(\bar{y}_{3\pi},\ \bar{\mathbf{u}}_{3\pi},\ \bar{\mathbf{x}}_{3\pi}) = \left(\sum_{i\in A_3} \pi_{3i}^{-1}\right)^{-1} \sum_{i\in A_3} \pi_{3i}^{-1}(y_i,\ \mathbf{u}_i,\ \mathbf{x}_i),$$

$\mathbf{r}_{3,ux,i} = (\mathbf{u}_i - \bar{\mathbf{u}}_{3\pi},\ \mathbf{x}_i - \bar{\mathbf{x}}_{3\pi})$, and π_{3i} is the probability that element i enters the phase 3 sample. To obtain an alternative expression for the estimator, let

$$\hat{\mathbf{a}}_i = \mathbf{x}_i - \bar{\mathbf{x}}_{2\pi} - (\mathbf{u}_i - \bar{\mathbf{u}}_{2\pi})\hat{\boldsymbol{\beta}}_{2\pi,x\cdot u}, \quad (3.3.39)$$

where $\hat{\boldsymbol{\beta}}_{2\pi,x\cdot u}$ is as defined in (3.3.37). Then $\bar{y}_{3p,reg}$ can be written as

$$\bar{y}_{3p,reg} = \bar{y}_{3\pi} + (\bar{\mathbf{u}}_{1\pi} - \bar{\mathbf{u}}_{3\pi},\ -\bar{\mathbf{a}}_{r,3\pi})\hat{\boldsymbol{\beta}}_{3p,y\cdot(u,a)}, \quad (3.3.40)$$

where $\bar{\mathbf{a}}_{r,3\pi}$ is the phase 3 mean of $\hat{\mathbf{a}}_i$,

$$\hat{\boldsymbol{\beta}}_{3\pi,y\cdot(u,a)} = \left(\sum_{i\in A_3} \mathbf{r}'_{3,ua,i}\pi_{3i}^{-1}\mathbf{r}_{3,ua,i}\right)^{-1} \sum_{i\in A_3} \mathbf{r}'_{3,ua,i}\pi_{3i}^{-1}(y_i - \bar{y}_{3\pi})$$

and $\mathbf{r}_{3,ua,i} = (\mathbf{u}_i - \bar{\mathbf{u}}_{3\pi},\ \hat{\mathbf{a}}_i - \bar{\mathbf{a}}_{r,3\pi})$. In the form (3.3.40) containing the deviation, $\hat{\mathbf{a}}_i$, it can be seen that y must be correlated with the part of $\mathbf{x}$ that is orthogonal to $\mathbf{u}$ in order for phase 2 to add information beyond that contained in phase 1.

The variance of the three-phase estimator can be obtained by repeated application of conditional expectation arguments. We assume the existence of moments for the estimators. We also assume a design such that the bias in $\bar{y}_{3p,reg}$ is $O(n^{-1})$ and the variance of means is $O(n^{-1})$. Then the variance of $\bar{y}_{3p,reg}$ for a fixed A_1 is the variance of a two-phase sample estimator of $\bar{y}_{1\pi}$;

$$\begin{aligned} V\{(\bar{y}_{3p,reg} - \bar{y}_{1\pi}) \mid (A_1,\ \mathcal{F})\} \\ &= V\{(\bar{y}_{2p,reg} - \bar{y}_{1\pi}) \mid (A_1,\ \mathcal{F})\} \\ &\quad + E\{V\left[\bar{y}_{3p,reg} \mid (A_2,\ A_1,\ \mathcal{F})\right] \mid (A_1,\ \mathcal{F})\} \\ &\quad + O_p(n_3^{-2}), \end{aligned}$$

where

$$E\{\bar{y}_{3\pi,reg} \mid (A_2,\ A_1,\ \mathcal{F})\} = \bar{y}_{2p,reg} + O_p(n_3^{-1})$$
$$\bar{y}_{2p,reg} = \bar{y}_{2\pi} + (\bar{\mathbf{u}}_{1\pi} - \bar{\mathbf{u}}_{2\pi})\hat{\boldsymbol{\beta}}_{2\pi,y\cdot u}$$

and $\hat{\boldsymbol{\beta}}_{2\pi,y\cdot u}$ is defined by analogy to $\hat{\boldsymbol{\beta}}_{2\pi,x\cdot u}$ of (3.3.37). It follows that

$$\begin{aligned} V\{(\bar{y}_{3p,reg} - \bar{y}_N) \mid \mathcal{F}\} &= V\{E[(\bar{y}_{2p,reg} - \bar{y}_N) \mid (A_1,\ \mathcal{F})] \mid \mathcal{F}\} \\ &+ E\{V[(\bar{y}_{2p,reg} - \bar{y}_{1\pi}) \mid (A_1,\ \mathcal{F})] \mid \mathcal{F}\} \\ &+ E[E\{V[(\bar{y}_{3p,reg} - \bar{y}_{2p,reg}) \mid (A_2,\ A_1,\ \mathcal{F})] \mid (A_1,\ \mathcal{F})\} \mid \mathcal{F}] \\ &+ O_p(n_3^{-2}), \end{aligned} \tag{3.3.41}$$

where

$$E\{\bar{y}_{2p,reg} \mid (A_1,\ \mathcal{F})\} = \bar{y}_{1\pi}.$$

The first two terms in the variance expression are often difficult to estimate, for reasons given in the discussion of two-phase estimation in Section 3.3.1.

To obtain a second representation for the variance, we write the Taylor approximation to the error in the estimator as

$$\begin{aligned} \bar{y}_{3p,reg} - \bar{y}_N &= \bar{y}_{3\pi} - \bar{y}_N + (\bar{\mathbf{u}}_{1\pi} - \bar{\mathbf{u}}_N,\ \bar{\mathbf{x}}_{2p,reg} - \bar{\mathbf{x}}_N)\boldsymbol{\beta}_{y\cdot(u,x),N} \\ &\quad + O_p(n_3^{-1}) \\ &= \bar{e}_{3\pi} + (\bar{\mathbf{u}}_{1\pi} - \bar{\mathbf{u}}_N,\ \bar{\mathbf{a}}_{2\pi})\boldsymbol{\beta}_{y\cdot(u,a),N} + O_p(n_3^{-1}), \end{aligned} \tag{3.3.42}$$

where $\bar{\mathbf{a}}_N = \mathbf{0}$,

$$\begin{aligned} e_i &= y_i - \bar{y}_N - (\mathbf{u}_i - \bar{\mathbf{u}}_N,\ \mathbf{x}_i - \bar{\mathbf{x}}_N)\boldsymbol{\beta}_{y\cdot(u,x),N}, \\ a_i &= \mathbf{x}_i - \bar{\mathbf{x}}_N - (\mathbf{u}_i - \bar{\mathbf{u}}_N)\boldsymbol{\beta}_{x\cdot u,N}, \end{aligned}$$

and $\boldsymbol{\beta}_{y\cdot(u,x),N}$ and $\boldsymbol{\beta}_{x\cdot u,N}$ are population regression coefficients. If $\bar{e}_{3\pi}$, $\bar{\mathbf{u}}_{1\pi}$, and $\bar{\mathbf{a}}_{2\pi}$ are uncorrelated,

$$\begin{aligned} V\{\bar{y}_{3p,reg} - \bar{y}_N \mid \mathcal{F}\} &= V\{\bar{e}_{3\pi} \mid \mathcal{F}\} + \boldsymbol{\beta}'_{y\cdot u,N} V\{\bar{\mathbf{u}}_{1\pi} \mid \mathcal{F}\}\boldsymbol{\beta}_{y\cdot u,N} \\ &\quad + \boldsymbol{\beta}'_{y\cdot a,N} V\{\bar{\mathbf{a}}_{2\pi} \mid \mathcal{F}\}\boldsymbol{\beta}_{y\cdot a,N}, \end{aligned} \tag{3.3.43}$$

where $\boldsymbol{\beta}'_{y\cdot(u,a),N} = (\boldsymbol{\beta}'_{y\cdot u,N},\ \boldsymbol{\beta}'_{y\cdot a,N})$. It is reasonable to treat the means as uncorrelated if stratified sampling is used at each phase, all stratum indicators are included in the regression vectors, and the same sampling units are used at all phases.

An important special case of the three-phase estimator occurs when the first of the three phases is the entire population. Then $\bar{\mathbf{u}}_{1\pi} = \bar{\mathbf{u}}_N$, $V\{\bar{\mathbf{u}}_{1\pi} \mid \mathcal{F}\} = \mathbf{0}$, and the variance expression is that of a two-phase sample with the exception that $V\{\bar{\mathbf{x}}_{1\pi} \mid \mathcal{F}\}$ is replaced with $V\{\bar{\mathbf{a}}_{2\pi} \mid \mathcal{F}\}$.

3.3.3 Separate samples with common characteristics

Two-phase estimation procedures can be applied in situations where the second phase is not a subsample of a phase 1 sample. Consider a situation in which two samples are selected from the same population for two different studies, but some common data are collected on the two samples. Let the observations on the common elements for sample 1 be $\mathbf{x}_1$ and let the observations on the common elements for sample 2 be $\mathbf{x}_2$. Assume that we are interested in estimating the mean of characteristic y for sample 2. The estimation problem can be viewed as an estimated generalized least squares problem. Let the estimator of the mean of $\mathbf{x}$ for sample 1 be $\bar{\mathbf{x}}_{\pi,1}$, let the estimator of the mean of $(\mathbf{x}, y)$ for sample 2 be $(\bar{\mathbf{x}}_{\pi,2}, \bar{y}_{\pi,2})$, and let $(\bar{\mathbf{x}}_{\pi,1}, \bar{\mathbf{x}}_{\pi,2}, \bar{y}_{\pi,2})' = \mathbf{u}$. If we have an estimator of the covariance matrix of $\mathbf{u}$, the estimated generalized least squares estimator of the mean vector $\boldsymbol{\theta} = (\bar{\mathbf{x}}_N, \bar{y}_N)$ is

$$\hat{\boldsymbol{\theta}}_{reg} = (\mathbf{Z}'\hat{\mathbf{V}}_{uu}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{V}}_{uu}^{-1}\mathbf{u}, \tag{3.3.44}$$

where $\hat{\boldsymbol{\theta}}_{reg} = (\bar{\mathbf{x}}_{reg}, \bar{y}_{reg})$

$$\mathbf{Z}' = \begin{pmatrix} \mathbf{I} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{pmatrix}$$

and $\hat{\mathbf{V}}_{uu}$ is the estimated covariance matrix of $\mathbf{u}$. The covariance of $\bar{\mathbf{x}}_{\pi,1}$ with $\bar{\mathbf{x}}_{\pi,2}$ depends on the nature of the two samples. For large populations and independently drawn samples, a reasonable approximation is to assume zero correlation between the two samples. Then the estimator of the vector of means is

$$\bar{\mathbf{x}}'_{reg} = (\hat{\mathbf{V}}_{uu11}^{-1} + \hat{\mathbf{V}}_{uu22}^{-1})^{-1}(\hat{\mathbf{V}}_{uu11}^{-1}\bar{\mathbf{x}}'_{\pi,1} + \hat{\mathbf{V}}_{uu22}^{-1}\bar{\mathbf{x}}'_{\pi,2}), \tag{3.3.45}$$

$$\bar{y}_{reg} = \bar{y}_{\pi,2} + (\bar{\mathbf{x}}_{reg} - \bar{\mathbf{x}}_{\pi,2})\hat{\mathbf{V}}_{uu22}^{-1}\hat{\mathbf{V}}_{uu23}, \tag{3.3.46}$$

where $\hat{\mathbf{V}}_{uuij}$ is the ij block of $\hat{V}_{uu}$ and the variance blocks for $\bar{\mathbf{x}}_{\pi,1}$, $\bar{\mathbf{x}}_{\pi,2}$, and $\bar{y}_{\pi,2}$, are identified by 11, 22, and 33, respectively. The error in $\bar{y}_{reg}$ as an estimator of $\bar{y}_N$ can be written

$$\bar{y}_{reg} - \bar{y}_N = \bar{e}_{\pi,2} + (\bar{\mathbf{x}}_{reg} - \bar{\mathbf{x}}_N)\boldsymbol{\beta} + O_p(n^{-1}).$$

Under the assumption of independent samples, variance estimator (3.3.33) is relatively easy to compute because $\hat{V}\{\bar{e}_{\pi,2} \mid \mathcal{F}\}$ can be computed from sample 2 and the estimated variance of $\bar{\mathbf{x}}_{reg}$ follows from (3.3.43). Thus, an estimator of the variance is

$$\begin{aligned} \hat{V}\{\bar{y}_{reg} \mid \mathcal{F}\} &= \hat{V}\{\bar{\mathbf{x}}_{reg}\boldsymbol{\beta} \mid \mathcal{F}\} + \hat{V}\{\bar{e}_{\pi,2} \mid \mathcal{F}\} \\ &= \hat{\boldsymbol{\beta}}'\hat{V}\{\bar{\mathbf{x}}_{reg} \mid \mathcal{F}\}\hat{\boldsymbol{\beta}} + \hat{V}\{\bar{e}_{\pi,2} \mid \mathcal{F}\}, \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = \hat{\mathbf{V}}_{uu22}^{-1}\hat{\mathbf{V}}_{uu23}$, $\hat{V}\{\bar{\mathbf{x}}_{reg} \mid \mathcal{F}\} = (\hat{\mathbf{V}}_{uu11}^{-1} + \hat{\mathbf{V}}_{uu22}^{-1})^{-1}$, $\hat{V}\{\bar{e}_{\pi,2} \mid \mathcal{F}\}$ is an estimator of the variance of the estimated mean of e_i computed with $\hat{e}_i$, and $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$.

3.3.4 Composite estimation

The generalized least squares procedure of Section 3.3.3 is appropriate in the situation where there are two or more samples and a common set of characteristics is observed for the samples. When the samples have common elements, efficiency gains are possible without observing a common set of characteristics. The estimation procedures for samples with common elements are most often applied when sampling is carried out at several time periods.

Consider the relatively simple situation of a single characteristic and simple random sampling on two occasions. Let n_{11} be the number of elements observed only at time 1, let n_{12} be the number of elements observed at both times 1 and 2, and let n_{22} be the number of elements observed only at time 2. Assuming the three samples to be independent, the covariance matrix of $(\bar{y}_{1,11}, \bar{y}_{1,12}, \bar{y}_{2,12}, \bar{y}_{2,22}) = \bar{\mathbf{y}}$ is

$$\mathbf{V}_{\bar{y}\bar{y}} = \begin{pmatrix} n_{11}^{-1} & 0 & 0 & 0 \\ 0 & n_{12}^{-1} & n_{12}^{-1}\rho & 0 \\ 0 & n_{12}^{-1}\rho & n_{12}^{-1} & 0 \\ 0 & 0 & 0 & n_{22}^{-1} \end{pmatrix} \sigma^2, \qquad (3.3.47)$$

where $y_{t,jk}$ is the time t mean of sample jk and ρ is the correlation between observations made at times 1 and 2 on the same element. Given $\mathbf{V}_{\bar{y}\bar{y}}$, the generalized least squares estimator of the vector of time means $(\bar{y}_{1,N}, \bar{y}_{2,N})' =: \boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = (\mathbf{Z}'\mathbf{V}_{\bar{y}\bar{y}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}_{\bar{y}\bar{y}}^{-1}\bar{\mathbf{y}},$$

where

$$\mathbf{Z}' = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Estimators that are a function of parts of samples observed at several points in time are often called *composite estimators*.

To consider optimizing the design for the simple example, assume that the total budget for data collection is C. Sometimes it is cheaper to obtain the second determination, but we assume equal costs for the two types of observations. To determine an optimal allocation to the three types of samples, one must have an objective function. The two time means are two unique

parameters, but it is useful to think about the second period mean and the change from time 1 to time 2. If the only parameter of interest is the change from period 1 to period 2, and $\rho > 0$, the optimal strategy is to observe all sample elements at both time periods.

If the period means are also important, the design choice is less clear. Tables 3.7 and 3.8 contain the variances of the change and of the second period mean, respectively, for samples of size 100 at each of two time points. The variances are standardized so that the variance of the mean of the sample at a time point is 1.00. The impact of correlation on the variance of a difference is clear in Table 3.7, with the variance of estimated change decreasing as ρ increases.

Table 3.7 Variance of One-Period Change as a Function of Common Elements and Correlation

	n_{12}					
ρ	0	20	40	60	80	100
0.00	2.000	2.000	2.000	2.000	2.000	2.000
0.50	2.000	1.667	1.429	1.250	1.111	1.000
0.70	2.000	1.364	1.034	0.833	0.698	0.600
0.85	2.000	0.938	0.612	0.454	0.361	0.300
0.90	2.000	0.714	0.435	0.312	0.244	0.200

If $\rho > 0$, there is some $0 < n_{12} < n_{11}$ that gives the minimum variance for the estimator of the mean. For $\rho > 0$, as n_{12} is moved away from 100, the variance of the mean goes down and the variance of the change goes up. With large correlations the decrease in the variance of the mean is smaller in percentage terms but larger in absolute value than the increase in the variance of change. Thus, defining an optimum requires specification of the relative importance of the two types of estimators. In a survey conducted over a long period of time, such as a labor force survey, the number of times a sample person is asked to respond is an important determinant for the amount of overlap in the survey design.

3.4 REJECTIVE SAMPLING

We have discussed several methods of using auxiliary information to define and select samples so that the samples have desirable properties. The method of rejective sampling, introduced in Section 1.2.6, is a method of removing "undesirable" samples from a set with generally good properties. In fact, it is possible to design a sample-estimation procedure that rejects such samples

Table 3.8 Variance of Mean as a Function of Common Elements and Correlation

	n_{12}					
ρ	0	20	40	60	80	100
0.00	1.000	1.000	1.000	1.000	1.000	1.000
0.50	1.000	0.952	0.934	0.938	0.960	1.000
0.70	1.000	0.886	0.857	0.872	0.920	1.000
0.85	1.000	0.795	0.766	0.804	0.881	1.000
0.90	1.000	0.731	0.726	0.777	0.866	1.000

and is design consistent. The properties of the procedure depend on the method of initial selection, the nature of the rejection rule, the estimator, and the parameter being estimated.

A sample selection procedure related closely to the rejective procedure we describe is that of Deville and Tillé (2004), called by them *balanced sampling*. The Deville-Tillé sampling procedure uses an algorithm that attempts to select samples with selection probabilities close to prescribed selection probabilities and with the sample mean of the vector of auxiliary variables close to the population mean.

To study the design properties of a rejective procedure, let the design model for the population be

$$\begin{aligned} y_i &= \beta_0 + x_i\beta_1 + e_i, \\ e_i &\sim (0, \sigma^2), \end{aligned} \tag{3.4.1}$$

where x_i is the auxiliary variable observed at the design stage. Given a simple random sample of size n selected from a finite population of size N, a regression estimator for the mean of y is

$$\bar{y}_{reg} = \bar{y} + (\bar{x}_N - \bar{x})\hat{\beta}_1, \tag{3.4.2}$$

where

$$\hat{\beta}_1 = \left(\sum_{i\in A}(x_i - \bar{x})^2\right)^{-1} \sum_{i\in A}(x_i - \bar{x})(y_i - \bar{y}),$$

and the conditional variance, conditional on $\mathbf{X}' = (x_1', x_2', \ldots, x_n')$ and $\bar{x}_N$, is

$$V\{\bar{y}_{reg} - \bar{y}_N \mid \mathbf{X}, \bar{x}_N\} = \left[n^{-1} + (\bar{x}_N - \bar{x})^2 \left(\sum_{i\in A}(x_i - \bar{x})^2\right)^{-1}\right] \sigma^2. \tag{3.4.3}$$

The variance in (3.4.3) is minimized if $(\bar{x}_N - \bar{x})^2 = 0$. Thus, a nearly minimum conditional variance can be obtained if we can select a sample with $\bar{x}$ close to $\bar{x}_N$.

One approach is to select simple random samples rejecting the samples until one is obtained with

$$Q_{p,n} = (\bar{x}_p - \bar{x}_N)^2 V_{\bar{x}\bar{x}}^{-1} < \gamma^2, \tag{3.4.4}$$

where $\bar{x}_p$ is used to denote the mean of the simple random sample, γ is a specified value, and $V_{\bar{x}\bar{x}} = N^{-1}(N-n)n^{-1}S_{x,N}^2$ is the variance of $\bar{x}_p$. Let

$$\bar{y}_{rej} = n^{-1} \sum_{i \in A_{rej}} y_i$$

be the mean of the sample meeting the criterion, where A_{rej} is the set of indices for the sample selected. The sample that meets the criterion of (3.4.4) is sometimes called the *rejective sample*, the reference being to the procedure. From our introduction to rejective sampling in Section 1.2.6, and by the discussion in Section 2.5, we know that the inclusion probabilities for the rejective sample are not equal to the original selection probabilities. However, it is possible to show that the regression estimator of $\bar{y}_N$ calculated with the rejective sample has the same limiting variance as the regression estimator for the simple random sample. We give the argument for a sequence of simple random samples from a sequence of finite populations of size N generated by model (3.4.1). We assume a constant sampling rate. For such samples, $\bar{x}_p - \bar{x}_N$ is nearly normally distributed and $Q_{p,n}$ is approximately distributed as a chi-square random variable with 1 degree of freedom.

With no loss of generality, let $\bar{x}_N = 0$ and $S_x^2 = 1$, where S_x^2 is the finite population variance of x. Then

$$E\{\bar{x}_p \mid \mathcal{F}_N, i \in A\} = n^{-1}\phi x_i \tag{3.4.5}$$

and

$$V\{\bar{x}_p \mid \mathcal{F}_N, i \in A\} = n^{-2}(n-1)\phi[1 - N^{-1}(x_i^2 - 1)] + O_p(n^{-2}), \tag{3.4.6}$$

where $\phi = (N-1)^{-1}(N-n)$. It follows that

$$E\{Q_{p,n} \mid \mathcal{F}_N, i \in A\} = 1 - \zeta_{N,i} + o_p(n^{-1}),$$

where $\zeta_{N,i} = n^{-1} - 2N^{-1} + (N^{-1} - n^{-1}\phi)x_i^2$. The conditional distribution of $Q_{p,n}$, given $i \in A$, is approximately that of a multiple of a noncentral chi-square random variable with noncentrality parameter $n^{-2}\phi^2 V_{\bar{x}\bar{x}}^{-1} x_i^2$. Thus, $(1-\zeta_{N,i})^{-1}Q_{p,n}$ is approximately distributed as a chi-square random variable

with 1 degree of freedom. For the noncentral chi-square, the remainder in the approximation is $O(n^{-2})$ for a noncentrality parameter that is $O(n^{-1})$. See Johnson and Kotz (1970) and Cox and Reid (1987). Therefore, we can write

$$\begin{aligned} G_{N(i)}(\gamma^2) &= G_N\{(1+\zeta_{N,i})\gamma^2\} + o_p(n^{-1}) \\ &= G_N(\gamma^2) + \gamma^2 g_{1N}(\gamma^2)\zeta_{N,i} + o_p(n^{-1}), \end{aligned} \tag{3.4.7}$$

where $G_{N(i)}(\gamma^2)$ is the conditional distribution of $Q_{p,n}$ given $i \in A$, $G_N(\gamma^2)$ is the chi-square distribution evaluated at γ^2, and $g_{1N}(\gamma^2)$ is the chi-square density evaluated at γ^2.

Approximation (3.4.7) can be used to obtain an approximation for the inclusion probability by writing the probability that element i is included in the rejective sample as

$$\pi_{i,N} = \frac{P\{(\bar{x}_p - \bar{x}_N)'V_{\bar{x}\bar{x}}^{-1}(\bar{x}_p - \bar{x}_N) < \gamma^2 \mid \mathcal{F}_N, i \in A\}P(i \in A \mid \mathcal{F}_N)}{P\{(\bar{x}_p - \bar{x}_N)'V_{\bar{x}\bar{x}}^{-1}(\bar{x}_p - \bar{x}_N) < \gamma^2 \mid \mathcal{F}_N\}}. \tag{3.4.8}$$

Thus, for a simple random sample from a finite population of size N generated as a realization of $NI(0,1)$ random variables, the probability that element i is included in the rejective sample is

$$\pi_{i,N} = N^{-1}n\{1 + C_{1\gamma N}\zeta_{N,i}\} + o_p(n^{-1}), \tag{3.4.9}$$

where $C_{1\gamma N} = \{G_N(\gamma^2)\}^{-1}g_{1N}(\gamma^2)\gamma^2$. By similar arguments,

$$E\{Q_{p,n} \mid \mathcal{F}_N, (i,j) \in A\} = 1 + \zeta_{N,i} + \zeta_{N,j} + 2\phi n^{-1}x_i x_j + o_p(n^{-1})$$

and the joint selection probability for the rejective sample is

$$\pi_{ij,N} = p_{ij,N}\{1 + C_{1\gamma N}(\zeta_{N,i} + \zeta_{N,j} + 2\phi n^{-1}x_i x_j) + o_p(n^{-1}), \tag{3.4.10}$$

where $p_{ij,N} = N^{-1}n(N-1)^{-1}(n-1)$ for simple random sampling.

Now $\zeta_{N,i} = o_p(n^{-1})$ and it follows from (3.4.9) that

$$E\{\bar{y}_{rej} \mid \mathcal{F}_N\} = N^{-1}\sum_{i \in U}\pi_{i,N}p_i^{-1}y_i = \bar{y}_N + O_p(n^{-1}),$$

and $p_i = N^{-1}n$ for a simple random sample. By (3.4.9) and (3.4.10), the variance of $\bar{y}_{rej}$ is

$$\begin{aligned} V\{\bar{y}_{rej} \mid \mathcal{F}_N\} &= \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij,N} - \pi_{i,N}\pi_{j,N})n^{-2}y_i y_j \\ &= N^{-1}(N-n)n^{-1}S_y^2 - K_c\sum_{i=1}^{N}\sum_{j=1}^{N}x_i y_i x_j y_j + o_p(n^{-1}) \\ &= O_p(n^{-1}), \end{aligned} \tag{3.4.11}$$

where $K_c = 2\phi N^{-1}(N-1)^{-1}n^{-2}(n-1)C_{1\gamma_N}$ and $S_y^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2$.

Let $\hat{\beta}_{1,reg}$ be $\hat{\beta}_1$ of (3.4.2) calculated from the rejective sample, and let $\bar{y}_{reg,rej}$ be the corresponding regression estimator. By (3.4.9) and (3.4.10) and the development of (3.4.11),

$$\hat{\beta}_{1,rej} - \beta_1 = O_p(n^{-1/2})$$

and

$$\bar{y}_{reg,rej} = \bar{y}_{rej} + (\bar{x}_N - \bar{x}_{rej})\hat{\beta}_1 = \bar{e}_{rej} + O_p(n^{-1}).$$

Therefore, the variance of the approximate distribution of $\bar{y}_{reg,rej}$ is the variance of $\bar{e}_{rej}$. By substituting the probabilities into the Horvitz–Thompson variance expression for $\bar{e}_{rej}$, we have

$$\begin{aligned} V\{\bar{e}_{rej} \mid \mathcal{F}_N\} &= \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij,N} - \pi_{i,N}\pi_{j,N})w_i e_i w_j e_j \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}(p_{ij,N} - p_i p_i)w_i e_i w_j e_j \\ &\quad + K_c \sum_{i=1}^{N}\sum_{j=1}^{N} n^{-1}x_i x_j e_i e_j + o_p(n^{-1}) \\ &= V(\bar{e}_p \mid \mathcal{F}_N) + o_p(n^{-1}) \end{aligned} \tag{3.4.12}$$

because $\sum_{i=1}^{N} x_i e_i = 0$ by the properties of regression residuals. Thus, the large-sample variance of the regression estimator for the rejective sample is the large-sample variance of the regression estimator for the basic procedure.

If the inclusion probabilities for the rejective sample and the e_i were known, the Horvitz–Thompson estimator of the variance for $\bar{e}_{rej}$ would be

$$\tilde{V}_{HT}\{\bar{e}_{rej} \mid \mathcal{F}_N\} = \sum_{i\in A_{rej}}\sum_{j\in A_{rej}} \pi_{ij,N}^{-1}(\pi_{ij,N} - \pi_{i,N}\pi_{j,N})w_i e_i w_j e_j, \tag{3.4.13}$$

where $\pi_{i,N}$ is the true inclusion probability for the rejective sample and $\pi_{ij,N}$ is the true joint inclusion probability. It can be proven that

$$\begin{aligned} \hat{V}_{HT}\{\bar{e}_{rej} \mid \mathcal{F}_N\} &= \sum_{i\in A_{rej}}\sum_{j\in A_{rej}} p_{ij,N}^{-1}(p_{ij,N} - p_i p_j)w_i \hat{e}_i w_j \hat{e}_j, \\ &= \tilde{V}_{HT}(\bar{e}_{rej} \mid \mathcal{F}_N) + o_p(n^{-1}), \end{aligned} \tag{3.4.14}$$

where $\hat{e}_i = y_i - \bar{y}_{rej} - (x_i - \bar{x}_{rej})\hat{\beta}_{1,rej}$.

To consider rejection in a more general setting, assume that a vector of auxiliary variables is available at the design stage. Let the probability rule used to select initial samples be called the *basic procedure*. We restrict our discussion to Poisson sampling and stratified sampling. Let the sample be rejected unless

$$Q_p = (\bar{\mathbf{x}}_p - \bar{\mathbf{x}}_N)\mathbf{H}_N^{-1}(\bar{\mathbf{x}}_p - \bar{\mathbf{x}}_N)' < \gamma^2, \tag{3.4.15}$$

where $\gamma^2 > 0$, $\mathbf{H}_N$ is a positive definite symmetric matrix, $\mathbf{x}_i$ is a vector of auxiliary variables,

$$\bar{\mathbf{x}}_p = \sum_{i \in A} w_i \mathbf{x}_i,$$

$w_i = N^{-1}p_i^{-1}$, p_i is the selection probability for the basic procedure, and A is the set of indexes in a sample selected by the basic procedure. An obvious choice for $\mathbf{H}_N$ is $\mathbf{V}_{\bar{x}\bar{x}}$, where $\mathbf{V}_{\bar{x}\bar{x}} = V\{\bar{\mathbf{x}}_p\}$ is the positive definite variance of $\bar{\mathbf{x}}_p$, but the matrix $\mathbf{H}_N$ permits one to impose a tighter restriction on some variables than on others.

Let the estimator of the population mean of y be the regression estimator

$$\bar{y}_{reg,rej} = \bar{y}_{rej} + (\bar{\mathbf{z}}_N - \bar{\mathbf{z}}_{rej})\hat{\boldsymbol{\beta}}, \tag{3.4.16}$$

where $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{z}_{2i})$ is a vector containing design variables,

$$(\bar{\mathbf{z}}_{rej}, \bar{y}_{rej}) = \sum_{i \in A_{rej}} w_i(\mathbf{z}_i, y_i),$$

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in A_{rej}} \mathbf{z}_i' \phi_i p_i^{-2} \mathbf{z}_i \right)^{-1} \sum_{i \in A_{rej}} \mathbf{z}_i' \phi_i p_i^{-2} y_i,$$

A_{rej} is the set of elements in the rejective sample, $\phi_i = (1 - p_i)$ for Poisson sampling, $\phi_i = (N_h - 1)^{-1}(N_h - n_h)$ for stratified sampling, and h is the index for strata. The $\mathbf{z}$ can contain variables not used in the design. Also, one need not include all x-variables in the estimator, but all are required for a formally correct variance estimator. For example, one might include the determinant of $\mathbf{X}'\mathbf{X}$ in the rejection criterion to guarantee a positive definite matrix for regression estimation, but not use the variable in the regression estimation of the mean.

We assume that the basic procedure and sequence of finite populations are such that

$$V\{(\bar{\mathbf{x}}_p, \bar{y}_p) \mid \mathcal{F}_N\} = O(n^{-1}) \quad \text{a.s.} \tag{3.4.17}$$

Under mild conditions on the sequence of designs and populations, the estimator $\hat{\boldsymbol{\beta}}$ of (3.4.16) computed for the basic sample will converge to

$$\boldsymbol{\beta}_N = \left[\sum_{i=1}^{N} \mathbf{z}_i' \phi_i p_i^{-1} \mathbf{z}_i\right]^{-1} \sum_{i=1}^{N} \mathbf{z}_i' \phi_i p_i^{-1} y_i. \tag{3.4.18}$$

For stratified sampling and Poisson sampling, the regression estimator with a $\hat{\boldsymbol{\beta}}$ converging to $\boldsymbol{\beta}_N$ gives nearly the minimum large-sample design variance. See Theorem 2.2.3. For the regression estimator to be design consistent under the basic design, we assume that there is a vector $\mathbf{c}$ such that

$$\phi_i p_i^{-2} \mathbf{z}_i \mathbf{c} = p_i^{-1} \tag{3.4.19}$$

for all i.

The vector $\mathbf{x}_i$ of (3.4.15) is a subset of $\mathbf{z}_i$ for which $V_{\bar{x}\bar{x}}$ is nonsingular. For example, if the basic procedure is stratified sampling, the vector $\mathbf{z}_i$ contains stratum indicator variables and other auxiliary variables. The sample mean of the stratum indicator variable,

$$\begin{aligned} \psi_{hi} &= 1 \quad \text{if element } i \text{ is in stratum } h \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

is the population fraction of elements in stratum h for all samples. Therefore, ψ_{hi} is in $\mathbf{z}_i$ but not in $\mathbf{x}_i$. If this is a fixed-size design, the unit element does not appear in $\mathbf{x}_i$. However, the unit element can appear in $\mathbf{x}_i$ for random-size designs such as Poisson sampling.

The large-sample variance of the regression estimator for samples selected by the basic procedure is the large-sample variance of $\bar{e}_p$, where $e_i = y_i - \mathbf{z}_i \boldsymbol{\beta}_N$, $\bar{e}_p$ is defined by analogy to $\bar{\mathbf{x}}_p$ of (3.4.4), and $\boldsymbol{\beta}_N$ is defined in (3.4.18). Furthermore, the variance of the regression estimator can be estimated with the usual variance estimator using the selection probabilities of the basic procedure. See Fuller (2009).

Example 3.4.1. We use some generated data to illustrate the use of rejective sampling. Assume that our design model for the population of Table 3.9 is

$$\begin{aligned} y_i &= \beta_0 + x_i \beta_1 + e_i, \\ e_i &\sim (0, x_i^2 \sigma^2), \end{aligned}$$

where x_i is proportional to the probabilities p_i in the table.

Assume that we desire a sample of size 20 but we are willing to accept a sample as small as 18 or as large as 22. Then Poisson rejective sampling becomes an option. We can use a rejection rule based on p_i alone or a rule

Table 3.9 Selection Probabilities for a Population of Size 100

Number of Elements	Selection Probability	Number of Elements	Selection Probability
35	0.05714	7	0.28571
15	0.13333	5	0.40000
11	0.18182	4	0.50000
9	0.22222	6	0.66667
8	0.25000		

based on a larger vector. If we use p_i alone, we could reject any sample where $n < 18$ or $n > 22$.

If we define an x-variable to be p_i, the estimated total for p_i is

$$\sum_{i \in A} p_i^{-1} p_i = n,$$

the realized sample size. The standard deviation for the number in the sample is 1.5341. Therefore, rejecting a sample with $\mid n - 20 \mid \geq 3$ is equivalent to rejecting a sample with

$$(1.5341)^{-2}(n - 20)^2 \geq 0.6544.$$

For the estimator (3.4.5) to be consistent, $(1-p_i)^{-1}p_i$ must be in the column space of the matrix of auxiliary variables. We can reject using only sample size and construct an estimator with a $\mathbf{z}$ vector that contains $(1-p_i)^{-1}p_i$, or we can use a rejection rule based on a larger vector. We consider a rejection rule based on the vector $\mathbf{x}_i = [1, 10p_i, (1-p_i)^{-1}p_i]$, where $\bar{\mathbf{x}}_N = (1, 2, 0.34244)$. The covariance matrix of

$$\bar{\mathbf{x}}_{HT} = N^{-1} \sum_{i \in A} p_i^{-1}[1, 10p_i, p_i(1 - p_i)^{-1}]$$

for a Poisson sample based on the probabilities of the table is

$$\mathbf{V}_{\bar{x}\bar{x}} = V\{\bar{\mathbf{x}}_{HT} \mid \mathcal{F}\} = \begin{pmatrix} 4.716 & -1.790 & -1.011 \\ -1.790 & 2.353 & 1.353 \\ -1.011 & 1.353 & 1.241 \end{pmatrix} \times 10^2.$$

We transform the vector so that the estimated means are uncorrelated, letting

$$(z_1, z_2, z_3) = [10p_i,\ 1 + 5.0015p_i,\ p_i(1 - p_i)^{-1} + 0.0283 - 1.8823p_i].$$

Then the covariance matrix of $\bar{\mathbf{z}}_{HT}$ is

$$\begin{aligned}\mathbf{V}_{\bar{z}\bar{z}} &= \text{diag}\{(13.753, 3.183, 0.028) \times 10^{-2}\} \\ &=: \text{diag}\{v_{\bar{z}\bar{z}11}, v_{\bar{z}\bar{z}22}, v_{\bar{z}\bar{z}33}\},\end{aligned}$$

where

$$\bar{\mathbf{z}}_{HT} = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{z}_i.$$

If we reject the sample when

$$(\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N)\mathbf{D}_W \mathbf{V}_{\bar{z}\bar{z}}^{-1} (\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N)' > 0.6544,$$

where $\mathbf{D}_W = \text{diag}(1.0,\ 0.5,\ 0.5)$, then $|n < 20| \leq 2$ for all accepted samples and there are no very large deviations in $\bar{z}_2$ or $\bar{z}_3$ for the samples accepted. ■■

3.5 REFERENCES

Section 3.1. Bankier (1988), Brewer (1963b), Cassell, Särndal, and Wretman (1976), Fuller and Isaki (1981), Godambe and Joshi (1965), Godfrey, Roshwalb, and Wright (1984), Isaki (1970), Isaki and Fuller (1982), Kott (1985), Lavallée and Hidiroglou (1987), Nedyalkova and Tillé (2008), Park (2002), Rao and Nigam (1992).

Section 3.2. Cochran (1977), Jessen (1978).

Section 3.3. Chen and Rao (2007), Fay (1989, 1996), Fuller (1998), Hidiroglou (2001), Hidiroglou and Särndal (1998), Kim, Navarro, and Fuller (2006), Kott (1990b), Kott and Stukel (1997), Legg and Fuller (2009), Neyman (1938), Särndal, Swensson, and Wretman (1992), Rao (1973).

Section 3.4. Deville and Tillé (2004, 2005), Fuller (2009), Hájek (1964, 1981), Herson (1976), Legg and Yu (2009).

3.6 EXERCISES

1. (Section 3.2) Assume that a finite population is composed of N clusters of elements of size M_i. Assume that the elements of the population

were generated by the model

$$Y_{ij} = \mu + a_i + u_{ij},$$

where $a_i \sim NI(0,\ \sigma_a^2)$, $u_{ij} \sim NI(0,\ \sigma_u^2)$, and a_i is independent of u_{jk} for all i, j, and k.

(a) Assume that a sample of n elements is obtained by selecting a replacement sample of size n clusters with probabilities proportional to M_i, where M_i is the number of elements in cluster i. When a cluster is selected, an element is selected from the M_i by simple random sampling. What is the variance of the simple sample mean as an estimator of the finite population mean conditional on the finite population? What is the expected value of the variance under the model?

(b) Assume that a nonreplacement sample of size 2 clusters is selected with probabilities $\pi_i = 2M_i(\Sigma_{j=1}^{N} M_j)^{-1}$ and joint probability π_{ij} defined by (1.4.3). Assume that an element is selected from the cluster as in part (a). What is the variance of the weighted sample mean (ratio estimator) as an estimator of the finite population mean conditional on the finite population? What is the expected value of the design variance under the model?

2. (Section 3.3) Show that $\hat{S}_y^2$ of (3.3.10) is a consistent estimator of S_y^2. Give the conditions on your sequence of populations and estimators.

3. (Section 3.2) Compare, under model (3.2.1),

$$E\left\{\pi_i^{-2} M_i^2 \left(m_i^{-1} - M_i^{-1}\right) \sigma_e^2\right\}$$

for the sampling schemes of PPS fixed take and equal-probability proportional take.

4. (Section 3.2) Let a geographic area be divided into sections each of size 640 acres. Let each section be divided into four quarter-sections each of 160 acres. Call the quarter-sections segments. Using Table 3.10,

Table 3.10 Population Analysis of Variance

Source	d.f.	E.M.S.
Sections	$N-1$	$\sigma_e^2 + 4\sigma_u^2$
Segments/sections	$3(N-1)$	σ_e^2

compare the following two sampling schemes. Do not ignore the stage 2 finite population correction.

Scheme 1: Select a simple random sample of n segments.

Scheme 2: Select a simple random sample of n sections and select one segment in each.

5. (Section 3.2) Let the observations in Table 3.11 be for a two-stage sample composed of three stage 1 units and two stage 2 units per stage 1 unit. The sample was selected from a population of size 25 with probability proportional to size where the total size is 200.

Table 3.11 Two-Stage Sample

PSU	SSU	Size	y_{ij}	x_{ij}
1	1	10	16	18
	2		12	12
2	1	40	33	41
	2		39	47
3	1	90	105	94
	2		115	100

Let $\pi_{12}\pi_1^{-1}\pi_2^{-1} = 0.92$, $\pi_{13}\pi_1^{-1}\pi_3^{-1} = 0.98$, and $\pi_{23}\pi_2^{-1}\pi_3^{-1} = 0.95$. Estimate the total of y and estimate the variance of your estimator. Estimate the ratio of the total of y to the total of x and estimate the variance of your estimator. You may use large-sample approximations.

6. (Section 3.1) In Example 3.1.2 the simple stratified estimator was used. Construct Table 3.3 for the regression estimator with a vector $\mathbf{x}_i$ containing the order-of-x variable i of (3.1.32) and dummy variables for strata. Assume that the data are generated by model (3.1.32). Use the approximation

$$\hat{V}\{\bar{y}_{reg}\} = n^{-1}(n - H - 1)^{-1} \sum_{i \in A} \hat{e}_i^2,$$

where $\hat{e}_i$ is the regression residual.

7. (Section 3.1) Let finite populations of 100 elements be realizations of the stationary autoregressive process satisfying

$$y_t = \rho y_{t-1} + e_t,$$
$$e_t \sim NI(0, \sigma^2).$$

(a) Calculate the unconditional variance of the sample mean for an equal-probability systematic sample of size 10.

(b) Calculate the unconditional variance of the stratified mean for a one-per-stratum sample selected from 10 equal-sized strata.

8. (Section 3.1) In the controlled two-per-stratum design for four strata described in Section 3.1.3, two groups have two elements and four groups have one element. For equal-sized strata of size m, show that the joint probabilities of selection are $[2m(m-1)]^{-1}$ for elements in the same group of a stratum, $(2m^2)^{-1}$ for elements in different groups of the same stratum, $5(6m^2)^{-1}$ for elements in the same group of different strata, and $7(6m^2)^{-1}$ for elements in different groups of different strata.

9. (Section 3.3) Assume that we have a stratified phase 1 sample that is restratified for phase 2 sample selection. Let the two-phase estimator be the regression estimator, where the phase 2 regression contains indicators for both phase 1 and 2 strata. Let

$$\hat{V}\{\hat{T}_{e,1} \mid \mathcal{F}\} = \sum_{h=1}^{H} N_h^2 n_{1h}^{-1} \hat{s}_{e1h}^2$$

and

$$\hat{V}\{\hat{T}_{e,2} \mid (A_1, \mathcal{F})\} = \sum_{g=1}^{G} (1 - n_{1g}^{-1} n_{2g}) n_{2g}^{-2} \sum_{i \in A_{2g}} w_{2i}^2 \hat{e}_i^2,$$

where

$$\tilde{s}_{e1h}^2 = n_{1h}^{-1} \sum_{i \in A_{1h} \cap A_2} w_{2i|1i} \hat{e}_i^2$$

is an estimator of the phase 1 variance for stratum h, n_{1h} is the number of elements in phase 1 stratum h, A_{1h} is the set of indexes of elements in phase 1 stratum h, $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_2$, and

$$\hat{\boldsymbol{\beta}}_2 = \left(\sum_{i \in A_2} \mathbf{x}_i' \pi_{2i|1i}^{-1} \pi_{1i}^{-1} \mathbf{x}_i \right)^{-1} \sum_{i \in A_2} \mathbf{x}_i' \pi_{2i|1i}^{-1} \pi_{1i}^{-1} y_i.$$

Ignoring the phase 1 finite population correction, show that

$$\hat{V}\{\hat{T}_{e,1} \mid \mathcal{F}\} + \hat{V}\{\hat{T}_{e,2} \mid (A_1, \mathcal{F})\} = \hat{V}\{\hat{T}_{e,2} \mid \mathcal{F}\},$$

where

$$\hat{V}\{\hat{T}_{e,2} \mid \mathcal{F}\} = \sum_{i \in A_2} w_{2i}^2 \hat{e}_i^2 = \sum_{i \in A_2} w_{1i}^2 w_{2i|1i}^2 \hat{e}_i^2.$$

10. (Section 3.1) In Theorem 3.1.1, $e_i \sim ind(0,\ \gamma_{ii}\sigma^2)$. Isaki and Fuller (1982) made the weaker assumptions

$$\begin{aligned} E\{e_i\} &= 0, \\ E\{e_i e_j\} &= \gamma_{ii}\sigma^2, \qquad i = j \\ &= \rho\gamma_{ii}^{0.5}\gamma_{jj}^{0.5}\sigma^2, \quad i \neq j, \end{aligned}$$

where $-(N-1)^{-1} < \rho < 1$. Let $\gamma_{ii}^{0.5}$ be an element of $\mathbf{x}_i$. Show that the best model unbiased linear predictor of $\bar{y}_N$ conditional on $\mathbf{X}$ is

$$\bar{y}_{pred} = f_N\bar{y}_n + (1 - f_N)\bar{x}_{N-n}\hat{\boldsymbol{\beta}},$$

where $f_N = N^{-1}n$ and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{D}_\gamma^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\gamma^{-1}\mathbf{y}.$$

If γ_{ii} and $\gamma_{ii}^{0.5}$ are elements of $\mathbf{x}_i$, show that the predictor can be written as

$$\bar{y}_{pred} = \bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}.$$

Show that the result (3.1.13) of Theorem 3.1.1 holds under the weaker conditions.

11. (Section 3.1) Let a very large finite population, $\mathcal{F}$, be a realization of (x_i, y_i) random vectors, where $(x_i, y_i) \sim NI(\mathbf{0}, \boldsymbol{\Sigma})$ and $V\{y \mid x\} = \sigma^2$. Let a simple random sample of size n be selected from $\mathcal{F}$. The unconditional variance of the regression estimator of $\bar{y}_N$ constructed with known $\bar{x}_N$ is given in (2.2.16). Assume now that the population is ordered on x and formed into H equal-sized strata. Let a stratified sample be selected, where $n = mH, m \geq 2$. What is the approximate unconditional variance of the regression estimator for the mean of y, where the regression is constructed with indicator variables for the strata? To develop the approximation, assume that $\bar{x}_{st}$ is normally distributed and that $\bar{x}_{st}$ is independent of $\sum_{h=1}^{H}\sum_{i=1}^{m}(x_{hi} - \bar{x}_h)^2$. You may also assume that

$$\frac{H(m-1)\sum_{h=1}^{H}\sum_{i=1}^{m}(x_{hi} - \bar{x}_h)^2}{E\{\sum_{h=1}^{H}\sum_{i=1}^{m}(x_{hi} - \bar{x}_h)^2\}} \dot{\sim} \chi^2_{n-H},$$

where χ^2_d is the chi-square distribution with d degrees of freedom. What is the approximate unconditional variance of the regression estimator of the mean of y that ignores the strata? To develop the approximation, you may assume that $\bar{x}_{st}$ is normally distributed. Describe any other approximations that you use.

12. (Section 3.1) We used a two-per-stratum design in Example 3.1.1. On reading the example, an investigator argued that one should take three or more observations per stratum in case there is nonresponse. To assess this argument, assume that we wish to select a sample of size 4 from a very large population. Consider two options:

    (i) Select a simple random sample of size 4. If there is nonresponse, the mean of the respondents is taken as the estimator.

    (ii) Select an equal-probability stratified sample with two per stratum. If there is nonresponse, collapse the strata and take the mean of the respondents as the estimator.

    Assume that the variance of the two-per-stratum sample is 4.0 and the variance of the simple random sample is 5.0. Assume that the probability of nonresponse is 0.15, that the finite population correction can be ignored, and that response is independent of the y-values.

    (a) Explain why, under the independence assumption, both procedures are unbiased for the population mean.

    (b) Calculate the variance for each of the two procedures for samples with 0, 1, or 2 nonrespondents. Let $\bar{y}$ be the simple mean of the respondents for all cases,

    (c) Let n be the number of respondents, and let

    $$\hat{V}\{\bar{y}\} = n^{-1}s^2$$

    for the mean of the simple random sample and for any sample with a nonrespondent, where

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

    Let

    $$V\{\bar{y}\} = n^{-1}s_w^2$$

    for a complete stratified sample, where

    $$s_w^2 = 0.5\sum_{h=1}^{2}\sum_{j=1}^{2}(y_{hj} - \bar{y}_h)^2.$$

    What is $E\{\hat{V}(\bar{y})\}$ for each procedure given that $n = 4, 3$, and 2?

    (d) Construct an unbiased estimator of $V(\bar{y})$ for the stratified procedure with $n = 3$.

This Page Intentionally Left Blank

CHAPTER 4

REPLICATION VARIANCE ESTIMATION

4.1 INTRODUCTION

In our discussion of estimation, we have presented variance estimation formulas for a number of estimators. Essentially all of these formulas are quadratic functions of the observations and most can be expressed as a weighted sum of squares and products of the original observations. Some expressions that are relatively simple require a great deal of computation for a large-scale survey. For example, an estimator of the variance of the regression estimator for a simple random sample can be written

$$\hat{V}\{\bar{y}_{reg}\} = [n(n-1)]^{-1} \sum_{i=1}^{n} [y_i - \bar{y} - (x_i - \bar{x})\hat{\beta}]^2. \quad (4.1.1)$$

To compute this estimator for a 10×10 table requires defining 100 y-variables,

$$\begin{aligned} y_{Ci} &= y_i \quad \text{if } i \text{ is in cell } C \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

Sampling Statistics. By Wayne A. Fuller

computing the regression coefficient for each y-variable, and computing the estimator (4.1.1) using the deviations for the respective y-variable. Because of the computations required for this and similar variance estimators, other methods have been developed for variance calculations.

4.2 JACKKNIFE VARIANCE ESTIMATION

4.2.1 Introduction

We call a sample created from the original sample by deleting some elements or (and) changing the weight of some elements, a replicate sample. The use of a set of such replicate samples to estimate variances is now established practice. Quenouille (1949) noted that the average of samples with one unit deleted could be used to reduce the bias in a nonlinear estimator, and Tukey (1958) suggested that deleted samples could also be used for variance estimation.

Let the mean of $n-1$ units remaining after unit k is deleted from a simple random sample be denoted by $\bar{x}^{(k)}$, where

$$\bar{x}^{(k)} = (n-1)^{-1}\left(\sum_{j=1}^{n} x_j - x_k\right).$$

Note that the mean of the n samples of size $n-1$ is the original mean,

$$\bar{x} = n^{-1}\sum_{k=1}^{n} \bar{x}^{(k)}.$$

The difference between the mean of a simple random sample of size n with x_k deleted and the mean of the total sample is

$$d_k = \bar{x}^{(k)} - \bar{x} = n^{-1}(\bar{x}^{(k)} - x_k) \tag{4.2.1}$$

$$= -(n-1)^{-1}(x_k - \bar{x}). \tag{4.2.2}$$

If x_k, $k = 1, 2, \ldots, n$, is a simple random sample from an infinite population with variance σ^2, then by (4.2.1),

$$E\{d_k^2\} = n^{-2}\left[(n-1)^{-1} + 1\right]\sigma^2 = [n(n-1)]^{-1}\sigma^2,$$

because $\bar{x}^{(k)}$ is independent of x_k. Also, by (4.2.2),

$$\sum_{k=1}^{n} d_k^2 = (n-1)^{-2}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= (n-1)^{-1}s^2, \tag{4.2.3}$$

where

$$s^2 = (n-1)^{-1}\sum_{j=1}^{n}(x_j - \bar{x})^2.$$

Therefore, for a simple random sample selected from an infinite population, the jackknife estimator of the variance,

$$\hat{V}_J\{\bar{x}\} = (n-1)n^{-1}\sum_{k=1}^{n} d_k^2 = \hat{V}\{\bar{x}\}, \tag{4.2.4}$$

is the unbiased estimator of the variance of the mean. If the finite population correction cannot be ignored, the expression on the right of the equality in (4.2.4) is multiplied by $(1 - N^{-1}n)$. We shall see that the jackknife variance estimator has real advantages for some complex estimators and some complex designs.

Alternative replicates can be constructed by reducing the weight of an observation without setting it equal to zero. Such replicates are preferred for samples with rare items and complex estimators and can simplify the computational form for stratified samples. As an example, consider a weighted mean in which unit k is given a weight of ψ and the remaining $n-1$ observations are given equal weight so that the sum of the weights is equal to 1. The weighted mean is

$$\bar{x}_{\psi}^{(k)} = \psi x_k + (n-1)^{-1}(1-\psi)\left(\sum_{j=1}^{n} x_j - x_k\right) \tag{4.2.5}$$

and the corresponding deviation is

$$\begin{aligned} d_{\psi k} = \bar{x}_{\psi}^{(k)} - \bar{x} &= n^{-1}(1-n\psi)(\bar{x}^{(k)} - x_k) \\ &= (n-1)^{-1}(1-n\psi)(\bar{x} - x_k). \end{aligned} \tag{4.2.6}$$

In this case,

$$E\{d_{\psi k}^2\} = (1-n\psi)^2\left[n(n-1)\right]^{-1}\sigma^2. \tag{4.2.7}$$

If $\psi = n^{-1} - \left[n^{-3}(n-1)\right]^{1/2}$, then

$$\sum_{k=1}^{n} d_{\psi k}^2 = \hat{V}\{\bar{x}\}, \tag{4.2.8}$$

where $\hat{V}\{\bar{x}\}$ is the unbiased variance estimator defined in (4.2.4). If $\psi = n^{-1} - [n^{-3}(n-1)(1-f_N)]^{1/2}$, then

$$\sum_{k=1}^{n} d_{\psi k}^2 = \hat{V}\{\bar{x} - \bar{x}_N\} = (1-f_N)\hat{V}\{\bar{x}\}, \tag{4.2.9}$$

where $f_N = N^{-1}n$.

One can also create replicates by deleting a random group of elements from the sample. Let $\bar{x}_b^{(k)}$ be the mean of $n-b$ elements obtained by randomly deleting b elements from a simple random sample of size n. Then

$$\begin{aligned} E\left\{(\bar{x}_b^{(k)} - \bar{x})^2\right\} &= E\left\{n^{-2}(n-b)^{-2}\left((n-b)\sum_{i\in B_b} x_i - b\sum_{j\notin B_b} x_j\right)^2\right\} \\ &= (n-b)^{-1}n^{-1}b\sigma^2, \end{aligned} \tag{4.2.10}$$

where B_b is the set of elements deleted. Thus, if a simple random sample of size mb is split at random into m groups of size b and if m replicates are created by deleting each group in turn,

$$\tilde{V}_{JG}\{\bar{x}\} = (m-1)m^{-1}\sum_{k=1}^{m}(\bar{x}_b^{(k)} - \bar{x})^2 \tag{4.2.11}$$

is an unbiased estimator of $V\{\bar{x}\}$. Of course, this estimator has only $m-1$ degrees of freedom.

If one is willing to accept an estimator with fewer degrees of freedom, one can choose a random subset of the replicates of any of the types considered and modify the multiplier accordingly.

The jackknife procedure can often be used to simplify the computations for more complicated functions of the data. The jackknife is appropriate for differentiable functions of sample moments. We state and prove the theorem for simple random samples and scalars, but the result extends to other designs and to vector-valued functions of vectors. See also Theorem 1.3.6, Exercise 8, and Exercise 10.

Theorem 4.2.1. Let $\{\mathcal{F}_N\} = \{y_1, y_2, \ldots, y_N\}$ be a sequence of finite populations, where the y_i are $iid(\mu_y, \sigma_y^2)$ random variables with finite $4+\eta$, $\eta > 0$, moments. Let $(y_1, y_2, \ldots, y_n)$ be a simple random sample of size n selected from the Nth population where $n \to \infty$ as $N \to \infty$. Let $g(\bar{y})$ be a continuous function of the sample mean with continuous first and second derivatives at μ_y. Then

$$k_n\sum_{k=1}^{n}[g(\bar{y}^{(k)}) - g(\bar{y})]^2 = [g'(\bar{y})]^2\hat{V}\{\bar{y}\} + O_p(n^{-2}) \tag{4.2.12}$$

$$= \left[g'(\mu_y)\right]^2 V\{\bar{y}\} + O_p(n^{-1.5}), \quad (4.2.13)$$

where $k_n = n^{-1}(n-1)$, $\hat{V}\{\bar{y}\} = n^{-1}s^2$, s^2 is defined in (4.2.3), $g'(\mu_y)$ is the derivative of $g(\cdot)$ evaluated at μ_y, and $\bar{y}^{(k)}$ is the jackknife replicate created by deleting element k.

Proof. By a Taylor expansion,

$$\begin{aligned}\sum_{k=1}^{n}\left[g(\bar{y}^{(k)}) - g(\bar{y})\right]^2 &= \sum_{k=1}^{n}\left[g(\bar{y}) + g'(\bar{y}_k^*)(\bar{y}^{(k)} - \bar{y}) - g(\bar{y})\right]^2 \\ &= \sum_{k=1}^{n}\left[g'(\bar{y}_k^*)\right]^2(\bar{y}^{(k)} - \bar{y})^2, \quad (4.2.14)\end{aligned}$$

where $\bar{y}_k^*$ is between $\bar{y}^{(k)}$ and $\bar{y}$ and $g'(\bar{y}_k^*)$ is the derivative of $g(\cdot)$ evaluated at $\bar{y}_k^*$. Given $1 > \delta > 0$, there is an n_0 such that for $n > n_0$, the probability is greater than $1 - \delta$ that $\bar{y}$ and $\bar{y}^{(k)}$, $k = 1, 2, \ldots, n$, are all in a compact set D containing μ_y as an interior point.

For $\bar{y}$ and $\bar{y}^{(k)}$ in D,

$$\left|\left[g'(\bar{y}_k^*)\right]^2 - \left[g'(\bar{y})\right]^2\right| = \left|2g'(\bar{y}_k^{**})g''(\bar{y}_k^{**})(\bar{y}_k^* - \bar{y})\right| < K_1\left|\bar{y}_k^* - \bar{y}\right|$$

for some $K_1 > 0$, where $\bar{y}_k^{**}$ is between $\bar{y}_k^*$ and $\bar{y}$, because the second derivative is continuous on D. Note that $\bar{y}_k^* - \bar{y} = O_p(n^{-1})$. It follows that for $\bar{y}$ and all $\bar{y}^{(k)}$ in D,

$$\begin{aligned}k_n\sum_{k=1}^{n}\left[g'(\bar{y}_k^*)\right]^2(\bar{y}^{(k)} - \bar{y})^2 &= k_n\sum_{k=1}^{n}\left[g'(\bar{y})\right]^2(\bar{y}^{(k)} - \bar{y})^2 + O_p(n^{-2}) \\ &= \left[g'(\bar{y})\right]^2\hat{V}\{\bar{y}\} + O_p(n^{-2}).\end{aligned}$$

See Fuller (1996, p. 226).

Also, for $\bar{y}$ in D,

$$\left|g'(\bar{y}) - g'(\mu_y)\right| < K_2\left|\bar{y} - \mu_y\right| = O_p(n^{-1/2})$$

for some $K_2 > 0$. Given that $\hat{V}\{\bar{y}\} - V\{\bar{y}\} = O_p(n^{-1.5})$, result (4.2.13) is established. ■

4.2.2 Stratified samples

The basic ideas of element deletion can be used to construct jackknife replicates for more complicated designs. For stratified sampling, replicates can be constructed by deleting a unit from each stratum in turn and applying the size correction appropriate for the stratum.

Example 4.2.1. Table 4.1 gives a set of simple jackknife weights for estimated totals for a small stratified sample. The sampling rate is one-twelfth in stratum 1 and one-twentieth in stratum 2. Thus, if the weights of Table 4.1 are used to construct five replicates,

$$\begin{aligned}\hat{V}\{\hat{T}_{y,st}\} &= \sum_{k=1}^{3} 22(36)^{-1}(\hat{T}_{y,st}^{(k)} - \hat{T}_{y,st})^2 \\ &\quad + \sum_{k=4}^{5} 19(40)^{-1}(\hat{T}_{y,st}^{(k)} - \hat{T}_{y,st})^2\end{aligned}$$

is algebraically equivalent to the usual stratified variance estimator. The multipliers are $N_h^{-1}(N_h - n_h)\, n_h^{-1}(n_h - 1)$, where h is the stratum index. Observe that the deviations for replicates 4 and 5, the replicates for the stratum with two observations, give the same square. Therefore, only one of the replicates need be used, and the appropriate multiplier becomes the finite population correction of $(40)^{-1}(38)$.

Table 4.1 Replication Weights for a Stratified Sample

Stratum	Obs.	Original Weight	Replicate 1	2	3	4	5
1	1	12	0	18	18	12	12
	2	12	18	0	18	12	12
	3	12	18	18	0	12	12
2	1	20	20	20	20	0	40
	2	20	20	20	20	40	0

In Table 4.2 the zero weight for a "deleted" element in the first stratum is replaced with

$$\psi_1 = 12\left[1 - \left\{22(36)^{-1}\right\}^{1/2}\right]$$

and the zero weight in the second stratum is replaced with

$$\psi_2 = 20\left[1 - \left\{19(40)^{-1}\right\}^{1/2}\right].$$

See equation (4.2.9). With the weights of Table 4.2,

$$\hat{V}\{\hat{T}_{y,st}\} = \sum_{k=1}^{5}(\hat{T}_{y,st}^{(k)} - \hat{T}_{y,st})^2,$$

where the k index is for the replicates of Table 4.2.

The replicates of Table 4.2 were constructed to reproduce the standard estimator of variance for an estimated total. It is possible to construct a smaller number of replicates with the same expected value but fewer degrees of freedom. For example, the linear combinations in stratum 2 of replicates 4 and 5 could be added to the linear combinations of replicates 1 and 2. Then, for example,

$$E\{(2.62y_{11} + 16.69y_{12} + 16.69y_{13} + 6.22y_{21} + 33.78y_{22} - \hat{T}_{y,st})^2\}$$
$$= 0.333V\{\hat{T}_{y,1}\} + 0.500V\{\hat{T}_{y,2}\}.$$

Table 4.2 Alternative Replication Weights for a Stratified Sample

Stratum	Obs.	Original Weight	Replicate 1	2	3	4	5
1	1	12	2.6192	16.6904	16.6904	12.0000	12.0000
	2	12	16.6904	2.6192	16.6904	12.0000	12.0000
	3	12	16.6904	16.6904	2.6192	12.0000	12.0000
2	1	20	20.0000	20.0000	20.0000	6.2160	33.7840
	2	20	20.0000	20.0000	20.0000	33.7840	6.2160

The two new replicates and replicate 3 provide an unbiased estimator of the variance of the total. This type of procedure is appropriate when one has a large number of strata and is willing to accept an estimator with fewer degrees of freedom in order to reduce the number of computations. ■■

If the total sample is large and there are a large number of primary sampling units in each stratum, one can create a relatively small set of replicates by deleting one primary sampling unit (or a small number) from each stratum. The procedure is called *delete-a-group jackknife*. We outline one of the possible ways of defining replicates. See Kott (2001) and Lu, Brick, and Sitter (2006). Assume that the stratum with the largest number of observations has n_L primary sampling units and that L is the number of replicates desired. First, the elements in each stratum are arranged in random order and numbered. If

$n_h < L$, the sampling units are assigned multiple numbers until L is reached. The elements are rerandomized for each assignment. For example, if $L = 8$ and $n_h = 3$, numbers 1, 5, and 8 might be assigned to unit 1, numbers 2, 6, and 7 assigned to unit 2, and numbers 3 and 4 assigned to unit 3. Replicate k is created by reducing the weight of the element in each stratum that has been assigned number k. Let B_k be the set of numbers assigned number k for replication purposes, and let $w_{hi} = W_h n_h^{-1}$ be the original weight for element i in stratum h for the stratified mean. Then

$$E\left\{\sum_{k=1}^{L}(\bar{y}_{st}^{(k)} - \bar{y}_{st})^2 \mid \mathcal{F}\right\} = V\{\bar{y}_{st} \mid \mathcal{F}\}, \tag{4.2.15}$$

where

$$\begin{aligned}
\bar{y}_{st}^{(k)} &= \sum_{h=1}^{H}\sum_{i\in A_h} w_{hi}^{(k)} y_{hi}, \\
w_{hi}^{(k)} &= W_h\psi_h \qquad\qquad \text{if } i\in B_k, \\
&= W_h(n_h - 1)^{-1}(1 - \psi_h) \qquad \text{if } i\notin B_k,
\end{aligned}$$

and ψ_h is the smallest root of

$$(n_h\psi_h - 1)^2 = (1 - f_h)(n_h - 1)L^{-1}.$$

See Exercise 9. If the sample is a cluster sample or a two-stage sample, primary sampling units are deleted to create the jackknife replicates.

In many cases the jackknife variance estimator gives a larger estimated variance for a nonlinear function of means than the traditional Taylor variance estimator. The reason for this can be illustrated with the ratio estimator. Let

$$\hat{R} = \bar{x}^{-1}\bar{y} = \left(\sum_{i\in A} x_i\right)^{-1}\sum_{i\in A} y_i \tag{4.2.16}$$

and assume a simple random sample. The jackknife deviate is

$$\begin{aligned}
\hat{R}^{(k)} - \hat{R} &= (\bar{x}^{(k)})^{-1}\bar{y}^{(k)} - \bar{x}^{-1}\bar{y} \\
&= \left(\sum_{j\in A} x_j - x_k\right)^{-1}(y_k - \hat{R}x_k).
\end{aligned} \tag{4.2.17}$$

The jackknife variance estimator based on (4.2.17) is

$$\hat{V}_J\{\hat{R}\} = (n-1)^{-1}n^{-1}\sum_{k\in A}(\bar{x}^{(k)})^{-2}(y_k - \hat{R}x_k)^2. \tag{4.2.18}$$

The traditional Taylor estimator of the variance of $\hat{R}$ is

$$\hat{V}_T\{\hat{R}\} = (n-1)^{-1}n^{-1}\sum_{i\in A}\bar{x}^{-2}(y_i - \hat{R}x_i)^2. \qquad (4.2.19)$$

Assume that $x_i > 0$ for all i. Because $(\bar{x}^{(k)})^{-2}$ is a convex function of $\bar{x}$, $E\{(\bar{x}^{(k)})^{-2}\} \geq E\{\bar{x}^{-2}\}$. In many of the data configurations in which the ratio is computed, $| y_k - Rx_k |$ is positively correlated with x_k and hence $\bar{x}^{(k)}$ is negatively correlated with $| y_k - Rx_k |$. In such situations the expected value of $(\bar{x}^{(k)})^{-2}(y_k - Rx_k)^2$ exceeds the expected value of $\bar{x}^{-2}(y_k - Rx_k)^2$, and the jackknife variance estimator can be considerably larger than the Taylor variance estimator.

4.2.3 Quantiles

The simple jackknife variance estimator is appropriate for differentiable functions of sample means but is not appropriate for nonsmooth functions such as sample quantiles. One approach to quantiles is to delete relatively large numbers of units at a time. See Shao (1989a), Shao and Tu (1995), and Kott (2001, 2006a). A second approach is to define functions for the replicates that give proper estimated variances.

For one implementation of the second approach, we use the Bahadur representation introduced in Section 1.3.5 and employ a local approximation to the cumulative distribution function to define replicates so that no computation outside jackknife replication is required. Let ξ_b be the quantile of interest, let $\hat{\xi}_b$ be the full-sample estimator of ξ_b, and let $i(b)$ be the largest integer j such that $y_{(j)} \leq \hat{\xi}_b$, where the $y_{(j)}$ are the order statistics of the full sample. Let $i(s)$ and $i(t)$ be the largest integers less than or equal to

$$i(s) = max\{1,\ i(b) - 2\,[b(1-b)n]^{1/2}\}$$

and

$$i(t) = min\{n,\ i(b) + 2\,[b(1-b)n]^{1/2}\},$$

respectively. See Exercise 12. For replicate k, let the estimated quantile be

$$\hat{\xi}_b^{(k)} = \hat{\xi}_b + \hat{\gamma}[b - \hat{F}^{(k)}(y_{(i(b))})], \qquad (4.2.20)$$

where

$$\hat{\gamma} = [\hat{F}(y_{(i(t))}) - \hat{F}(y_{(i(s))})]^{-1}(y_{(i(t))} - y_{(i(s))}).$$

Calculation of $\hat{\xi}_b, i(b), i(s)$, and $\hat{\gamma}$ can be part of the calculation for each replicate if necessary to avoid external computations. The jackknife variance

estimator is then

$$\hat{V}_{JK}\{\hat{\xi}_b\} = \sum_{k=1}^{L} c_k(\hat{\xi}_b^{(k)} - \hat{\xi}_b^{(\cdot)})^2, \tag{4.2.21}$$

where $\hat{\xi}_b^{(\cdot)}$ is the average of the $\hat{\xi}^{(k)}$ for the L replicates and c_k is determined by the design. The jackknife deviate is

$$\hat{\xi}^{(k)} - \hat{\xi}^{(\cdot)} = \hat{\gamma}[\hat{F}^{(\cdot)}(y_{(i(b))}) - \hat{F}^{(k)}(y_{(i(b))})]$$

and

$$\hat{V}_{JK}\{\hat{\xi}_b\} = \hat{\gamma}^2\hat{V}\{\hat{F}(y_{(i(b))})\}.$$

We used a particular smoothed estimator of the quantile in (4.2.20), but there are alternative smoothed estimators for which the jackknife remains appropriate. See Sheather (2004) and references cited there.

4.3 BALANCED HALF-SAMPLES

In Section 4.2 we created replicates for a stratified sample by deleting individual primary sampling units in each stratum and applying the appropriate multiplier to the squared deviation. A special replication technique for two-per-stratum designs was developed at the U.S. Census Bureau using results of McCarthy (1969). Let a sample be composed of two units in each of H strata. Then the stratified estimator of the mean is

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h,$$

where $W_h = N^{-1}N_h$, $\bar{y}_h = 0.5(y_{h1}+y_{h2})$. If the finite population correction can be ignored, the estimated variance is

$$\begin{aligned}\hat{V}\{\bar{y}_{st}\} &= 0.5\sum_{h=1}^{H} W_h^2 s_h^2 \\ &= 0.25\sum_{h=1}^{H} W_h^2 (y_{h1} - y_{h2})^2.\end{aligned}$$

See Section 1.2.3.

Consider a *half-sample* created by selecting at random one element from each stratum, and define the half-sample estimator of the mean by

$$\bar{y}_{st(k)} = \sum_{h=1}^{H} W_h y_{h(k)},$$

where $y_{h(k)}$ is the unit selected in stratum h. For an alternative representation of the half-sample, let $\delta_{hj}^{(k)} = 1$ if element hj is selected for the sample and $\delta_{hj}^{(k)} = -1$ otherwise. Then the half-sample estimator is

$$\bar{y}_{st(k)} = 0.5\sum_{h=1}^{H} W_h \sum_{j=1}^{2}(\delta_{hj}^{(k)} + 1)y_{hj}.$$

It follows that

$$\begin{aligned} E\{(\bar{y}_{st(k)} - \bar{y}_{st})^2 \mid \mathcal{F}\} &= E\left\{\left(0.5\sum_{h=1}^{H} W_h \sum_{j=1}^{2} \delta_{hj}^{(k)} y_{hj}\right)^2\right\} \\ &= 0.5\sum_{h=1}^{H} W_h^2 S_h^2 \end{aligned}$$

and the squared difference is unbiased for $V\{\bar{y}_{st} \mid \mathcal{F}\}$. Of course, the use of one square gives a very inefficient estimator of the variance. One can improve efficiency by using more half-samples. As with the jackknife, it is possible to define a set of half-samples such that the average of the squared differences reproduces the usual variance estimator. See Wolter (2007, Chapter 3).

4.4 TWO-PHASE SAMPLES

Two-phase sampling was introduced in Section 3.3. The variance estimator given in (3.3.28) can be very difficult to compute for complex first-phase samples. We outline replication procedures that are often easier to implement.

Assume a two-phase sample in which the second-phase sampling rates and sampling procedure are defined prior to selection of the first-phase sample. An example is a population divided into G mutually exclusive and exhaustive groups with a second-phase sampling rate specified for each group. It may be, and usually is, the case that the first-phase sample is selected to identify membership in the groups. Assume that there is a replicate variance estimator that gives a consistent estimator of the full-sample first-phase direct-expansion estimator of the total. We write the replication estimator for the covariance matrix of $\hat{\mathbf{T}}_{1x}$ as

$$\hat{V}_{n1}\{\hat{\mathbf{T}}'_{1x}\} = \sum_{k=1}^{L} c_k(\hat{\mathbf{T}}_{1x}^{(k)} - \hat{\mathbf{T}}_{1x})'(\hat{\mathbf{T}}_{1x}^{(k)} - \hat{\mathbf{T}}_{1x}), \qquad (4.4.1)$$

where L is the number of replicates, $\hat{\mathbf{T}}_{1x}^{(k)}$ is the estimated total for the kth replicate, $\hat{\mathbf{T}}_{x1}$ is the full-sample estimator of the total computed from the

first-phase sample,

$$\hat{\mathbf{T}}_{1x} = \sum_{i \in A_1} \pi_{1i}^{-1} \mathbf{x}_i = \sum_{i \in A_1} w_{1i} \mathbf{x}_i, \tag{4.4.2}$$

$w_{1i}^{-1} = \pi_{1i}$ is the probability that element i is in the first-phase sample, A_1 is the set of indices in the first-phase sample, and c_k is a factor associated with replicate k. We assume that the kth replicate can be written as

$$\hat{\mathbf{T}}_{1x}^{(k)} = \sum_{i \in A_1} w_{1i}^{(k)} \mathbf{x}_i,$$

where $w_{1i}^{(k)}$ is the weight of element i for the kth replicate.

We define the two-phase regression estimator of the mean of y by

$$\bar{y}_{2p,reg} = \bar{y}_{2\pi} + (\bar{\mathbf{x}}_{1\pi} - \bar{\mathbf{x}}_{2\pi})\hat{\boldsymbol{\beta}}_2, \tag{4.4.3}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_2 &= \mathbf{M}_{xx}^{-1} \sum_{i \in A_2} w_{2i} (\mathbf{x}_i - \bar{\mathbf{x}}_{2\pi})'(y_i - \bar{y}_{2\pi}), \\
(\bar{\mathbf{x}}_{2\pi},\ \bar{y}_{2\pi}) &= \left(\sum_{i \in A_2} w_{2i} \right)^{-1} \sum_{i \in A_2} w_{2i} (\mathbf{x}_i,\ y_i), \\
\bar{\mathbf{x}}_{1\pi} &= \left(\sum_{i \in A_1} \pi_{1i}^{-1} \right)^{-1} \sum_{i \in A_1} \pi_{1i}^{-1} \mathbf{x}_i, \\
(\mathbf{M}_{xx}, \mathbf{M}_{xy}) &= \sum_{i \in A_2} w_{2i} (\mathbf{x}_i - \bar{\mathbf{x}}_{2\pi})' \left[(\mathbf{x}_i - \bar{\mathbf{x}}_{2\pi}), (y_i - \bar{y}_{2\pi}) \right]
\end{aligned}$$

$w_{2i}^{-1} = \pi_{2i} = \pi_{1i}\pi_{2i|1i}$ is the second-phase selection probability for element i, and $\pi_{2i|1i}$ is the conditional probability of selecting element i for the phase 2 sample given that i is in the phase 1 sample. The estimator can be written as

$$\bar{y}_{2p,reg} = \sum_{i \in A_2} w_{2reg,i} y_i,$$

where

$$w_{2reg,i} = N^{-1} \left[w_{2i} + (\bar{\mathbf{x}}_{1\pi} - \bar{\mathbf{x}}_{2\pi}) \mathbf{M}_{xx}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{2\pi})' w_{2i} \right].$$

Consider a replication procedure in which the kth replicate is created on the basis of the phase 1 sampling units. For example, if a jackknife replicate for the phase 1 sample is created by deleting a phase 1 unit from a sample of

n units, the kth replicate of the phase 2 sample is composed of all phase 2 elements in the remaining $n-1$ phase 1 units. Let the kth replicate estimator of the mean of y be

$$\bar{y}_{2p,reg}^{(k)} = \bar{y}_{2\pi}^{(k)} + (\bar{\mathbf{x}}_{1\pi}^{(k)} - \bar{\mathbf{x}}_{2\pi}^{(k)})\hat{\boldsymbol{\beta}}_2^{(k)}, \tag{4.4.4}$$

where $\bar{\mathbf{x}}_{1\pi}^{(k)}$ is the estimator for the kth phase 1 replicate,

$$\begin{aligned}(\bar{\mathbf{x}}_{2\pi}^{(k)},\ \bar{y}_{2\pi}^{(k)}) &= \left(\sum_{i\in A_2} w_{2i}^{(k)}\right)^{-1} \sum_{i\in A_2} w_{2i}^{(k)}(\mathbf{x}_i,\ y_i), \\ \hat{\boldsymbol{\beta}}_2^{(k)} &= \left(\mathbf{M}_{xx}^{(k)}\right)^{-1} \sum_{i\in A_2} w_{2i}^{(k)}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{2\pi}^{(k)})'(y_i - \bar{y}_{2\pi}^{(k)}),\end{aligned}$$

$$(\mathbf{M}_{xx}^{(k)}, \mathbf{M}_{xy}^{(k)}) = \sum_{i\in A_2} w_{2i}^{(k)}(\mathbf{x}_i - \bar{\mathbf{x}}_{2\pi}^{(k)})'[(\mathbf{x}_i - \bar{\mathbf{x}}_{2\pi}^{(k)}), (y_i - \bar{y}_{2\pi}^{(k)})]$$

and $w_{2i}^{(k)} = \pi_{2i|1i}^{-1} w_{1i}^{(k)}$ is the phase 2 weight for element i in replicate k. If the replicate estimator is written in the form

$$\bar{y}_{2p,reg}^{(k)} = \sum_{i\in A_2} w_{2reg,i}^{(k)} y_i, \tag{4.4.5}$$

the regression weights are

$$w_{2reg,i}^{(k)} = w_{2i}^{(k)} + (\bar{\mathbf{x}}_{1\pi}^{(k)} - \bar{\mathbf{x}}_{2\pi}^{(k)})(\mathbf{M}_{xx}^{(k)})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{2\pi}^{(k)})' w_{2i}^{(k)}.$$

It is shown in Theorem 4.4.1 that the variance estimator

$$\hat{V}\{\bar{y}_{2p,reg} \mid \mathcal{F}\} = \sum_{k=1}^{L} c_k(\bar{y}_{2p,reg}^{(k)} - \bar{y}_{2p,reg})^2, \tag{4.4.6}$$

has a negative bias of order N^{-1}. The finite population of the theorem is assumed to be a sample from a superpopulation and the conclusions are with respect to all such finite populations. Replication variance estimators such as balanced half-samples or the ordinary jackknife satisfy the assumptions for the variance estimator. See Kim, Navarro, and Fuller (2006).

Theorem 4.4.1. Assume a sequence of finite populations in which each population is composed of G groups, $G \geq 1$, with proportion W_g in the gth group. Let $\mathcal{F}_N = \{(\mathbf{x}_{1N},\ y_{1N}),\ (\mathbf{x}_{2N},\ y_{2N}),\ \ldots,\ (\mathbf{x}_{NN},\ y_{NN})\}$, where the first $G-1$ elements of $\mathbf{x}_{iN}$ are indicator variables for membership in $G-1$ of

the groups. Assume that the finite population in the gth group is a sample from an infinite population with $4 + \delta$, $\delta > 0$, moments. Let two-phase samples be selected, where the phase 1 selection probabilities are $\pi_{1i,N}$. Let a set of fixed probabilities $\pi_{2i|1i} = \kappa_{2i}$, constant within a group, be used to select a phase 2 stratified random sample or to select a phase 2 Poisson sample. Let

$$(\hat{\mathbf{T}}_{1x},\ \hat{T}_{1y}) \quad = \quad \sum_{i \in A_1} w_{1i,N}(\mathbf{x}_{iN},\ y_{iN}), \tag{4.4.7}$$

where $w_{1i,N} = \pi_{1i,N}^{-1}$. Assume that:

(i) The phase 1 selection probabilities satisfy

$$K_L < N n_{1N}^{-1} \pi_{1i,N} < K_U \tag{4.4.8}$$

for all N, where K_L and K_U are fixed positive constants.

(ii) The variance of $\hat{T}_{1y}$ for a complete phase 1 sample satisfies

$$V\{\hat{T}_{1y} \mid \mathcal{F}_N\} \quad \leq K_M V\{\hat{T}_{y,SRS} \mid \mathcal{F}_N\}, \tag{4.4.9}$$

for a fixed K_M, for any y with fourth moments, where $V\{\hat{T}_{y,SRS} \mid \mathcal{F}_N\}$ is the variance of the Horvitz–Thompson estimator of the total for a simple random sample of size n_{1N}.

(iii) The variance of a complete phase 1 linear estimator of the mean is a symmetric quadratic function, and

$$n_N V\left\{ N^{-1} \sum_{i \in A_1} \pi_{1i,N}^{-1} y_{iN} \mid \mathcal{F}_N \right\} \quad = \quad \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij,N} y_{iN} y_{jN} \tag{4.4.10}$$

for coefficients $\omega_{ij,N}$, where

$$\sum_{i=1}^{N} |\omega_{ij,N}| \quad = \quad O(N^{-1}). \tag{4.4.11}$$

(iv) The phase 1 replicate variance estimator of $\hat{T}_{1y}$, denoted by $\hat{V}_1\{\hat{T}_{1y}\}$, satisfies

$$E\{[(V\{\hat{T}_{1y} \mid \mathcal{F}_N\})^{-1} \hat{V}_1\{\hat{T}_{1y}\} - 1]^2 \mid \mathcal{F}_N\} \quad = \quad o(1) \tag{4.4.12}$$

for any y with bounded fourth moments.

(v) The replicates for the phase 1 estimator, $\hat{T}_{1y}$, satisfy

$$E\{[c_{kN}(\hat{T}_{1y}^{(k)} - \hat{T}_{1y})^2]^2 \mid \mathcal{F}_N\} \quad < \quad K_\gamma L_N^{-2} [V\{\hat{T}_{1y} \mid \mathcal{F}_N\}]^2 \tag{4.4.13}$$

uniformly in N for any variable with fourth moments, where K_γ is a fixed constant and L_N is the number of replicates.

Then the variance estimator (4.4.6) with replicates (4.4.4) satisfies

$$\begin{aligned}\hat{V}\{\bar{y}_{2p,reg}\} &= V\{\bar{y}_{2p,reg} \mid \mathcal{F}_N\} \\ &\quad - N^{-2}\sum_{i=1}^{N}\kappa_{2i}^{-1}(1-\kappa_{2i})e_{iN}^2 + o_p(n_{1N}^{-1}), \qquad (4.4.14)\end{aligned}$$

where $e_{iN} = y_{iN} - \bar{y}_N - (\mathbf{x}_{iN} - \bar{\mathbf{x}}_N)\boldsymbol{\beta}_N$, n_{1N} is the size of the phase 1 sample and

$$\boldsymbol{\beta}_N = \left(\sum_{i=1}^{N}(\mathbf{x}_{iN} - \bar{\mathbf{x}}_N)'(\mathbf{x}_{iN} - \bar{\mathbf{x}}_N)\right)^{-1}\sum_{i=1}^{N}(\mathbf{x}_{iN} - \bar{\mathbf{x}}_N)'(y_{iN} - \bar{y}_N).$$

Proof. To simplify the notation, we sometimes omit the N subscript. Let

$$\begin{aligned}a_i &= 1 \quad \text{if element } i \text{ is selected for the phase 2 sample} \\ &= 0 \quad \text{otherwise.} \qquad (4.4.15)\end{aligned}$$

Under the assumption that the phase 2 sampling rates, $\pi_{2i|1i} = \kappa_{2i}$, are fixed, we can conceptualize the sample selection process as composed of two steps. First, an a is generated for every element in the population, and then a phase 1 sample is selected from the population of $(a_i,\ a_i y_i,\ \mathbf{x}_i)$ vectors. See Fay (1991) and Rao and Shao (1992). Let $z_i = a_i y_i$ and $\mathbf{u}_i = a_i \mathbf{x}_i$. Then the regression estimator can be expressed as a function of phase 1 estimators,

$$\begin{aligned}\bar{y}_{2p,reg} &= g(\bar{z}_{1\pi}, \bar{\mathbf{u}}_{1\pi}, \bar{w}_{1\pi}, \mathbf{M}_{1,uu}, \mathbf{M}_{1,uz}) \\ &= \bar{z}_{1,\pi} + (\bar{\mathbf{u}}_{1\pi} - \bar{\mathbf{x}}_{1\pi})\hat{\boldsymbol{\beta}}, \qquad (4.4.16)\end{aligned}$$

where

$$(\bar{z}_{1\pi}, \bar{\mathbf{u}}_{1\pi}, \bar{w}_{1\pi}) = \left(\sum_{i\in A_1} a_i\pi_{2i}^{-1}\right)^{-1}\sum_{i\in A_1}\pi_{2i}^{-1}(z_i, \mathbf{u}_i, w_i),$$

$\mathbf{M}_{1,uu} = \mathbf{M}_{xx}$ and $\mathbf{M}_{1,uz} = \mathbf{M}_{xy}$. By assumptions (4.4.12) and (4.4.13), and by the extension of Theorem 4.2.1 to complex designs, the phase 1 replicate variance estimator is consistent for $V\{\bar{y}_{2p,reg} \mid (\mathbf{a}_N, \mathcal{F}_N)\}$, where $\mathbf{a}_N$ is the N-dimensional vector $(a_1, a_2, \ldots, a_N)$. We write the variance of $\bar{y}_{2p,reg}$ in terms of conditional expectations,

$$\begin{aligned}V\{\bar{y}_{2p,reg} \mid \mathcal{F}_N\} &= E\{V[\bar{y}_{2p,reg} \mid (\mathbf{a}_N,\ \mathcal{F}_N)] \mid \mathcal{F}_N\} \\ &\quad + V\{E[\bar{y}_{2p,reg} \mid (\mathbf{a}_N,\ \mathcal{F}_N)] \mid \mathcal{F}_N\}. \qquad (4.4.17)\end{aligned}$$

To show that the replicate variance estimator is consistent for the first term of (4.4.17), we must show that the estimator of the conditional expectation is a consistent estimator of the unconditional expectation. Because

$$\bar{y}_{2p,reg} = \bar{y}_{2\pi} + (\bar{\mathbf{x}}_{1\pi} - \bar{\mathbf{x}}_{2\pi})\boldsymbol{\beta}_N + O_p(n_N^{-1}), \tag{4.4.18}$$

the variance of the approximate distribution of $\bar{y}_{2p,reg}$ is the variance of $\bar{e}_{2\pi} + \bar{\mathbf{x}}_{1\pi}\boldsymbol{\beta}_N$. Because $\bar{\mathbf{x}}_{1\pi}$ does not depend on $\mathbf{a}_N$, we consider only $\bar{e}_{2\pi}$ and further simplify by considering $\bar{e}_{2,HT}$, where

$$\bar{e}_{2,HT} = N^{-1} \sum_{i \in A_1} \pi_{2i}^{-1} a_i e_i.$$

By (4.4.10),

$$n_N V\{\bar{e}_{2,HT} \mid (\mathbf{a}_N, \mathcal{F}_N)\} = \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij}(d_i + 1)(d_j + 1)e_i e_j, \tag{4.4.19}$$

where $d_i = \kappa_{2i}^{-1} a_i - 1$. We assume that $\mathbf{a}_N$ is a Poisson sample so that the a_j are independent Bernoulli random variables, expand $(d_i + 1)(d_j + 1)$, and consider the resulting four terms of (4.4.19). Now

$$V\left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} e_i e_j \mid \mathcal{F}_N \right\} = 0$$

by the conditioning. Also,

$$\begin{aligned} V\left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} d_i e_i e_j \mid \mathcal{F}_N \right\} &= \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \omega_{ij} e_j \right)^2 \gamma_{di}^2 e_i^2 \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} \gamma_{di}^2 e_i^2 e_j \sum_{k=1}^{N} \omega_{ik} e_k \\ &= O_p(N^{-1/2}), \end{aligned} \tag{4.4.20}$$

where $V\{d_i\} = \gamma_{di}^2 = \kappa_{2i}^{-1}(1 - \kappa_{2i})$, because the double sum in $e_i^2 e_j$ is an $O_p(1)$ covariance by (4.4.9) and (4.4.10), and $\Sigma_{k=1}^{N} \omega_{ik} e_k = O_p(N^{-1/2})$ by the assumptions on the sequence of finite populations. To evaluate the variance of the remaining term of (4.4.19), we use

$$\begin{aligned} Cov\,(d_i d_j, d_k d_m \mid \mathcal{F}_N) &= \gamma_{di}^2 \gamma_{dj}^2 \qquad \text{if } ij = km \text{ or } ij = mk \\ &= 0 \qquad \text{otherwise.} \end{aligned} \tag{4.4.21}$$

Then

$$V\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\omega_{ij}d_id_je_ie_j \mid \mathcal{F}_N\right\} = \sum_{i=1}^{N}\sum_{j=1}^{N}2\omega_{ij}^2\gamma_{di}^2\gamma_{dj}^2e_i^2e_j^2$$

$$\leq max\,|\omega_{ij}|\sum_{i=1}^{N}\sum_{j=1}^{N}2\,|\omega_{ij}|\,\gamma_{di}^2\gamma_{dj}^2e_i^2e_j^2$$

$$= O_p\left(N^{-1}\right), \tag{4.4.22}$$

by (4.4.21) and assumption (4.4.11). Therefore, $\hat{V}\{\bar{e}_{2,HT} \mid (\mathbf{a}_N,\ \mathcal{F})\}$ is consistent for $E\{V\,[\bar{e}_{2,HT} \mid (\mathbf{a}_N,\ \mathcal{F})] \mid \mathcal{F}\}$ and the replicate variance estimator is consistent for the first term of (4.4.17).

To evaluate the second term on the right side of equality (4.4.17), we have

$$E\left\{N^{-1}\sum_{i\in A_1}\pi_{1i}^{-1}\kappa_{2i}^{-1}a_ie_i \mid (\mathbf{a}_N,\ \mathcal{F})\right\} = N^{-1}\sum_{i=1}^{N}\kappa_{2i}^{-1}a_ie_i$$

and for Poisson sampling,

$$V\{N^{-1}\sum_{i=1}^{N}\kappa_{2i}^{-1}a_ie_i \mid \mathcal{F}\} = N^{-2}\sum_{i=1}^{N}\kappa_{2i}^{-1}(1-\kappa_{2i})e_i^2. \tag{4.4.23}$$

Combining the consistency of the replicate variance estimator for the first term of (4.4.17) with (4.4.23), we have conclusion (4.4.14) for phase 2 Poisson sampling. By the arguments used in the proof of Theorem 1.3.3, the conclusion also holds for stratified samples. ■

If the phase 1 sampling rate is small, the second term of (4.4.14) is small relative to the first term, and estimator (4.4.6) can be used for the two-phase regression estimator. If the second term of (4.4.14) is judged to be important, the term can be estimated directly or with replicates.

In Theorem 4.4.1 it is assumed that the error in $\hat{\beta}_2$ is small. The delete-one jackknife variance estimator will have a positive bias if the phase 2 samples within strata are small. To investigate the nature of the small-sample bias, consider the simple stratified estimator of (3.3.13), which we write as

$$\bar{e}_{2\pi} = \sum_{g=1}^{G}W_g\bar{e}_{2g},$$

where $\bar{e}_{2g}$ is the mean for phase 2 stratum g, and W_g is the fraction of the population in phase 2 stratum g. Then the replicate estimator of the mean is

$$\bar{e}_{2\pi}^{(k)} = \sum_{g=1}^{G} W_g \bar{e}_{2g}^{(k)},$$

where $\bar{e}_{2g} = \bar{e}_{2g}^{(k)}$ unless element k is in group g. By (4.2.3),

$$E\left\{\sum_{k=1}^{L}(\bar{e}_{2\pi}^{(k)} - \bar{e}_{2\pi})^2\right\} = \sum_{g=1}^{G} W_g^2 (n_g - 1)^{-1} S_g^2, \qquad (4.4.24)$$

where S_g^2 is the variance for phase 2 stratum g and n_g is the number of elements in phase 2 stratum g. Thus, the estimator (4.4.6) is biased to the degree that the c_k differ from the $n_{2,g(k)}^{-1}(n_{2,g(k)} - 1)$, where $g(k)$ is the phase 2 stratum containing element k.

For some designs and replication procedures, it is possible to modify the weights to reduce the bias. To improve the performance of the variance estimator, we recall expression (4.4.16), which identifies two components that contribute to the variance. It is possible to construct a replicate for each component. The squared deviate associated with the replicate

$$\bar{y}_1^{(k)} = \bar{y}_{2\pi} + (\bar{\mathbf{x}}_{1\pi}^{(k)} - \bar{\mathbf{x}}_{2\pi})\hat{\boldsymbol{\beta}} \qquad (4.4.25)$$

will estimate the variance of $(\bar{\mathbf{x}}_{1\pi} - \bar{\mathbf{x}}_N)\boldsymbol{\beta}$, and the squared deviate associated with the replicate

$$\bar{y}_2^{(k)} = \bar{y}_{2\pi}^{(k)} + (\bar{\mathbf{x}}_{1\pi} - \bar{\mathbf{x}}_{2\pi}^{(k)})\hat{\boldsymbol{\beta}}^{(k)} \qquad (4.4.26)$$

will estimate the variance of $\bar{y}_{2\pi} - (\bar{\mathbf{x}}_{2\pi} - \bar{\mathbf{x}}_N)\hat{\boldsymbol{\beta}}_{2\pi}$. In both cases the replicates are formed on the basis of the phase 1 units. The procedure associated with (4.2.6) can be used to reduce the bias in the estimator of the phase 2 variance. Let the replication method be the delete-one jackknife and let element k in group $g(k)$, where $g(k)$ is the phase 2 stratum containing element k, be the "deleted" element. Let

$$\begin{aligned}\bar{y}_2^{(k)} = {} & \psi_g w_{1k} n_{1,g(k)} y_k \\ & + (n_{2g} - 1)^{-1}(1 - \psi_g)\left(\sum_{j \in A_{2g}} w_{1j} n_{1,g(k)} y_j - w_{1k} n_{1,g(k)} y_k\right) \\ & + \sum_{g \neq g(k)} \sum_{j \in A_{2g}} w_{1j} n_{1,g} n_{2,g}^{-1} y_j,\end{aligned}$$

where ψ_g is the solution of the quadratic

$$(1 - n_{2g}\psi_g)^2(n_{2g} - 1)^{-1} = c_k^{-1}n_{2g}^{-1}. \tag{4.4.27}$$

If unit k is a phase 2 unit, the replicate deviate is $\bar{y}_1^{(k)} + \bar{y}_2^{(k)} - 2\bar{y}_{2p,reg}$. If unit k is not a phase 2 unit, the replicate deviate is $\bar{y}_1^{(k)} - \bar{y}_{2p,reg}$. One can use the adjusted weights in a stratum for the stratified estimator (3.3.13) because adjusting the phase 2 weights will not alter the replicate estimate of the fraction in the stratum. If the $\mathbf{x}$ vector contains variables other than the group indicators, the replicate mean of $\mathbf{x}$ must be calculated with the unmodified weights.

Example 4.4.1. We use a small artificial data set to illustrate the construction of jackknife replicates for two-phase samples and to demonstrate the nature of possible biases in the replication variance estimator. The eight observations in Table 4.3 are assumed to be a simple random sample. We assume that N is large and ignore the finite population correction. On the basis of the classification defined by x, a phase 2 stratified sample is selected with a sampling rate of one-half in each stratum. The original weights for the full-sample mean are 1/8, and the weights for the phase 1 jackknife replicates are zero for the deleted unit and 1/7 for the remaining units. The eight estimates of $\bar{x}_N$ for the eight phase 1 replicates are given in the second column of Table 4.4. Observe that

$$(7/8)\sum_{k\in A}(\bar{x}^{(k)} - \bar{x})^2 = (7/8)(0.040816) = 0.03571,$$

which equals the usual variance estimator of $(n-1)^{-1}\bar{x}(1-\bar{x})$.

Table 4.3 Two-Phase Sample

			Phase 1 Replicate Weight							
Ident.	x	y	1	2	3	4	5	6	7	8
1	1	1	0	1/7	1/7	1/7	1/7	1/7	1/7	1/7
2	1	4	1/7	0	1/7	1/7	1/7	1/7	1/7	1/7
3	1	–	1/7	1/7	0	1/7	1/7	1/7	1/7	1/7
4	1	–	1/7	1/7	1/7	0	1/7	1/7	1/7	1/7
5	0	5	1/7	1/7	1/7	1/7	0	1/7	1/7	1/7
6	0	7	1/7	1/7	1/7	1/7	1/7	0	1/7	1/7
7	0	–	1/7	1/7	1/7	1/7	1/7	1/7	0	1/7
8	0	–	1/7	1/7	1/7	1/7	1/7	1/7	1/7	0

Table 4.4 Statistics for Jackknife Replicates

Replicate	$\bar{x}^{(k)}$	$\bar{y}_1^{(k)}$	$\bar{y}_2^{(k)}$	$\hat{\beta}_1^{(k)}$	$\bar{y}_{2p,reg}^{(k)}$	Adjusted $\bar{y}_{2p,reg}^{(k)}$
1	3/7	4.0	6	–2.0	5.1429	4.9861
2	3/7	1.0	6	–5.0	3.8571	4.0141
3	3/7	2.5	6	–3.5	4.5000	4.5000
4	3/7	2.5	6	–3.5	4.5000	4.5000
5	4/7	2.5	7	–4.5	4.4286	4.3242
6	4/7	2.5	5	–2.5	3.5714	3.6762
7	4/7	2.5	6	–3.5	4.0000	4.0000
8	4/7	2.5	6	–3.5	4.0000	4.0000

The replicate weights in Table 4.5 are the replication weights defined by (4.4.5) and give the $\bar{y}_{2p,reg}^{(k)}$ of the penultimate column of Table 4.4. The estimated phase 2 variance (4.4.6) is

$$\begin{aligned} \hat{V}_{JK}\{\bar{y}_{2p,reg}\} &= (7/8)\sum_{k=1}^{8}(\bar{y}_{2p,reg}^{(k)} - \bar{y}_{2p,reg})^2 \\ &= 1.4823. \end{aligned} \tag{4.4.28}$$

Table 4.5 Phase 2 Replicate Weights

Ident.	Replicate 1	2	3	4	5	6	7	8
1	0	0.4286	0.2143	0.2143	0.2857	0.2857	0.2857	0.2857
2	0.4286	0	0.2143	0.2143	0.2857	0.2857	0.2857	0.2857
5	0.2857	0.2857	0.2857	0.2857	0	0.4286	0.2143	0.2143
6	0.2857	0.2857	0.2857	0.2857	0.4286	0	0.2143	0.2143

For this simple example an estimator of the population S_y^2 is

$$\hat{S}_y^2 = \hat{\beta}^2\bar{x}(1-\bar{x}) + 4.5\bar{x} + 2.0(1-\bar{x}) = 6.3125, \tag{4.4.29}$$

where 4.5 and 2.0 are the sample variances for the two phase 2 strata. Thus, an estimator of the variance of the two-phase mean is

$$\begin{aligned} \hat{V}\{\bar{y}_{2p,st}\} &= n_1^{-1}\hat{S}_y^2 + \sum_{g=1}^{2} \hat{W}_g^2(1-f_g)n_{2g}^{-1}s_g^2 \\ &= 0.7891 + 0.4063 = 1.1954, \end{aligned} \tag{4.4.30}$$

where $\hat{W}_1 = \bar{x} = 0.5$, $\hat{W}_2 = (1-\bar{x})$, and $f_g = n_{1g}^{-1}n_{2g}$. The considerable difference between the two estimates is due to the small sample sizes in the phase 2 strata.

Table 4.6 Adjusted Phase 2 Replicate Weights

	Replicate							
Ident.	1	2	3	4	5	6	7	8
1	0.0523	0.3763	0.2143	0.2143	0.2857	0.2857	0.2857	0.2857
2	0.3763	0.0523	0.2143	0.2143	0.2857	0.2857	0.2857	0.2857
5	0.2857	0.2857	0.2857	0.2857	0.0523	0.3763	0.2143	0.2143
6	0.2857	0.2857	0.2857	0.2857	0.3763	0.0523	0.2143	0.2143

For this sample design, we can adjust the weights to remove the bias. The ψ_g of (4.4.27) is 0.1220 for both strata because the stratum sample sizes are the same. The adjusted replicate weights are given in Table 4.6. Only the weights for the strata and the replicate in which a phase 2 element is "deleted" are adjusted. The adjusted replicate estimates are given in the last column of Table 4.4. The estimated variance computed with the adjusted replicates is 1.1873, which agrees well with the estimate (4.4.30) based on nearly unbiased components. ■■

4.5 THE BOOTSTRAP

The bootstrap is a replication procedure that creates replicates by selecting samples with replacement from the original sample. Consider a simple random sample of size n, from which a number, say L, of replacement samples are selected. By the results of Chapter 1, we know that the mean of a replacement sample is the original sample mean, and the variance of the replacement sample mean is the finite population variance divided by the sample size. Thus, the variance of replacement samples of size m selected from the original

sample is

$$V\{\bar{y}_{b,m}\} = m^{-1}n^{-1}\sum_{i\in A}(y_i-\bar{y})^2 = m^{-1}n^{-1}(n-1)s^2, \quad (4.5.1)$$

where $\bar{y}$ is the original sample mean and

$$s^2 = (n-1)^{-1}\sum_{i\in A}(y_i-\bar{y})^2. \quad (4.5.2)$$

If we set $m = n-1$, the variance of the mean of the replacement samples is equal to the usual estimator of the variance of the original sample mean.

It follows that an estimator of the variance of the original sample mean is

$$\hat{V}_b\{\bar{y}-\bar{y}_N\} = N^{-1}(N-n)L^{-1}\sum_{j=1}^{L}(\bar{y}_{bj}-\bar{y})^2, \quad (4.5.3)$$

where $\bar{y}_{bj}$ is the mean of the jth replacement sample of size $n-1$.

The bootstrap sample mean can be written as the weighted mean of the original sample,

$$\bar{y}_{bj} = \sum_{i\in A} w_{ji}y_i, \quad (4.5.4)$$

where $w_{ji} = n^{-1}r_{ji}$ and r_{ji} is the number of times unit i is selected for the jth bootstrap sample.

Equations (4.5.1) and (4.5.4) can be used to define samples for more complicated designs. Consider a stratified sample with sample sizes n_h and sampling rates f_h. We wish to create replicates so that the expected value of the bootstrap variance is the variance of the mean. Therefore, we desire m_h such that

$$(1-f_h)n_h^{-1} = m_h^{-1}n_h^{-1}(n_h-1). \quad (4.5.5)$$

If f_h is not zero, the m_h will, in general, not be an integer. There are two ways to modify the bootstrap samples to obtain the correct expectation. In one procedure, proposed by Rao and Wu (1988), the weights for bootstrap samples with size m_h^* are replaced by

$$w_{br,i} = w_i + [(1-f_h)(n_h-1)^{-1}m_h^*]^{0.5}(w_{bi}-w_i), \quad (4.5.6)$$

where w_{bi} are the bootstrap weights and w_i are the original weights for the full sample. The procedure is called the *rescaling bootstrap*. Typically, m_h^* is chosen equal to n_h-1, but Rao and Wu (1988) suggest alternatives. In a second procedure, a fraction of samples with n_h elements and a fraction with

$n_h - 1$ elements is chosen so that the resulting set of samples gives the correct expected value. See McCarthy and Snowden (1985) and Shao and Tu (1995, p. 247). The procedure, called the *with-replacement bootstrap*, is outlined in Example 4.5.1.

Example 4.5.1. We construct bootstrap samples for the stratified sample of Table 4.1. The stratum sample sizes are $(n_1, n_2) = (3, 2)$ and the sampling rates are $(f_1, f_2) = (0.0833, 0.0500)$. If we select samples of size 1 from the two elements in stratum 2, the sample variance of the means has expectation $0.5s_2^2$. If we select samples of size 2, the expectation is $0.25s_2^2$. We desire the expectation to be $(0.95)(0.5s_2^2)$. Thus, we create a mixture of size 1 and 2 samples, where δ_2, the fraction of size 1 samples, is given by the solution to

$$\delta_2(0.5s_2^2) + (1 - \delta_2)(0.25s_2^2) = 0.475s_2^2.$$

Therefore, 0.90 of the bootstrap samples should have one element selected in stratum 2, and 0.10 of the samples should have two elements.

For stratum 1, δ_1, the fraction of samples with $m_1 = 2$, is the solution to

$$\delta_1(0.3333s_1^2) + (1 - \delta_1)(0.2222s_1^2) = (11/12)(0.3333)s_1^2$$

and $\delta_1 = 0.75$.

Table 4.7 Bootstrap Weights for a Stratified Sample

Stratum	Obs.	Original Weight	Replicate									
			1	2	3	4	5	6	7	8	9	10
1	1	12	36	0	36	24	18	12	0	18	18	0
	2	12	0	18	0	12	0	12	18	18	0	36
	3	12	0	18	0	0	18	12	18	0	18	0
2	1	20	40	0	40	0	40	0	40	0	40	20
	2	20	0	40	0	40	0	40	0	40	0	20

Table 4.7 contains 10 possible bootstrap samples for our stratified sample. Ninety percent of the samples for stratum 2 are of size 1 and the first nine samples of Table 4.7 are of size 1. The tenth sample is a sample of size 2 and it happens to be of two different elements, but it could look exactly like one of the first nine. For stratum 1, samples 4 and 6 are of size 3. The remaining samples are of size 2. In sample 4, element 1 was selected twice and element 2 was selected once. In sample 6, each element was selected once. ■■

The original objective of the bootstrap was to approximate the distribution of statistics using a large number of bootstrap samples. See Efron (1982),

Efron and Tibshirani (1986), and Shao and Tu (1995). For example, confidence sets might be formed on the basis of the bootstrap distribution of a statistic constructed by analogy to Student's t. In survey sampling the bootstrap is most often used to obtain an estimator of variance and the confidence interval is then constructed on the basis of the normal approximation to the distribution.

The bootstrap estimator of variance is an estimator of the estimated variance. Thus, for simple random samples from a normal distribution,

$$\begin{aligned} V\{s_b^2 - s^2\} &= E\left\{ V\left(L^{-1} \sum_{j=1}^{L} (\bar{y}_{bj} - \bar{y})^2 - s^2 \right) \mid s^2 \right\} \\ &\doteq 2L^{-1}\sigma^4, \end{aligned} \tag{4.5.7}$$

where σ^2 is the variance of the normal distribution, and

$$s_b^2 = nL^{-1} \sum_{j=1}^{L} (\bar{y}_{bj} - \bar{y})^2.$$

The variance of s_b^2 as an estimator of σ^2 is

$$\begin{aligned} V\{s_b^2 - \sigma^2\} &= V\{s_b^2 - s^2\} + V\{s^2 - \sigma^2\} \\ &\doteq 2\left[L^{-1} + (n-1)^{-1}\right]\sigma^4. \end{aligned}$$

It follows that a variance estimator for a sample of size $n = 30$ from a $N(0,1)$ distribution based on 100 bootstrap samples has an approximate variance of 0.089, which is approximately equal to the variance of the sample variance for a sample of size 22. Therefore, a large number of bootstrap samples is required to approximate the efficiency of Taylor and jackknife variance estimators.

If the sample has a very large number of strata, on the order of 1000, and a small number of units per stratum, the bootstrap variance estimator becomes a viable estimation procedure.

The bootstrap variance estimator is consistent for the variance of quantiles without modification, but confidence intervals based on the Woodruff procedures of Section 1.3.5 generally perform better. See Shao and Rao (1994) and Shao and Tu (1995, p. 268).

4.6 REFERENCES

Section 4.2. Fay (1985), Kott (2001, 2006a), Krewski and Rao (1981), Quenouille (1949, 1956), Rao and Wu (1985), Shao and Tu (1995), Tukey (1958), Wolter (2007).

Section 4.3. McCarthy (1969), Wolter (2007).

Section 4.4. Fay (1996), Fuller (1998), Kim, Navarro, and Fuller (2006), Kott (1990b), Kott and Stukel (1997), Krewski and Rao (1981), Legg and Fuller (2009), Rao and Shao (1992).

Section 4.5. Efron (1982), Efron and Tibshirani (1986), McCarthy and Snowden (1985), Rao and Wu (1988), Shao and Tu (1995).

4.7 EXERCISES

1. (Section 4.2) Assume that a simple random sample of size 30 from an infinite population with variance σ^2 is available. For each element of the sample of 30, a replicate is created as

$$\bar{x}^{(i)} = \psi x_i + (29)^{-1}(1-\psi)\sum_{j\neq i}^{30} x_j.$$

Find the ψ such that

$$E\left\{\sum_{i=1}^{30}(\bar{x}^{(i)}-\bar{x})^2\right\} = 0.01\sigma^2.$$

2. (Section 4.2) Assume that a simple random sample of size $n = mb$ from an infinite population has been split at random into m groups of size b. Define a replicate estimator of the mean by

$$\bar{x}_b^{(k)} = \psi \sum_{i\in B_{bk}} x_i + (n-b)^{-1}(1-b\psi)\sum_{j\notin B_{bk}} x_j,$$

where B_{bk} is the set of b elements deleted for the kth replicate, and the replicate estimator of the variance is

$$\hat{V}\{\bar{x}\} = \sum_{k=1}^{m}(\bar{x}_b^{(k)}-\bar{x})^2.$$

Find a ψ such that $\hat{V}\{\bar{x}\}$ is unbiased for $V\{\bar{x}\}$.

3. (Section 4.2) Let a simple random sample of 10 (x, y) vectors be (1, 0.1), (1, 1.9), (2, 2.5),(3, 1.9), (4, 2.4), (4, 6.0), (5, 3.3), (6, 9.0), (7, 4.1), (7, 10.0).

   (a) Calculate the estimator of the ratio $R_N = \bar{x}_N^{-1}\bar{y}_N$ and calculate the Taylor estimator of the variance of the ratio estimator. Assume that the finite population correction can be ignored.

   (b) Calculate the jackknife variance estimator for the ratio of part (a).

   (c) Assume that the population mean of x is $\bar{x}_N = 4.5$. Calculate the regression estimator of the mean of y. Calculate the Taylor estimator of the variance of $\bar{y}_{reg}$.

   (d) Calculate the jackknife variance estimator of the $\bar{y}_{reg}$ of part (c).

4. (Section 4.2) Let the regression estimator for a simple random sample be

$$\bar{y}_{reg} = \bar{y} + (\bar{x}_N - \bar{x})\hat{\beta},$$

   where $\bar{x}$ is the simple sample mean and

$$\hat{\beta} = \left(\sum_{i\in A}(x_i - \bar{x})^2\right)^{-1} \sum_{i\in A}(x_i - \bar{x})(y_i - \bar{y}).$$

   Define a jackknife deviate by

$$\bar{y}_{reg}^{(i)} = \bar{y}^{(i)} + (\bar{x}_N - \bar{x}^{(i)})\hat{\beta}^{(i)}.$$

   Show that the jackknife deviate can be written

$$\begin{aligned}\bar{y}_{reg}^{(i)} - \bar{y}_{reg} &= -n^{-1}\hat{e}_i - (\bar{x}^{(i)} - \bar{x}_N)(D^{(i)})^{-1}(x_i - \bar{x})\hat{e}_i \\ &\quad + O_p(n^{-2.5}) \\ &= -\left[(\bar{x} - \bar{x}_N) - n^{-1}(x_i - \bar{x})\right](D^{(i)})^{-1}(x_i - \bar{x})\hat{e}_i \\ &\quad - n^{-1}\hat{e}_i + O_p(n^{-2.5}),\end{aligned}$$

   where $\hat{e}_i = y_i - \bar{y} - (x_i - \bar{x})\hat{\beta}$ and

$$D^{(i)} = \sum_{j\in A}(x_j - \bar{x}^{(i)})^2 - (x_i - \bar{x}^{(i)})^2.$$

5. (Section 4.2) In the proof of Theorem 4.2.1, it is stated: "Given $1 > \delta > 0$, there is an n_0 such that for $n > n_0$, the probability is greater

than $1-\delta$ that $\bar{y}$ and $\bar{y}^{(k)}$, $k = 1, 2, \ldots, n$, are all in a compact set D containing μ_y as an interior point." Prove this assertion.

6. (Section 4.2) Write the jackknife variance estimator for a linear estimator as

$$\hat{V}_J\left\{\sum_{i\in A} w_i y_i\right\} = \sum_{k=1}^{L} c_k \left(\sum_{i\in A}(w_i^{(k)} - w_i)y_i\right)^2.$$

Let $(y_1, y_2, \ldots, y_n)$ be independent (μ, σ_i^2) random variables and consider the jackknife variance estimator with $c_k = n^{-1}(n-1)$ for all $k, L = n$, and

$$\begin{aligned} w_j^{(k)} &= n(n-1)^{-1}w_j \qquad \text{if } j \neq k \\ &= 0 \qquad\qquad\qquad\quad \text{if } j = k. \end{aligned}$$

What is the expected value of this jackknife variance estimator?

7. (Section 4.5) Compute the rescaling bootstrap weights for samples 1, 2, 3, 5, 7, 8, 9 of Table 4.7 under the assumption that (m_1, m_2) is always equal to $(2, 1)$.

8. (Section 4.2) Prove the extension of Theorem 4.2.1. Let $\{\mathcal{F}_N\} = \{y_1, y_2, \ldots, y_N\}$ be a sequence of finite populations, where the y_i are $iid(\mu_y, \sigma_y^2)$ random variables with finite $4+\eta$, $\eta > 0$ moments. Let a sequence of probability samples be selected from the sequence of populations such that

$$\begin{aligned} E\{\bar{y}_\pi \mid \mathcal{F}_N\} &= \bar{y}_N + O_p(n_N^{-1}) \quad \text{a.s.}, \\ V\{\bar{y}_\pi \mid \mathcal{F}_N\} &= O_p(n_N^{-1}) \quad \text{a.s.}, \end{aligned}$$

and $\lim_{N\to\infty} n_N = \infty$. Let $\bar{y}_\pi^{(k)}$, $k = 1, 2, \ldots, L_N$, be replicate estimators such that

$$E\{(\bar{y}_\pi^{(k)} - \bar{y}_\pi)^2 \mid \mathcal{F}_N\} = O_p(n_N^{-2}) \quad \text{a.s.}$$

for $k = 1, 2, \ldots, L_N$, and

$$\sum_{k=1}^{L_N}(\bar{y}_\pi^{(k)} - \bar{y}_\pi)^2 = V\{\bar{y}_\pi \mid \mathcal{F}_N\} + O_p(n^{-1.5}) \quad \text{a.s.}$$

Let $g(\bar{y}_\pi)$ be a continuous function of the mean with continuous first and second derivatives at μ_y. Then

$$\sum_{k=1}^{L_N} [g(\bar{y}_\pi^{(k)}) - g(\bar{y})]^2 \mid \mathcal{F}_N = [g'(\mu_y)]^2 V\{\bar{y}_\pi \mid \mathcal{F}_N\} + O_p(n^{-1.5}) \quad \text{a.s.}$$

9. (Section 4.2) Prove expression (4.2.15). You can use the generalization of (4.2.6) to show that $E\{(\bar{y}_{st}^{(k)} - \bar{y}_{st})^2 \mid \mathcal{F}\} = L^{-1}V\{\bar{y}_{st} \mid \mathcal{F}\}$.

10. (Section 4.2) Consider a Poisson sample with selection probabilities π_i. Let an estimator of the total be

$$\hat{T}_{HT} = \sum_{j \in A} \pi_j^{-1} y_j,$$

and let an estimator of the mean be

$$\bar{y}_\pi = \left(\sum_{j \in A} \pi_j^{-1} \right)^{-1} \sum_{j \in A} \pi_j^{-1} y_j.$$

Let

$$\hat{T}_{HT}^{(k)} = \sum_{j \in A} \psi_j^{(k)} y_j,$$

and

$$\bar{y}_\pi^{(k)} = \left(\sum_{j \in A} \psi_j^{(k)} \right)^{-1} \sum_{j \in A} \psi_j^{(k)} y_j,$$

where

$$\begin{aligned} \psi_j^{(k)} &= \pi_j^{-1} \qquad \text{if } j \neq k \\ &= [1 - (1 - \pi_k)^{0.5}]\pi_k^{-1} \qquad \text{if } j = k. \end{aligned}$$

(a) Show that

$$\sum_{k \in A} (\hat{T}_{HT}^{(k)} - \hat{T}_{HT})^2$$

is an unbiased estimator of $V(\hat{T}_{HT})$.

(b) Give c_k such that $\hat{V}_J\{\bar{y}_\pi\} = \Sigma_{k \in A} c_k (\bar{y}_\pi^{(k)} - \bar{y}_\pi)^2$ is a design consistent estimator of $V(\bar{y}_\pi \mid \mathcal{F})$.

11. (Section 4.5) For the data of Example 4.5.1, how many bootstrap samples are required for the variance estimator to have an efficiency equal to 95% of that of the usual variance estimator? You may use variances for normal variables as approximations and assume common stratum variances.

12. (Section 4.2.3) In Section 4.2.3 the slope of the estimated cumulative distribution function used the order statistics with indexes $i(t)$ and $i(s)$. Show that for a simple random sample and $b = 0.5$, these points correspond to $\hat{Q}(b - t_\alpha \sigma_{ca})$ and $\hat{Q}(b + t_\alpha \sigma_{ca})$, where σ_{ca}^2 is the variance of $\hat{F}(a)$ and $t_\alpha = 2.0$. In Section 1.3 it is stated that $t_\alpha = 2.0$ works well. If it is known that the design effect is about d_e for the survey, what would you use for $i(s)$ and $i(t)$? Recall that the design effect is the ratio of the variance of the estimator under the design to the variance of the estimator for a simple random sample with the same number of sample elements.

This Page Intentionally Left Blank

CHAPTER 5

MODELS USED IN CONJUNCTION WITH SAMPLING

5.1 NONRESPONSE

5.1.1 Introduction

Most surveys of human respondents suffer from some degree of nonresponse. One reason for nonresponse is a failure to contact some elements of the sample. In addition, some people may refuse to participate or fail to respond to certain items in the data collection instrument. Nonresponse is also common in other surveys. An instrument used to record physical data may fail or it may be impossible to record certain data. For example, in an aerial survey of land use it may not be possible to photograph certain selected sampling units where the air space is restricted.

Nonresponse is generally placed in two categories: unit nonresponse and item nonresponse. *Unit nonrespondents*, as the name implies, are those sample elements for which none of the questionnaire information is collected. However, often some information is available. For example, the address of the household is generally available in household surveys, and other information,

Sampling Statistics. By Wayne A. Fuller

such as the physical condition of the residence or the number of residents, may be collected.

Item nonresponse occurs when responses for some items are missing from a questionnaire that is generally acceptable. Such nonresponse is common in self-administered surveys where the respondent can skip questions or sections. Some collection procedures are designed recognizing that people may be reluctant to answer certain questions. For example, questions about income may be placed near the end of the interview and interval categories given as possible answers.

We have introduced the topic of this section using examples of data collected from human respondents, where some people neglect, or refuse, to report for some items. Often, as in two-phase sampling, some data are missing on the basis of the design. Such missing data is called *planned*, or *designed, missingness*. Also, some data collection may not require active participation from the sample unit, as in photo interpretation of an area segment. Nonetheless, with analogy to human respondents, we call an element with a reported value a *respondent* and call an element with a missing value a *nonrespondent*.

The analysis of data with unplanned nonresponse requires the specification of a model for the nonresponse. Models for nonresponse address two characteristics: the probability of obtaining a response and the distribution of the characteristic. In one model it is assumed that the probability of response can be expressed as a function of auxiliary data. The assumption of a second important model is that the expected value of the unobserved variable is related to observable auxiliary data. In some situations models constructed under the two models lead to the same estimator. Similarly, specifications containing models for both components can be developed.

5.1.2 Response models and weighting

A model specifying the probability of responding is most common for unit nonresponse, with the complexity of the model depending on the data available. In two-phase estimation in which the vector $(\mathbf{x}, \mathbf{y})$ is collected on phase 2 units but only $\mathbf{x}$ is observed on the remainder of the phase 1 sample, the probabilities of observing y given $\mathbf{x}$ are known. If the nonresponse is unplanned, it is common to assume that the probability of response is constant in a subpopulation, often called a *cell*. The response cell might be a geographic area or a subpopulation defined by demographic characteristics.

Under the cell response model, the sample is formally equivalent to a two-phase sample and we use the notation of Section 3.3 in our discussion. Assume that the original sample was selected with selection probabilities π_{1i}, that the population is divided into G mutually exclusive and exhaustive response cells,

and that every element in a cell has the same probability of responding. Then the two-phase estimated mean of the form (3.3.13) is

$$\bar{y}_{2p,reg} = \sum_{g=1}^{G} \bar{x}_{1\pi,g} \bar{y}_{2\pi,g}, \tag{5.1.1}$$

where

$$\bar{x}_{1\pi,g} = \left(\sum_{g=1}^{G} \sum_{j \in A_g} \pi_{1j}^{-1} \right)^{-1} \sum_{j \in A_g} \pi_{1j}^{-1},$$

$$\bar{y}_{2\pi,g} = \left(\sum_{j \in A_{Rg}} \pi_{1j}^{-1} \right)^{-1} \sum_{j \in A_{Rg}} \pi_{1j}^{-1} y_j,$$

A_g is the set of sample indices in cell g, A_{Rg} is the set of indices for the respondents in cell g, and $\bar{x}_{1\pi,g}$ is the estimated fraction of the population in cell g. Under the cell response model, the estimated variance of (5.1.1) can be computed with the two-phase formulas of Section 3.3. Of course, the validity of the variance estimator rests on the validity of the cell response model.

If the fractions of the population in the cells are known, the estimated mean

$$\bar{y}_r = \sum_{g=1}^{G} N^{-1} N_g \bar{y}_{2\pi,g} \tag{5.1.2}$$

can be treated as a poststratified estimator under the cell response model. See Section 2.2.3 for variance formulas.

The cell mean model is a special case of the regression model and (5.1.2) is the corresponding special case of the regression estimator. To consider general regression estimation, let a vector of auxiliary variables, $\mathbf{x}$, be available for both respondents and nonrespondents, and let the population mean of $\mathbf{x}$, denoted by $\bar{\mathbf{x}}_N$, be known. Then a regression estimator using the inverses of the original probabilities as weights is

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}, \tag{5.1.3}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_R' \mathbf{D}_{\pi R}^{-1} \mathbf{X}_R \right)^{-1} \mathbf{X}_R \mathbf{D}_{\pi R}^{-1} \mathbf{y}_R,$$

$\mathbf{X}_R$ is the $n_R \times k$ matrix of observations on the respondents, n_R is the total number of respondents, $\mathbf{D}_{\pi R}$ is the diagonal matrix of original selection probabilities for respondents, and $\mathbf{y}_R$ is the vector of observations for respondents.

Assume that there is a vector α such that

$$\mathbf{x}_i\alpha = \pi_{2i|1i}^{-1}, \tag{5.1.4}$$

where $\pi_{2i|1i}$ is the conditional probability that element i responds given that it is selected for the original sample. Note that condition (5.1.4) holds if a vector of indicator variables is used to construct estimator (5.1.2). Given (5.1.4), the regression estimator (5.1.3) is consistent. Furthermore, there is an appropriate regression estimator of variance if the finite population correction can be ignored.

Theorem 5.1.1. Let a sequence of finite populations and samples be such that the variance of the Horvitz–Thompson estimator of a mean for a complete sample has a variance that is $O_p(n^{-1})$, the Horvitz–Thompson estimator of the variance of a mean for a complete sample has a variance that is $O_p(n^{-3})$, and the limiting distribution of the properly standardized Horvitz–Thompson mean of a complete sample is normal. Assume that for a sample with nonresponse, (5.1.4) holds and that

$$K_L < \pi_{2i|1i} < K_U \tag{5.1.5}$$

for positive constants K_L and K_U. Assume that responses are independent, that $\bar{\mathbf{x}}_N$ is known, that there is a $\boldsymbol{\lambda}$ such that $\mathbf{x}_i\boldsymbol{\lambda} = 1$ for all i, and let the regression estimator be defined by (5.1.3). Then

$$\bar{y}_{reg} - \bar{y}_N = N^{-1} \sum_{i \in A_R} \pi_{2i}^{-1} e_i + O_p(n^{-1}), \tag{5.1.6}$$

where A_R is the set of indices of the respondents, $e_i = y_i - \mathbf{x}_i\beta_N$, $\pi_{2i} = \pi_{1i}\pi_{2i|1i}$, π_{1i} is the probability that element i is included in the original sample, and

$$\boldsymbol{\beta}_N = \left(\sum_{i \in U} \mathbf{x}_i' \pi_{2i|1i} \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i' \pi_{2i|1i} y_i. \tag{5.1.7}$$

Furthermore,

$$\hat{V}_{HT} \left\{ n^{1/2} \sum_{j \in A_R} \hat{b}_j \right\} = V_\infty \{ n^{1/2} (\bar{y}_{reg} - \bar{y}_N) \mid \mathcal{F}_N \} - N^{-2} \sum_{i \in U} (\pi_{1i} - \pi_{2i}) \pi_{2i}^{-1} e_i^2 + O_p(n^{-3/2}), \tag{5.1.8}$$

where

$$\hat{b}_j = \bar{\mathbf{x}}_N \left(\sum_{i \in A_R} \mathbf{x}_i' \pi_{1i}^{-1} \mathbf{x}_i \right)^{-1} \mathbf{x}_j' \pi_{1j}^{-1} \hat{e}_j,$$

$$\hat{V}_{HT} \left(\sum_{j \in A_R} \hat{b}_j \right) = \sum_{i \in A_R} \sum_{j \in A_R} \pi_{1ij}^{-1} \left(\pi_{1ij} - \pi_{1i} \pi_{1j} \right) \hat{b}_i \hat{b}_j, \tag{5.1.9}$$

$\hat{e}_j = y_j - \mathbf{x}_j \hat{\boldsymbol{\beta}}$, and $V_\infty \left\{ n^{1/2} \left(\bar{y}_{reg} - \bar{y}_N \right) \mid \mathcal{F}_N \right\}$ is the variance of the limiting distribution of $n^{1/2} \left(\bar{y}_{reg} - \bar{y}_N \right)$ conditional on $\mathcal{F}_N$.

Proof. The conditional expectations of the components of $\hat{\boldsymbol{\beta}}$ of (5.1.3) are

$$E \left\{ \sum_{i \in A_R} \mathbf{x}_i' \pi_{1i}^{-1} \mathbf{x}_i \mid \mathcal{F} \right\} = \sum_{i \in U} \mathbf{x}_i' \pi_{2i|1i} \mathbf{x}_i$$

and

$$E \left\{ \sum_{i \in A_R} \mathbf{x}_i' \pi_{1i}^{-1} y_i \mid \mathcal{F} \right\} = \sum_{i \in U} \mathbf{x}_i' \pi_{2i|1i} y_i.$$

By the assumption that the Horvitz–Thompson estimators of means have errors that are $O_p \left(n^{-1/2} \right)$,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = O_p(n^{-1/2}). \tag{5.1.10}$$

By (5.1.4), $\Sigma_{i \in U} e_i = 0$, and by Theorem 2.2.1, $n^{1/2} \left(\bar{y}_{reg} - \bar{y}_N \right)$ has a normal distribution in the limit.

To obtain representation (5.1.6), assume, without loss of generality, that the first element of $\mathbf{x}$ is $\pi_{2i|1i}^{-1}$. Define a transformation of the original $\mathbf{x}$-vector by $\mathbf{z}_i = \mathbf{x}_i \hat{\boldsymbol{\Lambda}}$, where

$$\begin{aligned} z_{1i} &= \pi_{2i|1i}^{-1}, \\ z_{ji} &= x_{ji} + z_{1i} \hat{\lambda}_{1j} \end{aligned}$$

and

$$\hat{\lambda}_{1j} = - \left(\sum_{i \in A_R} z_{1i}^2 \pi_{1i}^{-1} \right)^{-1} \sum_{i \in A_R} z_{1i} x_{ji} \pi_{1i}^{-1}$$

for $j = 2, 3, \ldots, k$. Then

$$\hat{\mathbf{\Lambda}}^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right) = \begin{pmatrix} \sum_{i \in A_R} z_{1i}^2 \pi_{1i}^{-1} & \mathbf{0} \\ \mathbf{0}' & \sum_{i \in A_R} \mathbf{z}_{2i}' \pi_{1i}^{-1} \mathbf{z}_{2i} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i \in A_R} z_{1i} \pi_{1i}^{-1} e_i \\ \sum_{i \in A_R} \mathbf{z}_{2i}' \pi_{1i}^{-1} e_i \end{pmatrix},$$

where $\bar{y}_{reg} - \bar{y}_N = \bar{\mathbf{x}}_N \hat{\mathbf{\Lambda}} \hat{\mathbf{\Lambda}}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)$, and $\mathbf{z}_i = (z_{1i}, \mathbf{z}_{2i}) = (z_{1i}, z_{2i}, \ldots, z_{ki})$. Now

$$\begin{aligned} \hat{\lambda}_{1j} &= -\left(\textstyle\sum_{i \in U} \pi_{2i|1i}^{-1}\right)^{-1} \sum_{i \in U} x_{ji} + O_p\left(n^{-1/2}\right) \\ &= -\bar{z}_{1,N}^{-1} \bar{x}_{j,N} + O_p\left(n^{-1/2}\right) \end{aligned}$$

for $j = 2, 3, \ldots, k$, and

$$\bar{\mathbf{x}}_N \hat{\mathbf{\Lambda}} = (\bar{z}_{1,N}, \mathbf{0}) + O_p(n^{-1/2}).$$

It follows that

$$\bar{y}_{reg} - \bar{y}_N = N^{-1} \sum_{i \in A_R} \pi_{1i}^{-1} z_{1i} e_i + O_p(n^{-1})$$

because

$$\begin{aligned} \bar{z}_{1,N} \left(\sum_{i \in A_R} \pi_{1i}^{-1} z_{1i}^2 \right)^{-1} &= N^{-1} \bar{z}_{1,N} \bar{z}_{1,HT}^{-1} \\ &= N^{-1}[1 + O_p(n^{-1/2})], \end{aligned}$$

where

$$\bar{z}_{1,HT} = N^{-1} \sum_{i \in A_R} \pi_{2i}^{-1} z_{1i},$$

and result (5.1.6) is proven.

To prove (5.1.8), note that

$$\sum_{i \in A_R} N^{-1} \pi_{1i}^{-1} z_{1i} e_i = \sum_{i \in A_R} w_{2i} e_i,$$

where $w_{2i} = N^{-1} \pi_{2i}^{-1}$, is a design linear estimator and

$$V\left\{ \sum_{i \in A_R} w_{2i} e_i \mid \mathcal{F} \right\} = \sum_{i \in U} \sum_{j \in U} (\pi_{2ij} - \pi_{2i}\pi_{2j})\, w_{2i} w_{2j} e_i e_j$$

$$= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \pi_{2i|1i}\pi_{2j|1j}\left(\pi_{1ij}-\pi_{1i}\pi_{1j}\right) w_{2i}w_{2j}e_ie_j + \sum_{j\in U}\left(\pi_{2j}-\pi_{2j}^2\right) w_{2j}^2e_j^2,$$

where π_{1ij} is the probability that elements i and j are in the original sample, $\pi_{2i} = \pi_{1i}\pi_{2i|1i}$, and π_{2ij} is the probability that both i and j are in the sample and both respond. We have $\pi_{2ij} = \pi_{1ij}\pi_{2i|1i}\pi_{2j|1j}$ because responses are independent.

The expectation of the Horvitz–Thompson variance estimator for $\Sigma w_{2i}e_i$ constructed with π_{1i} and π_{1ij} is

$$E\left\{\sum_{i\in A_R}\sum_{j\in A_R}\pi_{1ij}^{-1}\left(\pi_{1ij}-\pi_{1i}\pi_{1j}\right)w_{2i}e_iw_{2j}e_j \mid \mathcal{F}\right\}$$

$$= \sum_{i\in U}\left(\pi_{1i}-\pi_{1i}^2\right)\pi_{2i|1i}w_{2i}^2e_i^2 + \sum_{i\in U}\sum_{\substack{j\in U\\ i\neq j}}\pi_{2i|1i}\pi_{2j|1j}\left(\pi_{1ij}-\pi_{1i}\pi_{1j}\right)w_{2i}e_iw_{2j}e_j$$

$$= \sum_{i\in U}\sum_{j\in U}\left(\pi_{2ij}-\pi_{2i}\pi_{2j}\right)w_{2i}e_iw_{2j}e_j + \sum_{i\in U}\pi_{2i}\left(\pi_{2i}-\pi_{1i}\right)w_{2i}^2e_i^2.$$

The variance estimator constructed with $\hat{e}_t$ is asymptotically equivalent to that constructed with e_t. See the proofs of Theorems 2.2.1 and 2.2.2. Therefore, result (5.1.8) is proven. ■

In the variance estimator (5.1.9), $\hat{b}_i$ is of the form $\tilde{w}_i\hat{e}_i$, where $\tilde{w}_i$ is the regression weight. The $\tilde{w}_i$ must be retained in the variance calculations because the error in $\hat{\beta}$ contributes an $O(n^{-1})$ term to the variance.

The second term in (5.1.8) can be written as

$$-N^{-2}\sum_{i\in U}\pi_{2i|1i}^{-1}\left(1-\pi_{2i|1i}\right)e_i^2 \tag{5.1.11}$$

and will be relatively small for small sampling rates. Given (5.1.4), the $\pi_{2i|1i}$ can be estimated by expressing the response indicator as a function of $(\mathbf{x}_i\boldsymbol{\alpha})^{-1}$ and estimating $\boldsymbol{\alpha}$. Then, using the $\hat{e}_i$ of (5.1.9), expression (5.1.11) can be estimated.

5.2 IMPUTATION

5.2.1 Introduction

If a modest number of variables are missing from otherwise complete questionnaires, one method of implementing estimation is to replace individual missing values with "estimates." The objective is to use the replacement values as if they were observed values in a full-sample estimation procedure. The replacement values are called *imputed values*.

A goal of the imputation procedure is to construct imputed values that will yield efficient estimators of parameters for which estimators would be available from the full sample. Second, it should be possible to estimate the variance of the imputed estimators. In a typical survey situation, the survey statistician makes available to analysts a data set with weights and the values of a set of characteristics for the sample elements. The statistician may know some of the estimates that will be constructed from the data set, but seldom will the full set of possible estimates be known. Thus, the objective is to design an imputation procedure such that the imputed data set will be appropriate for both planned and unplanned estimates.

One may ask; "If one must build a model for the imputation, why not simply use the estimator obtained from the model?" The answer is in the many ways in which survey sample data are used. If a single variable is of importance, a model will be developed for that variable and estimates generated directly from the model. If the objective is to create a data set for general use, replacing the missing values with model-imputed values gives such a data set. Of course, the imputed values must be identified, and the model used for imputation must be made available to the end users.

Consider a simple random sample of n elements in which the y value for m elements is not observed and $r = n - m$ are observed. Assume that the fact that an element is not observed is independent of y. Then the r observations are a simple random sample of size r and the natural estimator of the mean of y is

$$\hat{\mu}_y = r^{-1} \sum_{i \in A_R} y_i, \tag{5.2.1}$$

where A_R is the set of indices of units observed and responding. Now assume that we wish to impute values for the missing values so that estimates based on the entire set of n elements will be equal to estimates based only on the responding units. If the only parameter to be estimated is the mean, replacing the missing values with the mean of the responding values will give a mean of the completed sample that is equal to the mean of the responding units.

However, if other characteristics of the distribution are of interest, estimates based on the mean-imputed data set will be seriously biased. For example, the large fraction of imputed values equal to the mean will bias all estimated quantiles. To meet the goal of multiple use, the imputed data set should provide a good estimate of any function of the variables. That is, the imputed data set should give a good estimate of the distribution function.

There are a number of imputation procedures that furnish good estimates of the distribution function. For a simple random sample and random nonresponse, one procedure is to choose randomly one of the respondents for each nonrespondent and use the respondent value for the missing value. Let y_{iI}, $i = r+1, r+2, \ldots, n$, be the m imputed values and let the mean computed with imputed data be

$$\begin{aligned} \bar{y}_I &= n^{-1}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1} y_{iI}\right) \\ &=: n^{-1}\left(r\bar{y}_r + m\bar{y}_{m,I}\right), \end{aligned} \tag{5.2.2}$$

where

$$\bar{y}_{m,I} = m^{-1}\sum_{i=r+1}^{n} y_{iI}.$$

Because y_{iI} is a random selection from the respondents, the expected value for any percentile is that for the respondents.

Procedures that use values from the sample as imputed values are called *hot deck imputation procedures.* In a situation such as that just described, the element with a missing value is called the *recipient* and the element providing the value for imputation is called the *donor.* The hot deck name was originally used by the U.S. Census Bureau to describe an imputation procedure when computer cards were used in processing data. The donor was an element that was close to the recipient in the deck of cards. An advantage of hot deck procedures is that the imputed values are values that appear in the data set. It is possible for some imputation procedures to generate impossible responses.

The random selection of donors gives an imputed data set with the correct expectation under the model, but the random selection increases the variance of an estimator relative to an estimator constructed directly from the respondents. If response is independent of y, and we use a random replacement selection of donors for a simple random sample, the conditional variance of the mean of the imputed values is

$$V\{\bar{y}_{m,I} - \bar{y}_r \mid \mathbf{y}_r\} = r^{-1}m^{-1}\sum_{i=1}^{r}(y_i - \bar{y}_r)^2, \tag{5.2.3}$$

where $\mathbf{y}_r = (y_1, y_2, \ldots, y_r)$ is the set of respondents and $\bar{y}_r$ is the mean of the respondents. See Section 1.2.5. Then

$$\begin{aligned} V\{\bar{y}_I \mid (\mathcal{F}, m)\} &= V\{E(\bar{y}_I \mid \mathbf{y}_r)\} + E\{V(\bar{y}_I \mid \mathbf{y}_r)\} \\ &= \left(r^{-1} - N^{-1}\right) S_y^2 + n^{-2} m r^{-1} (r-1) S_y^2, \quad (5.2.4) \end{aligned}$$

where the conditioning notation denotes the variance for samples of size n with exactly m missing.

There are a number of ways to select donors to reduce the imputation variance. One possibility is to use a more efficient sampling method, such as nonreplacement sampling, to select the donors. If m is an integer multiple of r, the imputed estimator of the mean based on without-replacement sampling is equal to the mean of the y values for the respondents. If m is not an integer multiple of r, there is an increase in variance relative to the mean of the respondents. See Exercise 1. Also, one can reduce the variance by using stratified or systematic selection of donors.

Another way to reduce the variance due to imputation is to impute more than one value for each respondent. In a procedure proposed by Rubin (1987) and called *multiple imputation*, the imputation operation is repeated a number of times to create multiple sets of imputed data. Also see Little and Rubin (2002) and Schafer (1997). In the next section we consider a procedure called *fractional imputation*, suggested by Kalton and Kish (1984).

5.2.2 Fractional imputation

In fractional imputation, a number, say M, of donors is used for each recipient and each donor is given a fractional weight, where the fractions sum to 1. Consider a simple random sample with random nonresponse and r respondents. Instead of selecting a single donor for each recipient we assign all respondents to each recipient and give a relative weight of r^{-1} to each donor value. The resulting data set has $r + mr$ vectors where there are now r vectors for each of the elements with missing y. For such a data set, the estimate for any function of y is exactly the same as that obtained by tabulating the sample composed of the respondents. Kim and Fuller (2004) call the procedure *fully efficient fractional imputation* because there is no variance due to the selection of imputed values. Fully efficient fractional imputation is not common because of the size of the resulting data set. However, very efficient procedures can be constructed with two to five imputed values per respondent.

Table 5.1 Sample with Missing Data

Observation	Weight	Cell for x	Cell for y	x	y
1	0.10	1	1	1	7
2	0.10	1	1	2	M
3	0.10	1	2	3	M
4	0.10	1	1	M	14
5	0.10	1	2	1	3
6	0.10	2	1	2	15
7	0.10	2	2	3	8
8	0.10	2	1	3	9
9	0.10	2	2	2	2
10	0.10	2	1	M	M

Example 5.2.1. We use a small artificial data set to illustrate the use of fractional imputation for the calculation of fully efficient estimators and for the calculation of estimated variances. Assume that the data in Table 5.1 constitute a simple random sample and ignore any finite population correction. Variable x is a categorical variable with three categories, identified as 1, 2, and 3. The sample is divided into two imputation cells for this variable. In imputation cell 1 the fraction in the three categories is 0.50, 0.25, and 0.25 for categories 1, 2, and 3, respectively. In imputation cell 2 the fractions are 0.00, 0.50, and 0.50 for categories 1, 2, and 3, respectively. For the missing value of x for observation 4, we impute three values, one for each category, and assign weights for the fractions equal to the observed fractions. All other data are the same for each "observation" created. See the three lines for original observation 4 in Table 5.2. The estimated fraction in a category for imputation cell 1 calculated using the imputed data and the fractional weights is the same as the fraction for the respondents.

The fully efficient fractional imputation of y for y-imputation cell 1 would require four imputed values. That would not be a problem for this small data set, but to illustrate the computation of efficient estimators with a sample of donors, we select a sample of three of the four available donors. See the imputed values for observation 3 in Table 5.2.

Several approaches are possible for the situation in which two items are missing, including the definition of a third set of imputation cells for such cases. Because of the small size of our illustration, we impute under the assumption that x and y are independent within cells. Thus, we impute four values for observation 10. For each of the two possible values of x we impute two possible values for y. One of the pair of imputed y values is chosen to be

less than the mean of the responses, and one is chosen to be greater than the mean. See the imputed values for observation 10 in Table 5.2.

Table 5.2 Fractionally Imputed Data Set

Observation	Donor x	Donor y	w^*_{ij0}	Final Weight	Cell for x	Cell for y	x	y
1	0	0	–	0.1000	1	1	1	7
2	0	1	0.3333	0.0289	1	1	2	7
	0	6	0.3333	0.0396	1	1	2	15
	0	8	0.3333	0.0315	1	1	2	9
3	0	5	0.3333	0.0333	1	2	3	3
	0	7	0.3333	0.0333	1	2	3	8
	0	9	0.3333	0.0333	1	2	3	2
4	†	0	0.5000	0.0500	1	1	1	14
	†	0	0.2500	0.0250	1	1	2	14
	†	0	0.2500	0.0250	1	1	3	14
5	0	0	–	0.1000	1	2	1	3
6	0	0	–	0.1000	2	1	2	15
7	0	0	–	0.1000	2	2	3	8
8	0	0	–	0.1000	2	1	3	9
9	0	0	–	0.1000	2	2	2	2
10	†	8	0.2500	0.0225	2	1	2	9
	†	4	0.2500	0.0275	2	1	2	14
	†	1	0.2500	0.0209	2	1	3	7
	†	6	0.2500	0.0291	2	1	3	15

†All relevant values of x are imputed for every missing observation.

To create fully efficient estimates of the mean of y, the cell mean of the imputed data should be the same as the mean of the respondents in the cell. To define such a data set, we use the regression estimator and require the fractional weights to sum to 1 for each observation. For observation 10, we require the two weights for each category to sum to the fraction (0.5) for the category. In using regression to adjust the fractional weights, one can adjust all weights subject to the restriction that the sum of the fractional weights is 1 for each person or one can adjust the weights for each person. Because of the small number of imputed values per person we use the second approach.

Let B_g be the set of indices of elements in cell g that have at least one characteristic imputed, let $\mathbf{z}_{g[i]j}$ be the ith imputed vector of characteristics for which at least one value has been imputed, let w_j be the weight for observation j, and let w^*_{ij0} be an initial fractional weight for the ith imputed

vector for element j, where

$$\sum_{i \in A_{Ij}} w^*_{ij0} = 1$$

and A_{Ij} is the set of indices of imputed values for observation j. The fractional weight for imputed value i of observation j in cell g is

$$w^*_{ij} = w^*_{ij0} + (\bar{\mathbf{z}}_{FE,g} - \bar{\mathbf{z}}_g)\mathbf{S}^{-1}_{\mathbf{zz}g} w^*_{ij0}(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})', \qquad (5.2.5)$$

where

$$\mathbf{S}_{\mathbf{zz}g} = \sum_{j \in B_g} \sum_{i \in A_{Ij}} w^*_{ij0}(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})'(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j}),$$

$$\bar{\mathbf{z}}_{g\cdot j} = \sum_{i \in A_{Ij}} w^*_{ij0}\mathbf{z}_{g[i]j},$$

$$\bar{\mathbf{z}}_g = \left(\sum_{j \in B_g} w_j\right)^{-1} \sum_{j \in B_g} w_j \bar{\mathbf{z}}_{g\cdot j},$$

$\bar{\mathbf{z}}_{FE,g}$ is the fully efficient weighted mean of the respondents in cell g, and $\bar{\mathbf{z}}_{g\cdot j}$ is the mean of imputed values for observation j. If the rth characteristic is observed, $\bar{\mathbf{z}}_{g\cdot jr}$ is the value observed and $\bar{\mathbf{z}}_{g[i]jr} - \bar{\mathbf{z}}_{g\cdot jr} = 0$.

Because all values of y in cell 2 were used to impute for observation 3, we need only compute weights for cell 1 for y. The mean of imputed values for observation 2 is 10.333 and the two means for observation 10 are 11.500 and 11.000. The mean of the observed values of y for cell 1 is 11.25, the weighted mean of the imputed values is $\bar{z}_1 = 10.7917$, and the weighted sum of squares is 23.1618. The adjusted fractions are 0.2886, 0.3960, and 0.3154 for observation 2 and 0.2247, 0.2753, 0.2095, and 0.2905 for observation 10, in the order that they appear in the table. The weights, called *final weights* in the table, are the products $w^*_{ij}w_j$.

We use jackknife replicates to illustrate variance estimation with fractional imputation. The procedure is analogous to that for two-phase samples. As the first step we create a standard jackknife replicate by deleting an observation. Tables 5.3 and 5.4 give the estimates for the cell means for the observed values for the 10 replicates. For example, when observation one is deleted, the replicate mean of y for y-cell 1 is 12.67, and the fractions for x-cell 1 are $0.3\bar{3}$, $0.3\bar{3}$, and $0.3\bar{3}$ for categories 1, 2, and 3, respectively.

Using the regression procedure, the fractional weights of each replicate are adjusted to give the mean for that replicate. For example, the fractional weights of the imputed y-values for observations 2 and 10 of replicate 1 are

modified so that

$$\frac{\hat{y}_{2I} + y_3 + y_6 + y_8 + \hat{y}_{10}}{5} = 12.67,$$

where $\hat{y}_{jI} = \Sigma_{i \in A} w_{ij}^* y_i$. Table 5.5 contains the replicate weights.

Table 5.3 Jackknife Replicate Cell Means for y-Variable

	Replicate									
Cell	1	2	3	4	5	6	7	8	9	10
1	12.67	11.25	11.25	10.33	11.25	10.00	11.25	12.00	11.25	11.25
2	4.33	4.33	4.33	4.33	5.00	4.33	2.50	4.33	5.50	4.33

Table 5.4 Jackknife Replicate Fractions for x-Categories

	Cat.	Replicate									
Cell	of x	1	2	3	4	5	6	7	8	9	10
1	1	0.33	0.67	0.67	0.50	0.33	0.50	0.50	0.50	0.50	0.50
	2	0.33	0.00	0.33	0.25	0.33	0.25	0.25	0.25	0.25	0.25
	3	0.33	0.33	0.00	0.25	0.33	0.25	0.33	0.25	0.25	0.25
2	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.50	0.50	0.50	0.50	0.50	0.33	0.67	0.67	0.33	0.50
	3	0.50	0.50	0.50	0.50	0.50	0.67	0.33	0.33	0.67	0.50

The final data set with the weights of Table 5.2 and the replicate weights of Table 5.5 can be used to compute all estimators and all estimated variances for which the jackknife is appropriate. For example, the estimated cumulative distribution function for y and its variance could be computed.

The jackknife estimated variance for the mean of y is

$$\hat{V}_{JK}(\bar{y}_{FI}) = \sum_{k=1}^{10} 0.9(\bar{y}_{FI}^{(k)} - \bar{y}_{FI})^2 = 3.1095$$

and the two-phase variance estimator is

$$\hat{V} = \frac{1}{n}\sum_{g=1}^{2} \frac{n_g}{n}(\bar{y}_{Rg} - \bar{y}_{FE})^2 + \sum_{g=1}^{2} \left(\frac{n_g}{n}\right)^2 \frac{1}{r_g} s_{Rg}^2 = 3.043,$$

where s^2_{Rg} is the within-cell sample variance for cell g. The two estimates differ by the amount

$$\sum_{g=1}^{2} \left[(r_g - 1)^{-1} r_g (n-1)^{-1} n - 1\right] s^2_{Rg}.$$

See Section 4.4. ■■

Table 5.5 Jackknife Weights† for Fractionally Imputed Data

	Replicate									
Obs	1	2	3	4	5	6	7	8	9	10
1	0	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
2	0.17	0	0.32	0.42	0.32	0.46	0.32	0.24	0.32	0.27
	0.66	0	0.44	0.30	0.44	0.25	0.44	0.55	0.44	0.51
	0.29	0	0.35	0.39	0.35	0.40	0.35	0.32	0.35	0.33
3	0.37	0.37	0	0.37	0.32	0.37	0.50	0.37	0.29	0.37
	0.37	0.37	0	0.37	0.50	0.37	0.01	0.37	0.60	0.37
	0.37	0.37	0	0.37	0.29	0.37	0.60	0.37	0.22	0.37
4	0.37	0.74	0.74	0	0.37	0.56	0.56	0.56	0.56	0.56
	0.37	0	0.37	0	0.37	0.28	0.28	0.28	0.28	0.28
	0.37	0.37	0	0	0.37	0.28	0.28	0.28	0.28	0.28
5	1.11	1.11	1.11	1.11	0	1.11	1.11	1.11	1.11	1.11
6	1.11	1.11	1.11	1.11	1.11	0	1.11	1.11	1.11	1.11
7	1.11	1.11	1.11	1.11	1.11	1.11	0	1.11	1.11	1.11
8	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0	1.11	1.11
9	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0	1.11
10	0.16	0.28	0.28	0.31	0.28	0.23	0.35	0.30	0.15	0
	0.39	0.28	0.28	0.25	0.28	0.14	0.39	0.44	0.22	0
	0.09	0.28	0.28	0.32	0.28	0.44	0.15	0.07	0.32	0
	0.46	0.28	0.28	0.23	0.28	0.30	0.22	0.30	0.42	0

†Multiply entries by 0.10 for mean estimation. Weights are rounded.

5.2.3 Nearest-neighbor imputation

Nearest-neighbor imputation is a hot deck procedure in which a distance measure, defined on the basis of observed characteristics, is used to define the donor. The respondents closest to the element with a missing value act as donors. It is common practice to use a single donor, but we suggest that two

or more donors be used for each recipient. The use of more than one donor facilitates variance estimation and generally improves efficiency.

Assume that the finite universe is generated by a stochastic mechanism and that a distance measure is defined for the elements. Let a neighborhood of element g be composed of elements that are close to element g, and let

$$\begin{aligned} \mu_g &= E\{y_j \mid j \in B_g\}, \\ \sigma_g^2 &= E\{(y_j - \mu_j)^2 \mid j \in B_g\}, \end{aligned}$$

where B_g is the set of indices for the elements in the neighborhood of element g. One might suppose that there would be some correlation among elements in the neighborhood, with elements that are close having a positive correlation, but we will assume that neighborhoods are small enough so that the correlation can be ignored. We assume that an adequate approximation for the distribution of elements in the neighborhood is

$$y_j \sim ii(\mu_g,\ \sigma_g^2), \qquad j \in B_g, \tag{5.2.6}$$

where $\sim ii$ denotes independent identically distributed. We assume that response is independent of the y values so that the distribution (5.2.6) holds for both recipients and donors. Our results are obtained under the working assumption (5.2.6). For the assumption to hold exactly for every neighborhood, the assumption must hold globally or the neighborhoods must be mutually exclusive. See Chen and Shao (2000, 2001) for conditions under which it is reasonable to use (5.2.6) as an approximation.

Let a probability sample be selected from the finite universe with selection probabilities π_j. Let $\hat{\theta}_n$ be a design linear estimator based on the full sample,

$$\hat{\theta}_n = \sum_{i \in A} w_i y_i, \tag{5.2.7}$$

and let $V\{\hat{\theta}_n\}$ be the variance of the full-sample estimator. Under model (5.2.6) we can write

$$y_i = \mu_i + e_i, \tag{5.2.8}$$

where the e_i are independent $(0,\ \sigma_i^2)$ random variables and μ_i is the neighborhood mean. Then, under model (5.2.6), the variance of $\hat{T}_y = \Sigma_{i \in A} w_i y_i$ is

$$V\left\{\sum_{i \in A} w_i y_i - T_y\right\} = V\left\{\sum_{i \in A} w_i \mu_i - T_\mu\right\} + E\left\{\sum_{i \in A} (w_i^2 - w_i)\sigma_i^2\right\}, \tag{5.2.9}$$

where T_y is the population total of the y_i and T_μ is the population total of the μ_i. Note that the variance is an unconditional variance.

Assume that y is missing for some elements and assume that there are always at least M observations on y in the neighborhood of each missing value. Let an imputation procedure be used to assign M donors to each recipient. Let w_{ij}^* be the fraction of the weight allocated to donor i for recipient j. Then

$$\alpha_i = \sum_{j \in A} w_j w_{ij}^* \tag{5.2.10}$$

is the total weight for donor i, where it is understood that $w_{ii}^* = 1$ for a donor donating to itself. Thus, the imputed linear estimator is

$$\hat{\theta}_I = \sum_{j \in A} w_j y_{Ij} = \sum_{i \in A_R} \alpha_i y_i, \tag{5.2.11}$$

where A_R is the set of indices of the respondents, the mean imputed value for recipient j is

$$y_{Ij} = \sum_{i \in A} w_{ij}^* y_i, \tag{5.2.12}$$

and $y_{Ij} = y_j$ if j is a respondent. Then, under model (5.2.6),

$$V\{\hat{T}_{yI} - T_y\} = V\left\{ \sum_{i \in A} w_i \mu_i - T_\mu \right\} + E\left\{ \sum_{i \in A_R} (\alpha_i^2 - \alpha_i)\sigma_i^2 \right\}, \tag{5.2.13}$$

where A_R is the set of indices of the respondents and $\hat{T}_{yI}$ is the estimated total based on imputed data. The increase in variance due to imputing for missing values is, from (5.2.9),

$$\sum_{i \in A_R} (\alpha_i^2 - \alpha_i)\sigma_i^2 - \sum_{i \in A} (w_i^2 - w_i)\sigma_i^2.$$

To use replication to estimate the variance of the imputed estimator, let a replication variance estimator for the complete sample be

$$\hat{V}_1\{\hat{\theta}\} = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \tag{5.2.14}$$

where $\hat{\theta}$ is the full-sample estimator, $\hat{\theta}^{(k)}$ is the kth estimate of θ_N based on the kth replicate, L is the number of replicates, and c_k is a factor associated

with replicate k determined by the replication method. If imputed data are used in (5.2.14) for $\hat{\theta} = \hat{T}_{yI}$,

$$E\{\hat{V}_1(\hat{T}_{yI} - T_y)\} = V\left\{\sum_{i\in A} w_i\mu_i - T_\mu\right\} + E\left\{\sum_{k=1}^{L}\sum_{i\in A_R} c_k(\alpha_{i1}^{(k)} - \alpha_i)^2\sigma_i^2\right\}, \quad (5.2.15)$$

where $\alpha_{i1}^{(k)} = \sum_j w_j^{(k)} w_{ij}^*$ and $w_j^{(k)}$ is the weight for element j in replicate k. The estimator $\hat{V}_1(\hat{\theta})$ computed as if imputed data were observed is sometimes called the *naive variance estimator*. We outline a replication procedure that produces unbiased variance estimates.

For nearest-neighbor imputed data, there are three types of observations in the data set:

1. Respondents that act as donors for at least one recipient
2. Respondents that are never used as donors
3. Recipients

The original full-sample replicate weights will be used for the last two types. For donors, the initial fractional weights w_{ij}^* in replicate k will be modified so that we obtain the correct expectation. Let superscript k denote the replicate where element k is in the deleted set. Following Kim and Fuller (2004), the fractions assigned to donor k are changed so that the expected value of the sum of squares is changed by the proper amount. First, the full-sample replicates for the variance estimator (5.2.14) are computed, and the sum of squares for element i computed as

$$\sum_{k=1}^{L} c_k(\alpha_{i1}^{(k)} - \alpha_i)^2 = \Phi_i, \qquad i \in A_R, \quad (5.2.16)$$

where $\alpha_{i1}^{(k)}$ is defined following (5.2.15).

In the second step, the fractions for replicates for donors are modified. Let R_k be the set of indices of recipients for which k is a donor. We use k as the index for the replicate and for the donor. Let the new fractional weight in replicate k for the value donated by element k to recipient j be

$$w_{kj}^{*(k)} = w_{kj}^* b_k, \quad (5.2.17)$$

where b_k is to be determined. For two donors to each recipient the new fractional weight for the other donor, denoted by t, is

$$w_{tj}^{*(k)} = 1 - b_k w_{kj}^*. \tag{5.2.18}$$

For $w_{kj}^* = 0.5$, $w_{kj}^{*(k)} = 0.5b_k$ and $w_{tj}^{*(k)} = (1 - 0.5b_k)$. Then, by (5.2.15), the b_k that gives the correct sum of squares is the solution to the quadratic equation

$$\begin{aligned} & c_k \left(w_k^{(k)} + b_k \sum_{\substack{j \in R_k \\ j \neq k}} w_j^{(k)} w_{ij}^* - \alpha_k \right)^2 \\ & + \sum_{t \in D_{Rk}} c_t \left(w_t^{(k)} + \sum_{j \in R_t \cap R_k} w_j^{(k)} (1 - w_{kj}^* b_k) - \alpha_t \right)^2 \\ & - c_k (\alpha_{k1}^{(k)} - \alpha_k)^2 - \sum_{j \in D_{Rk}} c_j (\alpha_{j1}^{(k)} - \alpha_j)^2 \\ & = \alpha_k^2 - \alpha_k - \Phi_k, \end{aligned} \tag{5.2.19}$$

where t is used as the index for the donors other than k that donate to j, and D_{Rk} is the set of donors other than k that donate to recipients that receive a value from donor k. The difference $\Phi_k - (\alpha_k^2 - \alpha_k)$ is the difference between the sum of squares for the naive estimator and the sum of squares desired for observation k. Under the assumption of a common variance in a neighborhood and the assumption that the full-sample variance estimator $\hat{V}_1(\hat{\theta})$ of (5.2.14) is unbiased, the variance estimator with b_k defined by (5.2.19) is unbiased for the variance of the mean of the imputed sample. The procedure corrects weights within each replicate and does not force the sum of squares over replicates for observation i to be equal to α_i^2.

Example 5.2.2. Table 5.6 contains an illustration data set of six observations. The variable x_i is observed on all six, but the variable y is missing for observations 3 and 6. The variable x is used to determine distance and, in Euclidean distance, observation 2 is closest to observation 3. Therefore, using the nearest-neighbor rule, we replace the missing value for observation 3 by the value of observation 2. In the same way, observation 5 is closest to observation 6, so the missing value for observation 6 is replaced by 2.3, the value of y for observation 5. If only the nearest neighbor is used for imputation, we obtain the imputed data set of the last column. The weight of 1/6 would be the weight for a simple mean.

Table 5.6 Data Set

Obs.	Weight	x_i	y_i	y_{Ii}
1	$0.16\bar{6}$	0.9	0.7	0.7
2	$0.16\bar{6}$	1.1	1.0	1.0
3	$0.16\bar{6}$	1.3	M	1.0
4	$0.16\bar{6}$	2.2	1.9	1.9
5	$0.16\bar{6}$	2.6	2.3	2.3
6	$0.16\bar{6}$	3.1	M	2.3

Table 5.7 contains the imputed observations when two imputations are made for each missing value. The second nearest neighbor for observation 2 is observation 4, where $\mid x_2 - x_4 \mid = 0.6$. Observation 6 has the largest x-value, so the two nearest neighbors are observations 4 and 5. In some situations one might impose the restriction that a donor is used only once when that is possible. We use the strict nearest-neighbor rule and use observation 4 as a donor for both observations 3 and 6. Each imputed value is weighted by one-half of the weight of the original observation. For observation 3 there are two new lines in the imputed data set, each with a weight of 1/12. The x-value is the same for both lines. If we had additional variables in the data set, those data are also repeated for the two lines.

Table 5.7 Jackknife Data

				Naive Replicate Weights					
Obs.	Donor	Weight	y_{Ii}	1	2	3	4	5	6
1	–	$0.16\bar{6}$	0.7	0	0.2	0.2	0.2	0.2	0.2
2	–	$0.16\bar{6}$	1.0	0.2	0	0.2	0.2	0.2	0.2
3	2	$0.08\bar{3}$	1.0	0.1	0.1	0	0.1	0.1	0.1
	4	$0.08\bar{3}$	2.2	0.1	0.1	0	0.1	0.1	0.1
4	–	$0.16\bar{6}$	2.2	0.2	0.2	0.2	0	0.2	0.2
5	–	$0.16\bar{6}$	2.3	0.2	0.2	0.2	0.2	0	0.2
6	4	$0.08\bar{3}$	2.2	0.1	0.1	0.1	0.1	0.1	0
	5	$0.08\bar{3}$	2.3	0.1	0.1	0.1	0.1	0.1	0

We construct jackknife replicate weights for variance estimation. The weights for six naive jackknife replicates are given in Table 5.8, where the weights are constructed as if the sample were complete. The two imputed values for an observation are treated as two observations from a primary

Table 5.8 Naive Weights for Respondents

Obs.	α_i	Naive Respondent Replication Weights					
		$\alpha_{1i}^{(1)}$	$\alpha_{1i}^{(2)}$	$\alpha_{1i}^{(3)}$	$\alpha_{1i}^{(4)}$	$\alpha_{1i}^{(5)}$	$\alpha_{1i}^{(6)}$
1	$0.16\bar{6}$	0.0	0.2	0.2	0.2	0.2	0.2
2	0.250	0.3	0.1	0.2	0.3	0.3	0.3
4	$0.33\bar{3}$	0.4	0.4	0.3	0.2	0.4	0.3
5	0.250	0.3	0.3	0.3	0.3	0.1	0.2

sampling unit. Ignoring the finite population correction, $c_k = 5/6$. The full-sample replicate weights for the respondents are given in Table 5.8. We have

$$(5/6)\sum_{k=1}^{L}(\alpha_{1i}^{(k)} - \alpha_i)^2 \quad = \quad (0.0278,\ 0.0292,\ 0.0278,\ 0.0292)$$

for $i = 1,\ 2,\ 4,\ 5$, respectively. From Table 5.8,

$$(\alpha_1^2, \alpha_2^2, \alpha_3^2, \alpha_4^2) = (0.0278, 0.0625, 0.1111, 0.0625).$$

The use of the naive replicates severely underestimates most of the coefficients for σ_i^2. Only for observation 1, the observation not used as a donor, is the sum of squares from the naive replicates equal to α_i^2.

Using the Φ_2 defined in (5.2.16) and the $w_{tj}^{*(k)}$ of (5.2.17), the quadratic equation for b_2 is

$$[0 + b_2(0.2)(0.5) - 0.25]^2 + [0.2 + 0.2(0.5) + (1 - 0.5b_2)(0.2) - 0.3333]^2 - 0.0350 - 0.0269 \quad = \quad 0.0750,$$

where $0.0750 - c_2^{-1}\alpha_2^2$, $c_k = 5/6$, and $\Phi_2 = 0.0350$. The simplified quadratic equation is

$$0.02b_2^2 - 0.0833b_2 + 0.0234 = 0$$

and $b_2 = 0.3033$. The equation for b_5 is the same as that for b_2. The quadratic equation for b_4 is

$$[0 + 2(0.1)b_4 - 0.3333]^2 + 2[0.2 + (1 - 0.5b_4)(0.2) - 0.25]^2 + 0.0333 - 0.0228 \quad = \quad 0.1333$$

and $b_4 = 0.1827$. The final jackknife replicates are given in Table 5.9 and the respondent weights in Table 5.10.

Table 5.9 Jackknife Weights for Fractional Imputation

			Weights for Unbiased Variance Estimator					
Obs.	Donor	Weight	1	2	3	4	5	6
1	–	$0.16\bar{6}$	0	0.2	0.2	0.2	0.2	0.2
2	–	$0.16\bar{6}$	0.2	0	0.2	0.2	0.2	0.2
3	2	$0.08\bar{3}$	0.1	0.030	0	0.183	0.1	0.1
	4	$0.08\bar{3}$	0.1	0.170	0	0.017	0.1	0.1
4	–	$0.16\bar{6}$	0.2	0.2	0.2	0	0.2	0.2
5	–	$0.16\bar{6}$	0.2	0.2	0.2	0.2	0	0.2
6	4	$0.08\bar{3}$	0.1	0.1	0.1	0.017	0.170	0
	5	$0.08\bar{3}$	0.1	0.1	0.1	0.183	0.030	0

Table 5.10 Final Weights for Respondents

	Final Respondent Replication Weight						
Obs.	α_i	$\alpha_i^{(1)}$	$\alpha_i^{(2)}$	$\alpha_i^{(3)}$	$\alpha_i^{(4)}$	$\alpha_i^{(5)}$	$\alpha_i^{(6)}$
1	$0.16\bar{6}$	0.0	0.2000	0.2	0.2000	0.2000	0.2
2	0.250	0.3	0.0303	0.2	0.3817	0.3000	0.3
4	$0.33\bar{3}$	0.4	0.4697	0.4	0.0366	0.4697	0.3
5	0.250	0.3	0.3000	0.3	0.3817	0.0303	0.2

The reader may check that

$$(5/6)\sum_{i\in A_R}\sum_{k=1}^{6}(\alpha_i^{(k)}-\alpha_i)^2 \;=\; \sum_{i\in A_R}\alpha_i^2,$$

where $A_R = (1,\ 2,\ 4,\ 5)$. Only for $i = 1$ is $\sum_k(\alpha_1^{(k)}-\alpha_1)^2 = \alpha_1^2$. For other observations the individual sums deviate slightly. Under our assumptions the neighborhoods that share a common donor have the same variance and hence the variance estimator is unbiased. Of course, the unbiased result requires the model assumptions of (5.2.6). ■■

5.2.4 Imputed estimators for domains

One of the reasons imputation is used in place of weighting is to improve estimates for domains. If the imputation model includes items such as age

and gender but not local geography, it is reasonable to believe that imputation will give an estimator for a small geographic area that is superior to the mean of the respondents in that area. If the model used for imputation is true, the imputed estimator for the small area may be superior to the simple estimator constructed from the full sample.

To illustrate the last point, consider a simple random sample and assume that the imputation model is

$$y_i = \mu + e_i, \tag{5.2.20}$$

where the e_i are $iid(0, \sigma^2)$ random variables. Let there be m nonrespondents and let the imputed value for each nonrespondent be the mean of the respondents. Let z_i be an indicator variable for membership in a domain, where a domain might be a cell in a two-way table. Assume that z_i is observed for all elements of the sample, and let the imputed estimator for domain a be

$$\hat{\mu}_a = \left(\sum_{i \in A} z_i \right)^{-1} \sum_{i \in A} z_i y_{iI}, \tag{5.2.21}$$

where y_{iI} is the imputed value for the rth element and $y_{iI} = y_i$ for respondents. Let domain a contain r_a respondents and m_a nonrespondents. Then the estimated domain mean based on imputed data is

$$\hat{\mu}_a = (m_a + r_a)^{-1} \left(\sum_{i \in A_{R,a}} y_i + m_a \bar{y}_r \right), \tag{5.2.22}$$

where $A_{R,a}$ is the set of indices of respondents in the domain and $\bar{y}_r$ is the mean of all respondents.

The model (5.2.20) is assumed to hold for all observations and hence holds for observations in the domain. Under model (5.2.20), mean imputation, and a negligible finite population correction,

$$\begin{aligned} V\{\hat{\mu}_a\} &= (m_a + r_a)^{-2}(r_a + 2m_a r_a r^{-1} + m_a^2 r^{-1})\sigma^2 \\ &= [(m_a + r_a)^{-1} + (m_a + r_a)^{-2} r^{-1} m_a (m_a + 2r_a - r)]\sigma^2 \\ &= [r^{-1} + (m_a + r_a)^{-2}(r_a - r_a^2 r^{-1})]\sigma^2. \end{aligned} \tag{5.2.23}$$

The second set of the expressions in (5.2.23) demonstrates that the imputed domain estimator is superior to the full-sample estimator if $r > m_a + 2r_a$, a condition easy to satisfy if the domain is small.

Under the model, the best estimator for the domain is $\bar{y}_r$. The last expression in (5.2.23) contains the increase in variance for the imputed domain estimator relative to the grand mean of the respondents. Often, practitioners are willing to use the model for imputation but unwilling to rely on the model to the degree required to use the model estimated mean for the cell. When the practitioner is willing to use the model estimator for the domain, the procedure is more often called *small area estimation.* See Section 5.5.

The use of donors from outside the domain produces a bias in the fractional replicated variance estimator for the domain. For nearest-neighbor imputation and an estimator linear in y, we constructed replicate weights that met the unbiasedness requirement (5.2.16) or an equivalent requirement. Because the weights for a domain estimator are not the same as the weights for the overall total, (5.2.16) will, in general, not hold for the domain mean.

Example 5.2.3. We use the imputed data of Table 5.7 to illustrate the nature of domain estimation. Assume that observations 1, 2, and 3 are in a domain. Then the imputed estimator for the domain total of y is

$$\hat{T}_d = N \sum_{j \in A} \sum_{i \in A_{Ij}} w_{ij}^* y_{[i]j} \delta_{dj}$$

and the imputed estimator for the domain mean is

$$\bar{y}_d = \left(\sum_{j \in A} \sum_{i \in A_{Ij}} w_{ij}^* \delta_{dj} \right)^{-1} \sum_{j \in A} \sum_{i \in A_{Ij}} w_{ij}^* y_{[i]j} \delta_{dj},$$

where $y_{[i]j}$ is the imputed value from donor i to recipient j, A_{Ij} is the set of indices of donors to j, w_{ij}^* are the weights of Table 5.7, and

$$\begin{aligned} \delta_{dj} &= 1 \quad \text{if observation } j \text{ is in domain } d \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Table 5.11 Respondent Weights for Alternative Estimators

	Observation			
Estimator	1	2	4	5
Total ($N = 1000$)	167	250	333	250
Imputed domain	167	250	83	
Poststrat. domain	250	250		

The line "imputed domain" in Table 5.11 gives the weights for the three respondents that contribute to the estimate for the domain. The weights are for a total under the assumption that $N = 1000$. Although observation 4 is not in the domain, it contributes to the domain estimate because, under the model, observation 4 has the same expectation as observation 2, which is in the domain. If only the two observations that fall in the domain are used to estimate the domain total, they receive the weights given in the last line of Table 5.11. Clearly, the weights of the second line of Table 5.11 will give a smaller sum of squares than those of the third line. ■■

5.3 VARIANCE ESTIMATION

Systematic sampling and one-per-stratum sampling produce unbiased estimators of totals, but design-unbiased variance estimation is not possible because some joint probabilities of selection are zero. One approach to variance estimation in these cases is to postulate a mean model and use deviations from the fitted model to construct a variance estimator.

Models can be divided into two types: local models and global models. For a population arranged in natural order, on a single variable x, a local model for y given x in the interval $q_j = (x_{Lj}, x_{Uj})$, is

$$\begin{aligned} y_i &= g\left(x_i, \boldsymbol{\beta}_j\right) + e_i, \; x_i \in q_j, \\ e_i &\sim ind(0, \sigma_i^2). \end{aligned} \tag{5.3.1}$$

The model is often simplified by letting x be the order number of the sample observations. A global model assumes that $g\left(x_i, \boldsymbol{\beta}\right)$ holds for the entire data set.

The most common local model assumption for a one-per-stratum design is that the means of two strata are the same. Thus, for an ordered set of an even number of strata,

$$g\left(x_i, \mu_h\right) = \mu_h, \qquad i = 1, \; 2, \; \ldots, \; 0.5n, \tag{5.3.2}$$

for $x_i = 2h$ and $x_i = 2h - 1$, where x_i is the order identification of the strata. Use of the mean model (5.3.2) leads to the procedure of collapsed strata discussed in Section 3.1.3.

Of the many variance estimation procedures that have been suggested for systematic sampling, the most popular is to form pairs of sample elements and assume a common mean for the pair. The pair is then treated as a set of two observations from a stratum. The created strata are sometimes called *pseudostrata*. Unlike the result for collapsed strata with one-per-stratum

sampling, the estimated variance for a systematic sample based on model (5.3.2) is not guaranteed to be an overestimate.

For a systematic sample or a one-per-stratum sample from a population arranged in natural order, a local model can be used to increase the number of degrees of freedom for the variance estimator relative to that for the collapsed strata procedure. Let $y_{[i]}$ denote the ith observation, where the order is that used in the sample selection, and let $x_i = i$. Then a local model that specifies adjacent observations to have the same mean is

$$\begin{aligned} y_{[i]} &= \mu_j + e_i, \\ e_i &\sim ind(0, \sigma_i^2), \end{aligned} \tag{5.3.3}$$

for $i \in [j, j+1]$. The associated variance estimator is

$$\begin{aligned} \hat{V}\{\bar{y}_\pi\} &= 0.5w_1^2(1-\pi_1)(y_{[2]} - y_{[1]})^2 \\ &\quad + 0.25\sum_{i=2}^{n-1} w_i^2(1-\pi_i)\left[(y_{[i-1]} - y_{[i]})^2 + (y_{[i]} - y_{[i+1]})^2\right] \\ &\quad + 0.5w_n^2(1-\pi_n)(y_{[n]} - y_{[n-1]})^2, \end{aligned} \tag{5.3.4}$$

where $w_i = N^{-1}\pi_i^{-1}$. If $w_i = n^{-1}$, the estimator reduces to

$$\begin{aligned} V\{\bar{y}\} &= N^{-1}(N-n)n^{-2}\{0.25\left[(y_{[2]} - y_{[1]})^2 + (y_{[n]} - y_{[n-1]})^2\right] \\ &\qquad + 0.5\sum_{i=2}^{n}(y_{[i]} - y_{[i-1]})^2\} \\ &\doteq N^{-1}(N-n)0.5n^{-1}(n-1)^{-1}\sum_{i=2}^{n}(y_{[i]} - y_{[i-1]})^2. \end{aligned} \tag{5.3.5}$$

The estimator (5.3.4) has nearly twice as many degrees of freedom as the collapsed strata procedure.

A second local model assumes a linear model for the center observation in a set of three adjacent observations. The model is

$$\begin{aligned} y_{[i]} &= \beta_{0j} + x_i\beta_{1j} + e_i \\ e_i &\sim ind(0, \sigma_i^2) \end{aligned} \tag{5.3.6}$$

for $x_i \in \{j-1,\ j,\ j+1\}$, where, as before, $x_i = i$ is the order identification. The use of local models for intervals greater than 2 usually requires an adjustment for the end observations. With model (5.3.6), for variance estimation purposes, we assume that the superpopulation mean associated with the first observation is equal to the superpopulation mean associated with the second

Table 5.12 Weights for Replicate Variance Estimation

Observation	Full Sample	Replicate 1	2	···	9	10
1	0.1000	0.1671	0.1387		0.1000	0.1000
2	0.1000	0.0329	0.0226		0.1000	0.1000
3	0.1000	0.1000	0.1387		0.1000	0.1000
4	0.1000	0.1000	0.1000		0.1000	0.1000
5	0.1000	0.1000	0.1000		0.1000	0.1000
6	0.1000	0.1000	0.1000		0.1000	0.1000
7	0.1000	0.1000	0.1000		0.1000	0.1000
8	0.1000	0.1000	0.1000		0.1387	0.1000
9	0.1000	0.1000	0.1000		0.0226	0.1671
10	0.1000	0.1000	0.1000		0.1387	0.0329

observation. Making the same assumption for the last two observations, the estimated variance of an estimator of the form $\Sigma_{i\in A} w_i y_i$ is

$$\begin{aligned}\hat{V}\{\bar{y}\} &= N^{-1}(N-n)[0.5w_1^2(y_{[1]}-y_{[2]})^2 \\ &\quad + \sum_{i=2}^{n-1} w_i^2 6^{-1}(y_{[i-1]}-2y_{[i]}+y_{[i+1]})^2 \\ &\quad + 0.5w_n^2(y_{[n-1]}-y_{[n]})^2]. \end{aligned} \tag{5.3.7}$$

The linear combination $(y_{[i-1]} - 2y_{[i]} + y_{[i+1]})$ is a multiple of the deviation from fit for the linear model estimated with the three observations $(y_{[i-1]}, y_{[i]}, y_{[i+1]})$.

The estimator (5.3.7) has a positive bias for the design variance of the one-per-stratum design. The bias expression can be obtained by replacing y_j with μ_j in (5.3.7), where μ_j is the mean for the jth stratum.

Example 5.3.1. Replication can be used to construct estimator (5.3.7). To illustrate, assume that a sample of 10 observations is selected, either by systematic sampling or by equal-probability one-per-stratum sampling, from an ordered population of 100 elements.

The replicate weights in Table 5.12 are such that the kth deviation $\bar{y}^{(k)} - \bar{y}$ is the kth linear combination in (5.3.7) normalized so that the square has the correct expectation. Thus, the entries in the first column for replicate 1 give

$$\begin{aligned}\bar{y}^{(1)} - \bar{y} &= \left[0.5N^{-1}(N-n)\right]^{0.5} n^{-1}(y_{[1]}-y_{[2]}) \\ &= 0.0671(y_{[1]}-y_{[2]}).\end{aligned}$$

and the entries in the second column give

$$\begin{aligned}\bar{y}^{(2)} - \bar{y} &= \left[6^{-1}N^{-1}(N-n)\right]^{0.5} n^{-1}(y_{[1]} - 2y_{[2]} + y_{[3]}) \\ &= 0.0387y_{[1]} - 0.0774y_{[2]} + 0.0387y_{[3]}.\end{aligned}$$

■■

Models for the covariance structure of the population can be used to estimate the variance of one-per-stratum designs. Let the population size for stratum h be N_h, and let U_h be the set of indices for stratum h. Let a simple random sample of size 1 be selected in each stratum. Then the variance of the estimated total for stratum h is

$$V\{\bar{y}_{st}\} = \sum_{h=1}^{H}(1 - N_h^{-1})W_h^2 S_h^2, \tag{5.3.8}$$

where $W_h = N^{-1}N_h$. Assume that the finite population is a realization of a stationary stochastic process, where the covariance is a function of the distance between observations. That is,

$$C\{y_j, y_k\} = \gamma(d), \tag{5.3.9}$$

where d is the distance between j and k and $\gamma(d) = \gamma(-d)$ is the covariance between two units that are a distance d apart. The distance can be defined in terms of auxiliary information and is often the distance between the indexes of a population arranged in natural order.

Under the model,

$$E\{S_h^2\} = (N_h - 1)^{-1}\left(N_h\gamma(0) - N_h^{-1}\sum_{d=-(N_h-1)}^{N_h-1}(N_h - d)\gamma(d)\right)$$

and

$$E\{V(y_j - \bar{y}_{h,N} \mid j \in U_h, \mathcal{F})\} = \gamma(0) - N_h^{-2}\sum_{d=-(N_h-1)}^{N_h-1}(N_h - d)\gamma(d). \tag{5.3.10}$$

If a parametric model for $\gamma(d)$ can be estimated, the estimated values for $\gamma(d)$ can be substituted in (5.3.10) and (5.3.8) to obtain an estimated variance for $\bar{y}_{st}$. It may be simpler to estimate the parameters of

$$E\{(y_j - y_k)^2\} = \psi(d, \boldsymbol{\theta}), \tag{5.3.11}$$

where $\psi(d, \boldsymbol{\theta})$ is called the *variogram* and $\boldsymbol{\theta}$ is the parameter vector.

For the one-per-stratum procedure, only differences with $d \leq N_M - 1$ enter the variance expression, where N_M is the maximum of the N_h. Thus, a simple procedure is to assume a constant variogram for $d \leq N_M - 1$ and estimate the variance with

$$\hat{V}\{\bar{y}_{st}\} = 0.5 \sum_{h=1}^{H} (1 - N_h^{-1}) W_h^2 n_\delta^{-1} \sum_{j \in A} \sum_{k \in A} (y_j - y_k)^2 \delta_{j,k}, \quad (5.3.12)$$

where

$$\begin{aligned} \delta_{j,k} &= 1 \quad \text{if } 0 < \mid j - k \mid \leq N_M - 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

and

$$n_\delta = \sum_{j \in A} \sum_{k \in A} \delta_{j,k}.$$

For a variogram that increases with distance, the estimator (5.3.12) will have a positive bias because differences with large d appear more frequently in the estimator than in the population.

A three-parameter variogram model based on the first-order autoregressive process is

$$\psi(d, \boldsymbol{\theta}) = \theta_0 + \theta_1 \theta_2^{|j-k|}, \quad (5.3.13)$$

where $\theta_0 \geq \theta_1$ and $\mid \theta_2 \mid < 1$. For the stationary first-order autoregressive process, $\theta_0 = \theta_1$. For a process with measurement error, $\theta_0 > \theta_1$. The parameters can be estimated using a nonlinear least squares procedure or by maximum likelihood. A large number of observations is required to obtain good estimates for the parameters of variance models. See Cressie (1991, Chapter 2) for a discussion of the variogram and for variogram models.

5.4 OUTLIERS AND SKEWED POPULATIONS

The problem of estimating the mean, or total, using a sample containing a few "very large" observations will be faced by almost every sampling practitioner. The definition of "very large" must itself be part of a study of estimation for such samples and the definition of "very large" that appears most useful is a definition that separates cases wherein the sample mean performs well as an estimator from those cases wherein alternative estimators are markedly

superior to the mean. Most editing procedures will have rules that identify unusual observations as part of the checking for errors. In this section we are interested in situations where the extreme observation, or observations, have been checked and the data are believed to be correct. These are typically situations where the population sampled is very skewed. Personal income and size measures of businesses are classical examples.

One approach to estimation for skewed populations is to specify a parametric model for the superpopulation and estimate the parameters of that model. Parametric estimation is discussed in Chapter 6. Our experience suggests that it is very difficult to specify a relatively simple model for the entire distribution. Robust procedures, as described by Huber (1981) and Hempel et al. (1986), are related estimation procedures but have heavy emphasis on symmetric distributions.

In applications, an observation can be extreme because the value of the characteristic is large, because the weight is large, or both. Estimators that make adjustments in the largest observations can be made by modifying the value or by modifying the weight. Because the value is believed to be correct, the modification is most often made by modifying the weight. We begin by considering the general estimation problem for simple random samples.

We present the procedure of Fuller (1991), in which it is assumed that the right tail of the distribution can be approximated by the right tail of a Weibull distribution. The Weibull density is

$$\begin{aligned} f(y;\alpha,\lambda) &= \alpha\lambda^{-1}y^{\alpha-1}exp\{-\lambda^{-1}y^{\alpha}\} \qquad \text{if } y > 0 \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{otherwise,} \end{aligned} \tag{5.4.1}$$

where $\lambda > 0$ and $\alpha > 0$. If x is defined by the one-to-one transformation $x = y^{\alpha}$, x is distributed as an exponential random variable with parameter λ. Conversely, the Weibull variable is the power of an exponential variable, x^{γ}, where $\gamma = \alpha^{-1}$. If $\alpha \leq 1$, the sample mean will perform well as an estimator of the population mean. If α is much larger than 1, there are alternative estimators that will perform better than the sample mean. We use the order statistics to test the hypothesis that $\alpha = 1$ against the alternative that $\alpha > 1$.

The distribution of the differences of order statistics from the exponential distribution are distributed as exponential random variables. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the order statistics of a sample of size n selected from an exponential distribution with parameter λ, and let $x_{(0)} = 0$. Then the random variables

$$z_k = (n-k+1)(x_{(k)} - x_{(k-1)}), \qquad k = 1, 2, \ldots, n, \tag{5.4.2}$$

are *iid* exponential random variables with parameter λ. See David (1981, p. 20). Postulating the exponential model for the largest m observations, we

construct the test

$$F_{mj} = \left((m-j)^{-1} \sum_{i=n-m+1}^{n-j} Z_i \right)^{-1} j^{-1} \sum_{i=n-j+1}^{n} z_i. \qquad (5.4.3)$$

If $\alpha = 1$, F_{mj} is distributed as Snedecor's F with $2j$ and $2(m-j)$ degrees of freedom. If the test rejects $\alpha = 1$, one has reason to believe that there are estimators superior to the mean. A relatively simple estimator constructed with the order statistics is

$$\begin{aligned} \hat{\mu}_{mj} &= \bar{y} && \text{if } F_m < K_j \\ &= n^{-1} \left(\sum_{i=1}^{n-j} y_{(i)} + j(y_{n-j} + K_j \bar{d}_{mj}) \right) && \text{otherwise,} \end{aligned} \qquad (5.4.4)$$

where F_m is defined in (5.4.3), K_j is a cutoff value, and

$$\bar{d}_{mj} = (m-j)^{-1} \left(\sum_{i=j}^{m-1} (y_{(n-i)} - y_{(n-m)}) + j(y_{(n-j)} - y_{(n-m)}) \right).$$

The estimator of (5.4.4) is a test-and-estimate procedure in which the estimator is a continuous function of the sums formed from different sets of order statistics. The sample mean is a special case of estimator (5.4.4) obtained by setting K_j equal to infinity.

It is difficult to specify the number of tail observations, m, the number of large order statistics, j, and the cutoff values, K_j, to use in constructing the estimator for the tail portion. It would seem that m approximately equal to one-fifth to one-third of the observations is reasonable for many populations and sample sizes. It also seems that one can reduce this fraction in large $(n > 200)$ samples. When the sample is large, setting $m \doteq 30$ seems to perform well.

In many applications $j = 1$, and K_j equal to the 99.5 percentile of the F distribution works well. The large value required for rejection means that the procedure has good efficiency for populations with modest skewness. See Fuller (1991a). The results of Rivest (1994) also support the use of $j = 1$.

5.5 SMALL AREA ESTIMATION

Sometimes estimates for a set of small domains are of considerable interest, but the sample sizes in the individual domains are not large enough to provide direct estimates with acceptable standard errors. In such situations models

and auxiliary variables can be used to construct improved estimates for the domains. The domains are often geographic areas, and the term *small area estimation* is used as a generic expression for such procedures in the survey literature. Rao (2003) describes a large number of procedures and models. We present one frequently used model.

Let the population be divided into M mutually exclusive and exhaustive areas. Let survey estimates be available for m, $m \leq M$, of the areas. Let $\bar{y}_g$ be the estimate of the mean for the gth area, and let $\bar{\mathbf{x}}_{gN}$ be a vector of known means for a vector of auxiliary variables for the gth area. For example, the areas might be metropolitan areas, and the means might be means per household. Assume that the $\bar{y}_g$ satisfy the model

$$\bar{y}_g = \theta_g + \bar{e}_g \tag{5.5.1}$$

and

$$\theta_g = \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} + u_g,$$

where u_g is the area effect, $\bar{e}_g$ is the sampling error,

$$u_g \sim ii(0, \sigma_u^2),$$

$$\bar{e}_g \sim ind(0, \sigma_{eg}^2),$$

and u_g is independent of $\bar{e}_h$ for all h and g. This model is also called a *mixed model* because the mean of y for area g is assumed to be the sum of a fixed part $\bar{\mathbf{x}}_{gN}\boldsymbol{\beta}$ and a random part u_g. The unknown mean for area g is $\theta_g = \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} + u_g$.

To be comfortable with model (5.5.1) the analyst should feel that the observable important differences among areas are included in the vector $\bar{\mathbf{x}}_{gN}$. That is, after adjusting for $\bar{\mathbf{x}}_{gN}$, there is no reason to believe that any area is particularly unusual relative to the others. We state our model for means, but the nature of the data will vary for different problems. The model could be stated in terms of mean per primary sampling unit or mean per element. The model could also be defined in terms of small area totals, but we find the model (5.5.1) more appealing when expressed in terms of means.

To introduce the estimation procedure, assume $\boldsymbol{\beta}$, σ_u^2, σ_{eg}^2 are known. Then $\bar{\mathbf{x}}_{gN}\boldsymbol{\beta}$ and $u_g + \bar{e}_g = \bar{y}_g - \bar{\mathbf{x}}_{gN}\boldsymbol{\beta}$ are known for the m sampled areas. If $(u_g, \bar{e}_g)$ is normally distributed, then $(u_g + \bar{e}_g, u_g)$ is normally distributed and the best predictor of u_g, given $u_g + \bar{e}_g$, is

$$\hat{u}_g = \gamma_g(u_g + \bar{e}_g), \tag{5.5.2}$$

where

$$\gamma_g = (\sigma_u^2 + \sigma_{eg}^2)^{-1}\sigma_u^2$$

is the population regression coefficient for the regression of u_g on $(u_g + \bar{e}_g)$. If $(u_g, \bar{e}_g)$ is not normal, (5.5.2) is the best linear unbiased predictor of u_g. Therefore, a predictor of the mean of y for the gth area is

$$\begin{aligned} \tilde{\theta}_g &= \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} + \gamma_g(\bar{y}_g - \bar{\mathbf{x}}_{gN}\boldsymbol{\beta}) \qquad \text{if } g \in A \\ &= \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} \qquad \text{if } g \notin A, \end{aligned} \tag{5.5.3}$$

where A is the index set for small areas in which $\bar{y}_g$ is observed.

The terms *small area estimator* and *small area predictor* are both used in the literature. We prefer to describe (5.5.3) as a predictor because u_g is a random variable. The variance of the prediction error is

$$\begin{aligned} V\{\tilde{\theta}_g - \theta_g\} &= (\sigma_u^2 + \sigma_{eg}^2)^{-1}\sigma_u^2\sigma_{eg}^2 \qquad \text{if } g \in A \\ &= \sigma_u^2 \qquad \text{if } g \notin A. \end{aligned} \tag{5.5.4}$$

See Exercise 8.

If $\boldsymbol{\beta}$ is unknown but σ_u^2 and σ_{eg}^2 are known, the generalized least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{m} \bar{\mathbf{x}}_{gN}' \left(\sigma_u^2 + \sigma_{eg}^2\right)^{-1} \bar{\mathbf{x}}_{gN} \right)^{-1} \sum_{g=1}^{m} \bar{\mathbf{x}}_{gN}' \left(\sigma_u^2 + \sigma_{eg}^2\right)^{-1} \bar{y}_g. \tag{5.5.5}$$

The estimator (5.5.5) of $\boldsymbol{\beta}$ can be substituted for $\boldsymbol{\beta}$ in (5.5.3) to obtain the unbiased predictor,

$$\begin{aligned} \hat{\theta}_g &= \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}} + \gamma_g(\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}) \qquad \text{if } g \in A \\ &= \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}} \qquad \text{if } g \notin A. \end{aligned} \tag{5.5.6}$$

The prediction variance has an added term due to the estimation of $\boldsymbol{\beta}$,

$$\begin{aligned} V\{\hat{\theta}_g - \theta_g\} &= \gamma_g\sigma_{eg}^2 + (1-\gamma_g)^2\bar{\mathbf{x}}_{gN}V\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}_{gN}' \quad \text{if } g \in A \\ &= \sigma_u^2 + \bar{\mathbf{x}}_{gN}V\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}_{gN}' \qquad \text{if } g \notin A, \end{aligned} \tag{5.5.7}$$

where

$$V\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{g=1}^{m} \bar{\mathbf{x}}_{gN}' \left(\sigma_u^2 + \sigma_{eg}^2\right)^{-1} \bar{\mathbf{x}}_{gN} \right)^{-1}. \tag{5.5.8}$$

See Exercise 9.

Estimation becomes even more difficult for the realistic situation in which σ_u^2 is unknown. Although an estimator of σ_{eg}^2 is often available, estimation of

σ_u^2 and β requires nonlinear estimation procedures. A number of statistical packages contain estimation algorithms for both Bayesian and classical procedures. For classical estimation, one procedure uses estimators of β and σ_{eg}^2 to construct an estimator of σ_u^2 and then uses the estimator of σ_u^2 and estimators of σ_{eg}^2 to construct an improved estimator of β. The predictor is (5.5.3) with the estimators of σ_{eg}^2, σ_u^2, and β replacing the unknown parameters.

An estimator of the prediction mean square error (MSE) is

$$\begin{aligned}\hat{V}\{\hat{\theta}_g - \theta_g\} &= \hat{\gamma}_g\hat{\sigma}_{eg}^2 + (1-\hat{\gamma}_g)^2\bar{\mathbf{x}}_{gN}\hat{V}\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}_{gN}' \\ &\quad + 2(\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)\hat{V}\{\hat{\gamma}_g\} \qquad \text{if } g \in A \\ &= \hat{\sigma}_u^2 + \bar{\mathbf{x}}_{gN}\hat{V}\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}_{gN}' \qquad \text{if } g \notin A, \quad (5.5.9)\end{aligned}$$

where

$$\hat{V}\{\hat{\gamma}_g\} = (\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)^{-4}[\,\hat{\sigma}_u^4\hat{V}\{\hat{\sigma}_{eg}^2\} + \hat{\sigma}_{eg}^4\hat{V}\{\hat{\sigma}_u^2\}],$$

$$\hat{V}\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{g=1}^{m}\bar{\mathbf{x}}_{gN}'(\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)^{-1}\bar{\mathbf{x}}_{gN}\right)^{-1},$$

$\hat{\sigma}_{eg}^2$ is an estimator of σ_{eg}^2 based on d_g degrees of freedom, and $d_g + 1$ is typically the number of primary sampling units in the small area. Many computer programs will provide an estimate of the variance of σ_u^2, where the estimated variance of $\hat{\sigma}_u^2$ will depend on the particular algorithm used.

One estimator of σ_u^2 is

$$\hat{\sigma}_u^2 = \sum_{g=1}^{m}\kappa_g[(m-k)^{-1}m(\bar{y}_g - \bar{\mathbf{x}}_g\hat{\boldsymbol{\beta}})^2 - \hat{\sigma}_{eg}^2],$$

with the estimated variance

$$\hat{V}\{\hat{\sigma}_u^2\} = \sum_{g=1}^{m}\kappa_g^2[\,2(\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)^2 + \hat{V}\{\hat{\sigma}_{eg}^2\}],$$

where k is dimension of $\mathbf{x}_g$ and

$$\kappa_g = \left(\sum_{j=1}^{m}\left[\hat{\sigma}_u^2 + d_j^{-1}(d_j+2)\hat{\sigma}_{ej}^2\right]^{-1}\right)^{-1}\left[\,\hat{\sigma}_u^2 + d_g^{-1}(d_g+2)\hat{\sigma}_{eg}^2\right]^{-1}.$$

Construction of the estimator requires iteration because κ_g depends on $\hat{\sigma}_u^2$. See Prasad and Rao (1990), Rao (2003), and Wang and Fuller (2003) for derivations and alternative estimators.

Example 5.5.1. We illustrate small area estimation with some data from the U.S. National Resources Inventory (NRI). The NRI was described in Example 1.2.2.

In this example we use data on wind erosion in Iowa for the year 2002. The analysis is based on that of Mukhopadhyay (2006). The data had not been released at the time of this study, so the data in Table 5.13 are a modified version of the original data. The general nature of the original estimates is preserved, but published estimates will not agree with those appearing in the table. The N_g is the population number of segments in the county and n_g is the sample number of segments. The variable y is the cube root of wind erosion. This variable is not of subject matter interest in itself but is used for illustrative purposes. There are 44 counties in Iowa for which wind erosion is reported. There were observations in all 44 counties in the study, but for purposes of this illustration we assume that there are four additional counties with no sample observations.

Wind erosion is a function of soil characteristics. The soils of Iowa have been mapped, so population values for a number of soil characteristics are available. The mean of the soil erodibility index for the county is used as the explanatory variable in our model. For our purposes, the sample of segments in a county is treated as a simple random sample. A preliminary analysis suggested that the assumption of a common population variance for the counties was reasonable. Therefore, we assume that the variance of the mean wind erosion for county g is $n_g^{-1}\sigma_e^2$, where n_g is the number of segments in county g and σ_e^2 is the common variance. We treat $\sigma_e^2 = 0.0971$ as known in our analysis. Thus, our model is

$$\bar{y}_g = \theta_g + e_g, \tag{5.5.10}$$

$$\theta_g = \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} + u_g,$$

$$u_g \sim ii(0, \sigma_u^2),$$

$$e_g \sim ind(0, n_g^{-1}\sigma_e^2),$$

where u_g is independent of e_j for all g and j, $\bar{\mathbf{x}}_{gN} = [1, 0.1(\bar{r}_{1,g,N} - 59)]$, $\bar{r}_{1,g,N}$ is the population mean erodibility index for county g and $\bar{y}_g$ is the estimated mean wind erosion for county g. The erodibility index was reduced by 59 in the regression to facilitate the numerical discussion.

Using a program such as PROC Mixed of SAS, the estimated model parameters are

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_u^2) = \underset{(0.026)}{(0.770,} \ \underset{(0.024)}{0.155,} \ \underset{(0.0062)}{0.0226),}$$

where the estimates are based on the 44 counties with erosion values and the estimator of σ_u^2 is the maximum likelihood estimator.

The estimated parameters were used to construct the predictions of the erosion measure given in Table 5.13. For the first county in the table,

$$\begin{aligned} \hat{\theta}_3 &= \hat{\gamma}_3 \bar{y}_3 + (1 - \hat{\gamma}_3)\bar{\mathbf{x}}_{3N}\hat{\boldsymbol{\beta}}, \\ &= 0.466, \end{aligned}$$

where $\hat{\gamma}_3 = [0.0226 + (13)^{-1}0.0971]^{-1}(0.0226) = 0.7516$ and $\bar{\mathbf{x}}_{3N} = (1, -1.232)$. The standard errors of Table 5.13 were computed using (5.5.9) treating σ_e^2 as known. For county 3 the model standard error of 0.077 is about 89% of the design standard error of 0.086. County 167 has 44 observations, $\hat{\gamma}_{167} = 0.942$, and the model standard error is about 99% of the design standard error. The difference between the design standard errors and the prediction standard errors are modest because the σ_{eg}^2 are small relative to $\hat{\sigma}_u^2$.

Table 5.13: **Data on Iowa Wind Erosion**

County	N_g	n_g	Erodibility Index	$\bar{y}_g$	$\hat{\theta}_g$	S.E. of Prediction
3	1387	13	46.683	0.429	0.466	0.077
15	2462	18	58.569	0.665	0.684	0.067
21	2265	14	65.593	1.083	1.034	0.074
27	2479	19	47.727	0.788	0.753	0.066
33	2318	18	52.802	0.869	0.831	0.067
35	1748	12	72.130	1.125	1.085	0.079
41	2186	16	59.079	0.683	0.701	0.070
47	3048	19	49.757	0.408	0.448	0.066
59	1261	12	53.694	0.839	0.799	0.079
63	1822	15	63.563	0.754	0.773	0.072
67	1597	11	44.947	0.690	0.651	0.082
71	1345	15	56.807	0.927	0.885	0.072
73	1795	12	54.182	0.945	0.879	0.079
75	2369	13	40.951	0.619	0.587	0.077
77	2562	15	48.605	0.475	0.504	0.072
79	1899	11	74.981	0.790	0.854	0.082
83	2486	16	57.455	0.647	0.668	0.070
85	2241	19	66.700	0.727	0.757	0.066
91	2066	15	56.118	1.120	1.032	0.072

Continued

County	N_g	n_g	Erodibility Index	$\bar{y}_g$	$\hat{\theta}_g$	S.E. of Prediction
93	1385	10	61.830	0.677	0.718	0.084
109	2752	18	64.255	0.968	0.945	0.067
119	1753	29	61.605	0.703	0.717	0.055
129	1270	12	58.739	0.616	0.656	0.079
131	1232	10	48.739	0.422	0.478	0.084
133	2943	24	73.121	1.045	1.037	0.060
135	1190	15	45.417	0.363	0.407	0.072
141	1567	11	81.911	1.424	1.340	0.083
143	1511	10	56.229	0.975	0.900	0.084
145	1772	16	40.862	0.451	0.459	0.071
147	2716	17	60.811	0.945	0.915	0.069
149	3877	16	80.541	1.065	1.073	0.071
151	1823	10	63.190	0.918	0.893	0.084
153	1580	18	48.503	0.670	0.658	0.067
155	4405	21	62.348	0.619	0.653	0.063
157	2121	13	44.462	0.578	0.570	0.077
161	2423	16	66.551	0.719	0.754	0.070
165	2327	12	47.496	0.376	0.432	0.079
167	3180	44	72.262	0.954	0.956	0.045
169	1862	16	54.794	0.583	0.609	0.070
187	3011	15	58.420	0.874	0.849	0.072
189	1644	10	76.335	1.256	1.191	0.085
193	2319	17	75.142	0.905	0.928	0.069
195	1290	16	46.488	0.599	0.594	0.071
197	1754	11	71.380	0.577	0.685	0.082
201	1822		63.563		0.841	0.153
202	1511		56.229		0.727	0.153
203	3877		80.541		1.104	0.161
204	3011		58.420		0.761	0.153

The prediction for a county with no observations is $\bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$ and has a variance

$$V\{\hat{\theta}_g - \theta_g\} = V\{\bar{\mathbf{x}}_g\hat{\boldsymbol{\beta}}\} + \sigma_u^2,$$

estimated by

$$\hat{V}\{\hat{\theta}_g - \theta_g\} = \bar{\mathbf{x}}_{gN}\hat{V}\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}'_{gN} + \hat{\sigma}_u^2,$$

where the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\hat{V}\{\hat{\boldsymbol{\beta}}\} = \begin{pmatrix} 6.65 & 0.04 \\ 0.04 & 5.77 \end{pmatrix} \times 10^{-4}.$$

Thus, the estimated prediction variance for county 201 is

$$\begin{aligned}\hat{V}\{\hat{\theta}_{201} - \theta_{201}\} &= (1, 0.456)\hat{V}\{\hat{\boldsymbol{\beta}}\}(1, 0.456)' + \hat{\sigma}_u^2 \\ &= 7.8894(10^{-4}) + 0.0226 = 0.0234.\end{aligned}$$

The estimated variance of u_g is the dominant term in the estimated prediction variance when there are no observations in the county. ■■

In many situations the standard error of the direct survey estimate for the overall total is judged to be acceptable, whereas those for the small areas are judged to be too large. In such situations the practitioner may prefer small area estimates that sum to a design consistent survey estimate of the total. That is, it is requested that

$$\sum_{g=1}^{M} N_g \hat{\theta}_g = \hat{T}_y, \tag{5.5.11}$$

where N_g is the number of elements in small area g, $\hat{\theta}_g$ is the small area predictor, and $\hat{T}_y$ is a design-consistent estimator of the total of y. If (5.5.11) is satisfied, the predictions are said to be *benchmarked* and the small area procedure becomes a method for allocating the design-consistent estimated total to the small areas. Two situations for benchmarking can be considered. In one the design-consistent estimator has been constructed using information not used for the small area estimation. Procedures appropriate for this situation have been reviewed by Wang, Fuller, and Qu (2009).

We consider benchmarking for the situation in which information to be used to construct the design consistent estimator is that used in the small area estimation. Under model (5.5.1), $\bar{x}_{gN}$ is known for all small areas and it follows that the population total

$$\mathbf{T}_x = \sum_{g=1}^{M} N_g \bar{\mathbf{x}}_{gN} \tag{5.5.12}$$

is known. Given the information available on $\mathbf{x}$, it is natural to use the regression estimator as an estimator for the total of y. If σ_{eg}^2 and σ_u^2 are known, the generalized least squares estimator (5.5.5) is the preferred estimator for $\boldsymbol{\beta}$, and a regression estimator of the total of y is

$$\hat{T}_{y,reg} = \mathbf{T}_x \hat{\boldsymbol{\beta}}, \tag{5.5.13}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

$\mathbf{V} = \text{diag}(\sigma_u^2 + \sigma_{eg}^2)$, $\mathbf{X}$ is the $m \times k$ matrix with gth row equal to $\bar{\mathbf{x}}_{gN}$, $\mathbf{y}' = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_g, \ldots, \bar{y}_m)$, and the estimator of $\boldsymbol{\beta}$ of (5.5.13) is identically equal to the estimator $\hat{\boldsymbol{\beta}}$ of (5.5.5). We investigate the design consistency of estimator (5.5.13) permitting the number of small areas with a direct observation $\bar{y}_g$ to be less than M. Let π_g denote the probability that area g is observed and assume that $(\bar{y}_g, \bar{\mathbf{x}}_g)$ is design unbiased for $(\bar{y}_{gN}, \bar{\mathbf{x}}_{gN})$. Then a design-unbiased estimator of the vector of totals is

$$(\hat{T}_y, \hat{\mathbf{T}}_x) = \sum_{g \in A} N_g \pi_g^{-1} (\bar{y}_g, \bar{\mathbf{x}}_g),$$

It follows that the regression estimator (5.5.13) will be design consistent for the total of y if there is a vector $\mathbf{c}_1$ such that

$$\bar{\mathbf{x}}_{gN} \mathbf{c}_1 = \pi_g^{-1} N_g (\sigma_u^2 + \sigma_{eg}^2) \tag{5.5.14}$$

for all g. See Corollary 2.2.3.1.

For the weighted sum of the small area predictors to be equal to the regression estimator of the total, we require that

$$\begin{aligned} \hat{T}_{y,reg} &= \sum_{g=1}^{M} N_g \hat{\theta}_g \\ &= \sum_{g=1}^{M} N_g [\bar{\mathbf{x}}_{gN} \hat{\boldsymbol{\beta}} + \gamma_g (\bar{y}_g - \bar{\mathbf{x}}_{gN} \hat{\boldsymbol{\beta}})], \end{aligned} \tag{5.5.15}$$

where it is understood that the predictor is $\bar{\mathbf{x}}_{jN} \hat{\boldsymbol{\beta}}$ if area j is not observed. Because $\mathbf{T}_x = \sum_{g=1}^{M} N_g \bar{\mathbf{x}}_{gN}$, the requirement (5.5.15) becomes

$$\sum_{g=1}^{m} N_g \gamma_g (\bar{y}_g - \bar{\mathbf{x}}_{gN} \hat{\boldsymbol{\beta}}) = 0,$$

where m is the number of small areas observed, and the requirement (5.5.15) is satisfied for estimator (5.5.13) if there is a vector $\mathbf{c}_2$ such that

$$\bar{\mathbf{x}}_{gN} \mathbf{c}_2 = N_g \tag{5.5.16}$$

for all g. See Exercise 12. Thus, if N_g and $\pi_g^{-1} N_g (\sigma_u^2 + \sigma_{eg}^2)$ are in the column space of $\mathbf{X}$, the weighted sum of the small area predictors is equal to the design-consistent regression estimator of the total.

Typically, σ_{eg}^2 will not be known for the unobserved areas, but $\pi_g^{-1} N_g$ will be known. In some situations the rule defining n_g as a function of N_g

is known and one may be willing to assume that σ_{eg}^2 is of the form $n_g^{-1}\sigma_e^2$. Then $\pi_g^{-1}N_g\sigma_{eg}^2 = \pi_g^{-1}N_g n_g^{-1}\sigma_e^2$ can be treated as known for the unobserved areas.

To consider the case where σ_{eg}^2 is unknown, we adopt some of the notation of Section 2.3 and let $\mathbf{z}_g = (\bar{\mathbf{x}}_{0,gN}, \mathbf{z}_{dg})$, where

$$\mathbf{z}_d = \mathbf{x}_d - \mathbf{X}_0(\mathbf{X}_0'\mathbf{V}^{-1}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{V}^{-1}\mathbf{x}_d, \tag{5.5.17}$$

$\mathbf{x}_d$ is the vector of observations on $x_{d,g} = \pi_g^{-1}N_g(\sigma_u^2 + \sigma_{eg}^2)$, $\mathbf{X}_0$ is the matrix of observations on the other explanatory variables, and $\mathbf{V}$ is defined in (5.5.13). The population mean of $\bar{\mathbf{x}}_0$ is known, but the population mean of z_d is unknown. Following the development of Section 2.3, we let

$$\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_0', \hat{\alpha}_d)' = (\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y}, \tag{5.5.18}$$

where $\mathbf{Z} = (\mathbf{X}_0, \mathbf{z}_d)$. Then the regression estimator (2.3.22) is

$$\bar{y}_{reg} = (\bar{\mathbf{x}}_{0,\cdot,N}, \bar{z}_{d\pi})\hat{\boldsymbol{\alpha}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{0,\cdot,N} - \bar{\mathbf{x}}_{0\pi})\hat{\boldsymbol{\alpha}}_0, \tag{5.5.19}$$

where

$$(\bar{y}_\pi, \bar{\mathbf{x}}_{0\pi}, \bar{z}_{d\pi}) = \left(\sum_{g\in A} N_g\pi_g^{-1}\right)^{-1} \sum_{g\in A} N_g\pi_g^{-1}(\bar{y}_g, \bar{\mathbf{x}}_{0g}, z_{dg})$$

and $\bar{\mathbf{x}}_{0,\cdot,N}$ is the population mean of $\mathbf{x}_0$. The small area predictors are

$$\begin{aligned}\hat{\theta}_g &= \bar{\mathbf{z}}_{gN}\hat{\boldsymbol{\alpha}} + \hat{\gamma}_g(\bar{y}_g - \bar{\mathbf{z}}_{gN}\hat{\boldsymbol{\alpha}}) \\ &= \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}} + \hat{\gamma}_g(\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}) \qquad \text{if } g \in A\end{aligned}$$

and

$$\hat{\theta}_g = (\bar{\mathbf{x}}_{0,g,N}, \bar{z}_{d\pi})\hat{\boldsymbol{\alpha}} \qquad \text{if } g \notin A. \tag{5.5.20}$$

When $\bar{y}_g$ is observed, the estimated MSE of $\hat{\theta}_g$ is of the form (5.5.9) and, in the current notation, is

$$\begin{aligned}\hat{V}\{\hat{\theta}_g - \theta_g\} &= \hat{\gamma}_g\hat{\sigma}_{eg}^2 + (1-\hat{\gamma}_g)^2\bar{\mathbf{z}}_{gN}\hat{V}\{\hat{\boldsymbol{\alpha}}\}\bar{\mathbf{z}}_{gN}' \\ &\quad + 2(\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)\hat{V}\{\hat{\gamma}_g\}.\end{aligned}$$

If $\bar{y}_g$ is not observed, the estimation error is

$$\hat{\theta}_g - \theta_g = (\bar{\mathbf{x}}_{0,g,N}, \bar{z}_{d\pi})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - (\bar{z}_{d,g} - \bar{z}_{d\pi})\alpha_d - u_g$$

and an estimator of the MSE is

$$\hat{V}(\hat{\theta}_g - \theta_g) = (\bar{\mathbf{x}}_{0,g,N}, \bar{z}_{d\pi})\hat{V}\{\hat{\boldsymbol{\alpha}}\}(\bar{\mathbf{x}}_{0,g,N}, \bar{z}_{d\pi})' + \hat{\alpha}_d^2 s_{zd}^2 + \hat{\sigma}_u^2, \quad (5.5.21)$$

where

$$s_{zd}^2 = \left(\sum_{g\in A} \pi_g^{-1}\right)^{-1} \sum_{g\in A} \pi_g^{-1}(\bar{z}_{dg} - \bar{z}_{d\pi})^2.$$

Example 5.5.2. To illustrate the construction of small area estimates constrained to match the regression estimator, we use the data of Table 5.13. We assume, for the purposes of this example, that the sample of 44 counties is a simple random sample selected from a population of 48 counties. Then $\pi_g = 44/48$ for all counties. To satisfy (5.5.14) and (5.5.16), we add multiples of $\pi_g^{-1}N_g(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)$ and N_g to model (5.5.10), letting

$$\begin{aligned}\bar{\mathbf{x}}_{gN} &= (1, \bar{x}_{2,g,N}, \bar{x}_{3,g,N}, \bar{x}_{4,g,N}) \\ &=: [1, 0.1(\bar{r}_{1,g,N} - 59), 0.01N_g, 0.01N_g(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)].\end{aligned}$$

We iterate the estimation procedure redefining $\bar{\mathbf{x}}_{gN}$ at each step and using $\hat{\sigma}_u^2$ from the previous step until

$$(m-k)^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 1.00, \quad (5.5.22)$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y},$$

$\hat{\mathbf{V}} = \text{diag}(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)$, the gth row of $\mathbf{X}$ is $\bar{\mathbf{x}}_{gN}$, and $m - k = 40$. The vector of estimates is

$$\begin{aligned}(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\sigma}_u^2) = &(0.800, \ 0.160, -0.016, \ 0.452, \ 0.0256), \\ &(0.111)\ (0.026)\ (0.026)\ (0.897)\ (0.0066)\end{aligned} \quad (5.5.23)$$

where the standard errors for the elements of $\hat{\boldsymbol{\beta}}$ are the square roots of the diagonal elements of $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$. The variance of $\hat{\sigma}_u^2$ was estimated by

$$\begin{aligned}\hat{V}\{\hat{\sigma}_u^2\} &= 2(m-k)^{-1}\left(m^{-1}\sum_{g=1}^{m}(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)^{-1}\right)^{-2} \\ &= (20)^{-1}(33.8031)^{-2} = (0.0066)^2, \quad (5.5.24)\end{aligned}$$

obtained from a Taylor expansion of equation (5.5.2). See Exercise 16. The estimate of σ_u^2 differs from that of Example 5.5.1 because of the additional

explanatory variables and because the current estimator differs from the maximum likelihood estimator used in Example 5.5.1.

If the small areas are not selected with equal probability, the estimator of σ_u^2 should recognize this fact. See Pfeffermann et al. (1998).

Because $x_{d,g} = x_{4,g}$ is unknown for the unobserved counties, we construct the design-consistent regression estimator (5.5.19) for the mean of y. We define $z_4 = z_d$ to be the deviations from the weighted regression of x_4 on $(1, \bar{x}_{2,gN}, \bar{x}_{3,g,N})$. Then the vector of regression coefficients for the weighted regression of y on $(1, \bar{x}_{2,gN}, \bar{x}_{3,g,N}, z_{4g})$, the $\hat{\boldsymbol{\alpha}}$ of (5.5.18), is

$$\begin{array}{rcl} \hat{\boldsymbol{\alpha}}' & = & (0.831, \ \ 0.161, \ -0.003, \ \ 0.453). \\ & & \ (0.090) \ (0.026) \ \ (0.004) \ (0.897) \end{array} \tag{5.5.25}$$

A design-consistent estimator of the mean of z_d is

$$\begin{aligned} \bar{z}_{d\pi} &= \left(\sum_{g=1}^{44} N_g \right)^{-1} \sum_{g=1}^{44} N_g z_{dg} \\ &= -0.0013 \end{aligned}$$

and the regression estimator (5.5.19) of the population mean of y is

$$\begin{aligned} \bar{y}_{reg} &= (1, \ \bar{x}_{2,\cdot,N}, \ \bar{x}_{3,\cdot,N}, \ \bar{z}_{d\pi})\hat{\boldsymbol{\alpha}} \\ &= (1, \ 0.1601, \ 23.9906, \ -0.0013)\hat{\boldsymbol{\alpha}} \\ &= 0.7887. \end{aligned}$$

The regression estimated mean of 0.7887 corresponds to an estimated total of 81,441.28.

The county predictions of (5.5.20) are given in Table 5.14. The values predicted differ slightly from those of Table 5.13, primarily because of the difference in $\hat{\sigma}_u^2$. The standard errors for the predictions for counties with a y observation are computed using (5.5.9), and the standard errors for counties with no y observation are computed using (5.5.21).

The term $\hat{\alpha}_d^2 s_{zd}^2 = (0.453)^2(0.00099)$ adds little to the variance of predictions for areas with no y observations because s_{zd}^2 is small. The standard errors of Table 5.14 are sometimes slightly larger than those of Table 5.13 because more parameters are being estimated and the estimate of σ_u^2 is larger than that used to construct Table 5.13.

Table 5.14: **Predictions with Sum Constrained to Regression Estimator**

County	x_{4g}	$\bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$	$\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$	$\hat{\theta}_g$	S.E. of Prediction
3	0.4084	0.5906	–0.1615	0.466	0.077
15	0.6739	0.7481	–0.0831	0.679	0.067
21	0.6549	0.8799	0.2034	1.040	0.074
27	0.6716	0.5711	0.2173	0.752	0.066
33	0.6345	0.6584	0.2105	0.832	0.067
35	0.5256	0.9995	0.1250	1.094	0.079
41	0.6131	0.7679	–0.0854	0.699	0.070
47	0.8257	0.5924	–0.1842	0.439	0.066
59	0.3792	0.7075	0.1312	0.807	0.079
63	0.5184	0.8486	–0.0950	0.773	0.072
67	0.4920	0.5708	0.1189	0.659	0.082
71	0.3827	0.7469	0.1802	0.891	0.072
73	0.5398	0.7120	0.2327	0.889	0.079
75	0.6976	0.4901	0.1287	0.590	0.077
77	0.7290	0.5993	–0.1246	0.500	0.073
79	0.5850	1.0505	–0.2605	0.857	0.083
83	0.6973	0.7374	–0.0902	0.664	0.070
85	0.6071	0.8793	–0.1525	0.752	0.066
91	0.5878	0.7262	0.3936	1.040	0.072
93	0.4389	0.8471	–0.1703	0.724	0.085
109	0.7533	0.8337	0.1339	0.944	0.067
119	0.4440	0.7934	–0.0907	0.713	0.055
129	0.3819	0.7881	–0.1719	0.658	0.079
131	0.3904	0.6374	–0.2157	0.481	0.085
133	0.7659	0.9541	0.0912	1.033	0.060
135	0.3386	0.5667	–0.2033	0.404	0.073
141	0.4827	1.1623	0.2614	1.357	0.083
143	0.4788	0.7576	0.2175	0.913	0.085
145	0.4970	0.4828	–0.0314	0.457	0.071
147	0.7521	0.7834	0.1619	0.915	0.069
149	1.0874	1.0855	–0.0208	1.069	0.072
151	0.5777	0.8694	0.0488	0.905	0.085
153	0.4325	0.6032	0.0670	0.658	0.067
155	1.1719	0.7575	–0.1386	0.640	0.064
157	0.6246	0.5485	0.0294	0.571	0.077

Continued

County	x_{4g}	$\bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$	$\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$	$\hat{\theta}_g$	S.E. of Prediction
161	0.6796	0.8839	–0.1653	0.750	0.070
165	0.6997	0.6018	–0.2261	0.430	0.080
167	0.7691	0.9081	0.0462	0.951	0.046
169	0.5223	0.7043	–0.1215	0.606	0.070
187	0.8567	0.7503	0.1239	0.849	0.073
189	0.5210	1.0795	0.1767	1.208	0.086
193	0.6422	1.0192	–0.1143	0.926	0.069
195	0.3618	0.5802	0.0189	0.595	0.071
197	0.5403	0.9933	–0.4164	0.684	0.082
201	0.5282	0.8528	NA	0.853	0.164
202	0.4504	0.7434	NA	0.743	0.165
203	1.0408	1.0683	NA	1.068	0.180
204	0.8240	0.7362	NA	0.736	0.167

Because

$$\sum_{g \in A} N_g \hat{\gamma}_g (\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}) = 0,$$

the weighted sum of the predicted values for the 48 counties in Table 5.14 is

$$\sum_{g=1}^{48} N_g \hat{\theta}_g = 81,441.28,$$

equal to the regression estimator of the total. The sum of the predictions in Table 5.13 is $81,605$.

In this example wind erosion has a small correlation with the sampling weight, so the weighted sum of predictions constructed with N_g and $N_g(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)$ included in the set of explanatory variables differs little from the weighted sum constructed with only $(1, \bar{x}_{1,g,N})$ as the explanatory vector. ■■

5.6 MEASUREMENT ERROR

5.6.1 Introduction

Essentially all data are collected subject to measurement error, and the design of collection instruments to minimize measurement error is an important part of the discipline of survey sampling. The mean and variance of the measurement process are both important. The mean is typically the most

difficult to evaluate, because some determination of the "truth" is required. In survey sampling, external sources, perhaps available at a later time, can sometimes be used to estimate bias. If a reliable external source is available, the collection instrument can be recalibrated.

The variance of the measuring operation can sometimes be estimated by repeated independent determinations on the same element. One measure of the relative magnitude of the variance of measurement error is the correlation between two independent determinations on a random sample. This measure is called the *attenuation coefficient* and is denoted by κ_{xx} for the variable x. We adopt the convention of using a lowercase letter to denote the true value and a capital letter to denote the observed value, where the observed value is the sum of the true value and the measurement error. The κ_{xx} gives the relative bias in the simple regression coefficient of y on X as an estimator of the population regression of y on x. For continuous variables, κ_{xx} is the ratio of the variance of the true x to the variance of observed X.

Fuller (1987b, p. 8) reports on a large study conducted by the U.S. Census Bureau in which determinations were made on a number of demographic variables in the Decennial Census and in the Current Population Survey. The two surveys gave two nearly independent determinations. The attenuation coefficient for education was 0.88, that for income was 0.85, and that for fraction unemployed was 0.77. Thus, for example, of the variation observed in a simple random sample of incomes, 15% is due to measurement error.

5.6.2 Simple estimators

Consider a simple measurement error model in which the observation is the true value plus a zero-mean measurement error. Let X_i be the observed value, and x_i be the true value, so that

$$\begin{aligned} X_i &= x_i + u_i, \\ u_i &\sim (0, \sigma_u^2), \end{aligned} \tag{5.6.1}$$

where u_i is the measurement error. Assume that u_i is independent of x_j for all i and j. Assume that a sample of size n is selected from a finite population of x values and that the measurement process satisfies (5.6.1). Let the Horvitz–Thompson estimator of the total be constructed as

$$\hat{T}_X = \sum_{i \in A} w_i X_i, \tag{5.6.2}$$

where $w_i = \pi_i^{-1}$. Because the estimator based on the true values,

$$\hat{T}_x = \sum_{i \in A} w_i x_i,$$

is unbiased for T_x, $\hat{T}_X$ is also unbiased for T_x. That is,

$$E\{E(\hat{T}_X \mid \mathcal{F}_X) \mid \mathcal{F}_x\} = E\{T_X \mid \mathcal{F}_x\} = T_x,$$

where $\mathcal{F}_X = (X_1, X_2, \ldots, X_N)$ and $\mathcal{F}_x = (x_1, x_2, \ldots, x_N)$ is the set of true x values.

Given that x_i is independent of u_j for all i and j.

$$V\{\hat{T}_X - T_x \mid \mathcal{F}_x\} = V\{\hat{T}_x - T_x \mid \mathcal{F}_x\} + V\left\{\sum_{i\in A} w_i u_i\right\}, \tag{5.6.3}$$

where $\hat{T}_x = \Sigma_{i\in A} x_i$. The variance of $\hat{T}_u = \Sigma_{i\in A} w_i u_i$ is a function of the covariance structure of $\mathbf{u} = (u_i, u_2, \ldots, u_n)$, and if $u_i \sim ind(0, \sigma_u^2)$, independent of (x_j, w_j) for all i and j, then

$$\begin{aligned} V\{\hat{T}_X - T_x \mid \mathcal{F}_x\} &= V\{\hat{T}_x - T_x \mid \mathcal{F}_x\} + E\left\{\sum_{i\in A} w_i^2 \sigma_u^2\right\} \\ &= V\{\hat{T}_x - T_x \mid \mathcal{F}_x\} + \sum_{i\in U} w_i \sigma_u^2. \end{aligned} \tag{5.6.4}$$

Furthermore,

$$\begin{aligned} E\{\hat{V}_{HT}(\hat{T}_X - T_X \mid \mathcal{F}_X) \mid \mathcal{F}_x\} &= V\{\hat{T}_x - T_x \mid \mathcal{F}_x\} \\ &\quad + E\left\{\sum_{k\in A}\sum_{j\in A} \pi_{jk}^{-1}(\pi_{jk} - \pi_j\pi_k) w_j u_j w_k u_k\right\} \\ &= V\{\hat{T}_x - T_x\} + \sum_{i\in U}(1 - \pi_i) w_i \sigma_u^2, \end{aligned} \tag{5.6.5}$$

where

$$\hat{V}_{HT}\left(\hat{T}_X - T_X \mid \mathcal{F}_X\right) = \sum_{k\in A}\sum_{j\in A} \pi_{jk}^{-1}(\pi_{jk} - \pi_j\pi_k) w_j X_j w_k X_k.$$

It follows that

$$E\{\hat{V}_{HT}(\hat{T}_X - T_X \mid \mathcal{F}_X) \mid \mathcal{F}_x\} - V\{\hat{T}_X - T_x \mid \mathcal{F}_x\} = -N\sigma_u^2. \tag{5.6.6}$$

Thus, if the measurement errors have zero means, are independent of x and w, and are independent, the expectation of a design linear estimator in the observed values is equal to the expectation of the estimator in the true variables.

Furthermore, the bias in the Horvitz–Thompson estimator of variance is small if the sampling rates are small. See Exercise 6.

The assumption of independent measurement errors is critical for result (5.6.6) and the result does not hold with personal interviews, where each interviewer collects data from several respondents. See Hansen et al. (1951). The traits of the interviewer lead to correlation among responses obtained by that interviewer. To illustrate the effect of correlated measurement error, assume that a simple random sample of size $n = mk$ is selected and each of k interviewers is given an assignment of m interviews, where assignment is at random. Assume that interviewers have an effect on the responses and that it is reasonable to treat the k interviewers as a random sample from the population of interviewers. With these assumptions a representation for the observations is

$$Y_{gj} = \mu + e_j + \alpha_g + \epsilon_{gj}, \tag{5.6.7}$$

where μ is the superpopulation mean, $e_j = y_j - \mu$, y_j is the true value, α_g is the interviewer effect for interviewer g, and ϵ_{gj} is the measurement error for person j interviewed by interviewer g. We consider the relatively simple specification

$$\alpha_g \sim ii(0, \sigma_\alpha^2),$$

$$\epsilon_{gj} \sim ii(0, \sigma_\epsilon^2),$$

and assume α_g, ϵ_{rj}, and e_i to be independent for all g, r, j, and i. Then

$$\begin{aligned} V\{\bar{Y}_{..} - \bar{y}_N \mid \mathcal{F}_y\} &= V\{\bar{e}_{.} + \bar{\alpha}_{.} + \bar{\epsilon}_{..} \mid \mathcal{F}_y\} \\ &= (1 - f_n)n^{-1}S_e^2 + k^{-1}\sigma_\alpha^2 + n^{-1}\sigma_\epsilon^2, \end{aligned} \tag{5.6.8}$$

where

$$\begin{aligned} (\bar{Y}_{..}, \bar{e}_{.}, \bar{\epsilon}_{..}) &= n^{-1} \sum_{gj \in A} (Y_{gj}, e_j, \epsilon_{gj}), \\ \bar{\alpha}_{.} &= k^{-1} \sum_{g=1}^{k} \alpha_g, \end{aligned}$$

and $f_n = N^{-1}n$. Because there is one observation per person, the summation over gj is a summation over persons. For large interviewer assignments the term $k^{-1}\sigma_\alpha^2$ can be very important even when σ_α^2 is relatively small.

The usual estimator of variance is seriously biased. For a simple random sample,

$$E\{s_Y^2\} = \sigma_e^2 + \sigma_\epsilon^2 + n(n-1)^{-1}m^{-1}(m-1)\sigma_\alpha^2, \tag{5.6.9}$$

where

$$s_Y^2 = (n-1)^{-1} \sum_{gj \in A} (Y_{gj} - \bar{Y}_{..})^2$$

and $\sigma_e^2 = E\{S_e^2\}$.

Example 5.6.1. Let model (5.6.7) hold and assume that σ_α^2 is 2% of σ_e^2 and that σ_ϵ^2 is 15% of σ_e^2. Let a simple random sample of 1000 be selected, and let each of 20 interviewers be given a random assignment of 50 interviews. If the finite population correction can be ignored, the variance of the sample mean is

$$V\{\bar{Y}_{..} - \bar{y}_N\} = n^{-1}(\sigma_e^2 + 0.15\sigma_e^2) + 0.02k^{-1}\sigma_e^2 = 0.00215\sigma_e^2.$$

Although the variance of the interviewer effect is small, it makes a large contribution to the variance because each interviewer has a large number of interviews. By (5.6.9)

$$E\{s_Y^2\} = 1.1696\sigma_e^2$$

and the usual estimator of variance of $\bar{Y}_n$ has a bias of -45.6%.

By treating interviewer assignments as the first stage of a two-stage sample, it is possible to construct an unbiased estimator of the variance of $\bar{Y}_{..}$. Under the assumptions that the interviewer assignments are made at random and that the finite population correction can be ignored, an unbiased estimator of $V\{\bar{Y}_{..} - \bar{y}_N\}$ is

$$\hat{V}\{\bar{Y}_{..} - \bar{y}_N\} = (380)^{-1} \sum_{g=1}^{20} (\bar{Y}_{g.} - \bar{Y}_{..})^2,$$

where $\bar{Y}_{g.}$ is the mean for the gth interviewer. ■■

Many large-scale surveys are stratified multistage samples. In such surveys the interviewer assignments are often primary sampling units. If an interviewer is not assigned to more than one primary sampling unit, and if the finite population correction can be ignored, the usual variance estimator for a design linear estimator remains appropriate.

Situations that seem simple at first can be quite difficult in the presence of measurement error. The estimation of the distribution function is an example. Assume that

$$Y_i = y_i + e_i,$$
$$y_i \sim (\mu, \sigma_y^2),$$

where $e_i \sim NI(0, \sigma_e^2)$, independent of y_i. Then the mean of Y_i for a simple random sample is an unbiased estimator of μ, but the estimator

$$\hat{F}(y_o) = n^{-1} \sum_{i \in A} \delta_i(y_o),$$

where

$$\begin{aligned} \delta_i(y_o) &= 1 \qquad \text{if } Y_i \leq y_o \\ &= 0 \qquad \text{otherwise,} \end{aligned}$$

is, in general, biased for the probability that $y_i < y_o$. The mean of e_i is zero for y_i, but the mean of the measurement error for $\delta_i(y_o)$ is not zero.

If $e_i \sim NI(0, \sigma_e^2)$ and if $y_i \sim N(\mu, \sigma_y^2)$ then $Y_i \sim N(\mu, \sigma_y^2 + \sigma_e^2)$ and the parameters of the distribution function of y_i are easily estimated. Similarly, one can use likelihood methods to estimate the parameters of the distribution if one can specify the form of the distribution of y_i and the form of the error distribution.

Nonparametric or semiparametric estimation of the distribution function in the presence of measurement error is extremely difficult. See Stefanski and Carroll (1990), Cook and Stefanski (1994), Nusser et al. (1996), Cordy and Thomas (1997), Chen, Fuller, and Breidt (2002), and Delaigle, Hall, and Meister (2008). If σ_e^2 is known, the variable

$$z_i = \bar{Y} + [\hat{\sigma}_Y^{-2}(\hat{\sigma}_Y^2 - \sigma_e^2)]^{-1/2}(Y_i - \bar{Y})$$

has sample mean and variance equal to estimators of the mean and variance of y, where those estimators are $(\hat{\mu}_y, \hat{\sigma}_y^2) = (\bar{Y}, \hat{\sigma}_Y^2 - \sigma_e^2)$ and $(\bar{Y}, \hat{\sigma}_Y^2)$ is an estimator of the mean and variance of Y. Therefore, the sample distribution function of z is a first approximation to the distribution function of y that can be used to suggest parametric models for the distribution of y_i.

5.6.3 Complex estimators

As demonstrated in the preceding section, measurement error with zero mean increases the variance of linear estimators, but the estimators remain unbiased. Alternatively, the expectation of nonlinear estimators, such as regression coefficients, can be seriously affected by zero-mean measurement error. Consider the simple regression model,

$$y_i = \beta_0 + x_{1i}\beta_1 + e_i, \tag{5.6.10}$$

where $e_i \sim (0, \sigma_e^2)$ independent of x_{1i}. Assume that the observation on the explanatory variable is $X_{1i} = x_{1i} + u_i$, where $u_i \sim ind(0, \sigma_u^2)$ is the

measurement error as specified in (5.6.1). Consider the finite population to be a simple random sample from an infinite population where (y_i, x_{1i}) satisfies (5.6.10). Given a probability sample, the weighted estimator

$$\hat{\gamma} = (\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}y, \tag{5.6.11}$$

where $\mathbf{D}_\pi = \text{diag}(\pi_1, \pi_2, \ldots, \pi_n)$, $\mathbf{X}' = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n')$, and $\mathbf{X}_i = (1, X_{1i})$, was discussed in Chapter 2. We call the estimator $\hat{\gamma}$ because, in the presence of measurement error, we shall see that $\hat{\gamma}$ is a biased estimator of β. The estimator of the coefficient for X_{1i} can be written

$$\hat{\gamma}_1 = \left(\sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})^2\pi_i^{-1}\right)^{-1} \sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})\pi_i^{-1}(y_i - \bar{y}_\pi). \tag{5.6.12}$$

Under the assumption that u_i is independent of (x_{1i}, π_i, e_i), and under the usual assumptions required for consistency of a sample mean,

$$E\left\{N^{-1}\sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})^2\pi_i^{-1} \mid \mathcal{F}_{(x,y),N}\right\} = S_{x,N}^2 + \sigma_u^2 + O_p(n^{-1})$$

and

$$E\left\{N^{-1}\sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})(y_i - \bar{y}_\pi)\pi_i^{-1} \mid \mathcal{F}_{(x,y),N}\right\} = S_{xy,N} + O_p(n^{-1}),$$

where $\mathcal{F}_{(x,y),N} = [(y_1, x_{1,1}), (y_2, x_{1,2}), \ldots, (y_N, x_{1,N})]$ is the finite population of values for the true x_1 and y. Then

$$\hat{\gamma}_1 = (S_{x,N}^2 + \sigma_u^2)^{-1}S_{xy,N} + O_p(n^{-1/2}) \tag{5.6.13}$$

and $\hat{\gamma}_1$ is a consistent estimator of $(\sigma_x^2 + \sigma_u^2)^{-1}\sigma_{xy} = \kappa_{xx}\beta_1$.

Consistent estimation of β_1 requires additional information beyond the set of (y_i, X_i) vectors. There are several forms for such information. If we know κ_{xx} or have an estimator of κ_{xx}, then

$$\tilde{\beta}_{1,\kappa} = \kappa_{xx}^{-1}\hat{\gamma}_1 \tag{5.6.14}$$

is a consistent estimator of β_1. For example, one could use the estimate of κ_{xx} from the U.S. census study if the explanatory variable is education. Similarly, if σ_u^2 is known, or estimated,

$$\tilde{\beta}_{1,\sigma} = \left(\sum_{i\in A}[(X_{1i} - \bar{X}_{1\pi})^2 - \sigma_u^2\,]\pi_i^{-1}\right)^{-1} \sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})\pi_i^{-1}(y_i - \bar{y}_\pi) \tag{5.6.15}$$

is a consistent estimator of β_1. The matrix expression for the estimator of $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}}_\sigma = (\mathbf{M}_{X\pi X} - \Sigma_{uu})^{-1}\mathbf{M}_{X\pi y},$$

where $\Sigma_{uu} = \text{diag}(0, \sigma_u^2)$, $\mathbf{X}_i = (1, X_{1i})$, and

$$(\mathbf{M}_{X\pi X}, \mathbf{M}_{X\pi y}) = N^{-1}\sum_{i\in A}\mathbf{X}_i'\pi_i^{-1}(\mathbf{X}_i, y_i).$$

Our usual Taylor approximation gives

$$\hat{V}\{\tilde{\boldsymbol{\beta}}_\sigma\} = (\mathbf{M}_{X\pi X} - \Sigma_{uu})^{-1}\hat{V}_{HT}\{\bar{\mathbf{b}}_{HT}\}(\mathbf{M}_{X\pi X} - \Sigma_{uu})^{-1},$$

where $\mathbf{b}_i = \mathbf{x}_i' a_i$,

$$\bar{\mathbf{b}}_{HT} = N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{x}_i' a_i$$

and

$$\hat{V}_{HT}\{\bar{\mathbf{b}}_{HT}\} = \hat{V}_{HT}\left\{\left(\sum_{i\in A}\pi_i^{-1}\right)^{-1}\sum_{i\in A}\mathbf{X}_i'\pi_i^{-1}a_i\right\}$$

is computed using $\hat{a}_i = y_i - \bar{y}_{HT} - (X_{1i} - \bar{X}_{1HT})\hat{\boldsymbol{\beta}}_\sigma$. In most situations one is interested in $\boldsymbol{\beta}$ as a superpopulation parameter. If so, finite population corrections are not appropriate. See Chapter 6.

Establishing the relationship between two error-prone measures is an important application of measurement error models. In some situations there is a relatively inexpensive procedure appropriate for large-scale data collection and an expensive procedure believed to be unbiased for the characteristic of interest. Then a subsample may be used to study the relationship between the two measures. Similarly, if one measuring procedure is to be replaced by another, it is important to establish the relationship between the two measures. Example 5.6.2 is an illustration.

Example 5.6.2. The National Resources Inventory is described in Example 1.2.2. In 2004, the Natural Resources Conservation Service changed the way in which data were collected for the segments. In 2003 and in prior years, the data collectors outlined, on a transparent overlay placed on an aerial photograph, the areas designated as developed. Developed land includes urban areas, built-up areas, and roads. Beginning in 2004 a digital process was used in which roads and certain types of developed land, such as manufacturing plants and cemeteries, were outlined digitally, but for single- or double-unit residences, the location of the residence was entered as a simple "dot" location. A computer program was designed to convert the digital information to area information. To calibrate the computer program, a study was conducted

in which two data collectors, using the new procedure, made independent determinations on segments that had been observed in 2003. The determinations were treated as independent because only unmodified photographs were available to the data collector for each determination. See Yu and Legg (2009).

We analyze data collected in the calibration study for the western part of the United States. The data are observations where at least one of the three determinations is not zero. The computer program has parameters that can be changed to improve the agreement between the old and new procedures. Our analysis can be considered to be a check on parameters determined on a different data set.

Let (Y_{1i}, Y_{2i}, X_i) be the vector composed of the first determination by the new procedure, the second determination by the new procedure, and the determination by the old procedure, respectively. In all cases, the variable is the fraction of segment acres that are developed. The old procedure is assumed to be unbiased for the quantity of interest. Our analysis model is

$$\begin{aligned} y_{ji} &= \beta_0 + \beta_1 x_i, \\ (Y_{1i}, Y_{2i}, X_i) &= (y_i, y_i, x_i) + (e_{1i}, e_{2i}, u_i), \\ x_i &\sim ind(0, \sigma_x^2), \\ e_{ji} &\sim ind(0, \sigma_{ei}^2), \\ u_i &\sim ind(0, \sigma_{ui}^2), \end{aligned} \tag{5.6.16}$$

for $j = 1,\ 2$, and it is assumed that u_i, e_{jt}, and x_i are mutually independent. It seems reasonable that the measurement error has smaller variance for segments with a small fraction of developed land than for segments with a fraction near 50%. One could specify a model for the error variance, but we estimate the model estimating the average variance of u_i, denoted by $\sigma_{a,u}^2$, and the average variance of e_i, denoted by $\sigma_{a,e}^2$.

To estimate the parameters, it is convenient to define the vector

$$\mathbf{Z}_i = (Z_{1i}, Z_{2i}, Z_{3i}) = [X_i,\ 0.5(Y_{1i} + Y_{2i}),\ (0.5)^{0.5}(Y_{1i} - Y_{2i})] \tag{5.6.17}$$

and let

$$\mathbf{m} = (n-1)^{-1} \sum_{i \in A} (\mathbf{Z}_i - \bar{\mathbf{Z}})'(\mathbf{Z}_i - \bar{\mathbf{Z}}).$$

Then

$$E\{\mathbf{m}\} = \begin{bmatrix} \sigma_x^2 + \sigma_{a,u}^2 & \beta_1 \sigma_x^2 & 0 \\ \beta_1 \sigma_x^2 & \beta_1^2 \sigma_x^2 + 0.5\sigma_{a,e}^2 & 0 \\ 0 & 0 & \sigma_{a,e}^2 \end{bmatrix}. \tag{5.6.18}$$

If the elements of $\mathbf{Z}_i$ are normally distributed, m_{13} and m_{23} contain no information about the parameters because $E\{m_{13}, m_{23}\} = \mathbf{0}$ and the covariance between (m_{13}, m_{23}) and the other elements of $\mathbf{m}$ is the zero vector. If the distribution is not normal, it is possible that (m_{13}, m_{23}) contains information, but we ignore that possibility and work only with $(m_{11}, m_{12}, m_{22}, m_{33})$. By equating the sample moments to their expectations, we obtain the estimators

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Z}_1 - \hat{\beta}_1 \bar{Z}_2, \\
\hat{\beta}_1 &= m_{12}^{-1}(m_{22} - 0.5m_{33}) = m_{12}^{-1}\hat{\sigma}_y^2, \\
\hat{\sigma}_x^2 &= (m_{22} - 0.5m_{33})^{-1} m_{12}^2 = \hat{\sigma}_y^{-2} m_{12}^2, \\
\hat{\sigma}_{a,e}^2 &= m_{33}, \\
\hat{\sigma}_{a,u}^2 &= m_{11} - (m_{22} - 0.5m_{33})^{-1} m_{12}^2 = m_{11} - \hat{\sigma}_x^2, \quad (5.6.19)
\end{aligned}$$

where $\hat{\sigma}_y^2 = m_{22} - 0.5m_{33}$.

For our sample of 382 segments,

$$\begin{aligned}
(\bar{Z}_1, \bar{Z}_2) &= \begin{matrix} (0.1433, & 0.1464), \\ (0.0073) & (0.0070) \end{matrix} \\
(m_{11}, m_{12}, m_{22}, m_{33}) &= \begin{matrix} (2.061, & 1.758, & 1.873, & 0.060) \times 10^{-2}, \\ (0.184) & (0.129) & (0.120) & (0.014) \end{matrix} \\
(\hat{\beta}_0, \hat{\beta}_1) &= \begin{matrix} (-0.0039, & 1.0482), \\ (0.0037) & (0.0374) \end{matrix}
\end{aligned}$$

and

$$\begin{aligned}
100(\hat{\sigma}_{a,e}^2, \hat{\sigma}_{a,u}^2, \hat{\sigma}_x^2) &= \begin{matrix} (0.0601, & 0.3837, & 1.6775), \\ (0.0139) & (0.0546) & (0.1583) \end{matrix}
\end{aligned}$$

where the standard errors were calculated with 382 delete-one jackknife replicates. The jackknife is appropriate under model (5.6.16) because all estimators are continuous differentiable functions of the moments. Also see Exercise 15.

The calibration sample was selected to have a much higher fraction of segments with developed land than the general population of segments. Because our analysis is conducted on unweighted data, $(\hat{\mu}_x, \hat{\sigma}_x^2)$ is not an estimate for the general population.

The estimated variance of $\hat{\sigma}_{a,e}^2$ is much larger than one would expect if $e_i \sim N(0, \sigma_e^2)$. There are two reasons: (1) the e_i have unequal variances, and (2) the distribution of the measurement errors has long tails.

The approximate F test for $H_0 : (\beta_0, \beta_1) = (0, 1)$ is

$$F(2, 380) = 0.5(\hat{\beta}_0, \hat{\beta}_1 - 1)[\hat{V}\{\hat{\beta}_0, \hat{\beta}_1\}]^{-1}(\hat{\beta}_0, \hat{\beta}_1 - 1)' = 0.83,$$

where the estimated covariance matrix for $(\hat{\beta}_0, \hat{\beta}_1)$ is

$$\hat{V}\{(\hat{\beta}_0, \hat{\beta}_1)\} = \begin{pmatrix} 0.1397 & -1.0746 \\ -1.0746 & 14.0068 \end{pmatrix} \times 10^{-4}.$$

Also, plots of the data give little indication of a nonlinear relationship between the old and new procedures. Therefore, the data are consistent with the hypothesis that the new measuring procedure produces values that are equal to the true value plus a zero-mean measurement error. ■■

5.7 REFERENCES

Section 5.1. Chang and Kott (2008), Fuller and An (1998), Fuller, Loughin, and Baker (1994), Kalton and Kasprzyk (1986), Kalton and Kish (1984), Kott (2006b), Little (1983b, 1988), Little and Rubin (2002), Meng (1994), Sande (1983), Särndal (1992).

Section 5.2. Chen and Shao (2000, 2001), Fay (1996, 1999), Fuller and Kim (2005), Kim and Fuller (2004), Kim, Fuller, and Bell (2009), Rancourt, Särndal, and Lee (1994), Rao and Shao (1992), Särndal (1992).

Section 5.3. Cressie (1991), Wolter (2007).

Section 5.4. Fuller (1991), Hidiroglou and Srinath (1981), Rivest (1994).

Section 5.5. Battese, Harter, and Fuller (1988), Fay and Herriot (1979), Ghosh and Rao (1994), Harville (1976), Kackar and Harville (1984), Mukhopadhyay (2006), Pfeffermann and Barnard (1991), Prasad and Rao (1990, 1999), Rao (2003), Robinson (1991), Wang and Fuller (2003), Wang, Fuller, and Qu (2009), You and Rao (2002).

Section 5.6. Biemer et al. (1991), Carroll, Ruppert, and Stephanski (1995), Fuller (1987b, 1991b, 1995). Hansen et al. (1951), Hansen, Hurwitz, and Pritzker (1964), Yu and Legg (2009).

5.8 EXERCISES

1. (a) (Section 5.2) Let a simple random sample of size n have m missing values and r respondents. Assume that response is independent of

y. Let a simple random replacement sample of M of the respondents be used as a set of donors for each missing value. Each donated value is given a weight of $n^{-1}M^{-1}$. Show that the variance of the imputed mean is

$$V\{\bar{y}_I \mid (\mathcal{F}, m)\} = (r^{-1} - N^{-1})S_y^2 + n^{-2}M^{-1}mr^{-1}(r-1)S_y^2.$$

(b) Assume that m values are missing and r are present in a simple random sample of size n and assume that the probability of response is independent of y. The data set is completed by imputing a single value for each missing value. Let $m = kr + t$, where k is the largest nonnegative integer such that $kr \leq m$. Consider a hot deck procedure in which $r - t$ respondents are used as donors k times, and t respondents are used as donors $k+1$ times. In human nonresponse, k is generally zero. The t respondents are chosen randomly from the r. Show that the variance of the imputed mean is

$$V\{\bar{y}_I \mid \mathcal{F}, m\} = \left(r^{-1} - N^{-1}\right) S_y^2 + n^{-2}tr^{-1}\left(r - t\right) S_y^2.$$

2. (Section 5.2) Using the data of Table 5.2, compute the estimated mean of y for each of the three x-categories. Using the replicates of Table 5.5, estimate the variance of your estimates.

3. (Section 5.2) Using the replicates of Table 5.7, estimate the variance of the estimated mean of x, where x is as given in Table 5.6. Is this a design-unbiased estimator?

4. (Section 5.2) Using the imputed data of Table 5.7, compute the weighted regression for y on x using the weights in the table. Using the replicates of Table 5.9, compute the estimated covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1)$, where $(\hat{\beta}_0, \hat{\beta}_1)$ is the vector of estimated coefficients and $\hat{y}_t = \hat{\beta}_0 + x_t\hat{\beta}_1$. In this simple example, do you think the regression coefficient based on the imputed estimator is a good estimator? Would your answer change if we had a sample of 100 observations with 30 y-values missing?

5. (Section 5.3) Assume that a one-per-stratum sample is selected from a population with stratum sizes N_h, $h = 1, 2, \ldots, H$. Let $\pi_{h,i}$, $i = 1, 2, \ldots, N$, be the selection probabilities, where the subscript h is redundant but useful. Assume that H is even and that the strata are collapsed to form $H/2$ strata. Let $y_{h,i}$ be the sample observation in the hth original stratum and let

$$\hat{V}\{\hat{T} \mid \mathcal{F}\} = 0.5 \sum_{h=1}^{H/2} (\pi_{2h,i}^{-1} y_{2h,i} - \pi_{2h-1,i}^{-1} y_{2h-1,i})^2,$$

where

$$\hat{T} = \sum_{h=1}^{H} \pi_{h,i}^{-1} y_{h,i}.$$

Assume that the finite population is a sample from a superpopulation with

$$y_{h,i} \sim ind(\mu_h,\ \sigma_h^2) \text{ for } (h,i) \in U_h.$$

What is the expected value of $\hat{V}\{\hat{T} \mid \mathcal{F}\}$?

6. (Section 5.6) Assume that it is known that the measurement error variance for a variable x is σ_ϵ^2. Let a stratified sample be selected, where the observations are $X_{hj} = x_{hj} + \epsilon_{hj}$. Let the usual estimator of variance of the estimated total be

$$\hat{V}\{\hat{T}_X \mid \mathcal{F}\} = \sum_{h=1}^{H} N_h^2 (1 - f_h) n_h^{-1} s_h^2,$$

where $f_h = N_h^{-1} N_h$, s_h^2 is the sample stratum variance of X, and N_h is the stratum size. Assume the model (5.6.1) holds for each h. Show that

$$\tilde{V}\{\hat{T}_X - T_x \mid \mathcal{F}_x\} = \sum_{h=1}^{H} N_h^2 n_h^{-1} [(1 - f_h) s_h^2 + f_h \sigma_\epsilon^2]$$

is an unbiased estimator of the variance of $\hat{T}_X - T_x$.

7. (Section 5.6) Let a simple random sample of size n be selected from a population of size N. Let duplicate measures be made on $m, m < n$ of the observations, where $X_{ij}, j = 1, 2$ are the measurements. Assume that

$$X_{ij} = x_i + u_{ij},$$
$$u_{ij} \sim ind(0, \sigma_u^2),$$

where u_{ij} is independent of x_t for all ij and t. Let

$$\hat{T}_X = N n^{-1} \left(\sum_{i \in A_r} \bar{X}_{i\cdot} + \sum_{i \in A_s} X_i \right),$$

where $\bar{X}_{i\cdot}$ is the mean of the two determinations on elements with two determinations, A_r is the set of indices for elements with two

determinations and A_s is the set of indices for elements with a single determination. What is the variance of $\hat{T}_X$ as an estimator of T_x? Give an unbiased estimator of $V\{\hat{T}_X - T_x \mid \mathcal{F}_x\}$. Include the finite population correction.

8. (Section 5.6) Let the predictor of u_g for known σ_e^2 and σ_{eg}^2 be given by (5.5.2). Show that $E\{(\hat{u}_g - u_g)^2\} = (\sigma_u^2 + \sigma_{eg}^2)^{-1}\sigma_u^2\sigma_{eg}^2$, where $\hat{u}_g$ is as defined in (5.5.2).

9. (Section 5.5) Prove that the estimator (5.5.6) is unbiased for θ_g in the sense that $E\{\hat{\theta}_g - \theta_g\} = 0$. Derive expression (5.5.7).

10. (Section 5.5) The y_g values for counties 201, 202, 203, and 204 were treated as unobserved in Table 5.13. Assume that the values are 0.754, 0.975, 1.065, and 0.874, with n_g values of 15, 10, 16, and 15, respectively. Are the predicted values in the table consistent with these observations? Using the estimated parameters given in Example 5.5.1, compute predicted values using the given observations.

11. (Section 5.5) The standard error for $\hat{\sigma}_u^2$ of Example 5.5.1 is 0.0062. Use a Taylor approximation to estimate the variance of $\hat{\gamma}_3$ for county 3 of Table 5.13, treating σ_e^2 as known. Use a Taylor expansion to find the leading term in the bias of $\hat{\gamma}_g$ under the assumption that $\hat{\sigma}_u^2$ is unbiased and that σ_e^2 is known.

12. (Section 5.5) Use the facts that $\mathbf{V} = \text{diag}(\sigma_u^2 + \sigma_{eg}^2)$ and that $\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$, to show that (5.5.16) is sufficient for (5.5.15).

13. (Section 5.6) Assume the model

$$y_i = \beta_0 + x_i\beta_1 + e_i,$$
$$X_i = x_i + u_i,$$

where (e_i, u_i) is independent of x_i, and e_i is independent of u_i. It is desired to estimate β_1 using a simple random sample of n values of (y_i, X_i). How small must the true κ_{xx} be for the ordinary least squares estimator to have a MSE smaller than that of $\kappa_{xx}^{-1}\hat{\beta}_{1,ols}$? Assume that

$$\begin{pmatrix} y \\ X \end{pmatrix} \sim NI\left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} 1.4 & 1.0 \\ 1.0 & 1.0 \end{pmatrix}\right].$$

14. (Section 5.6) In Example 5.6.1 the finite population correction was ignored. Calculate the variance under the assumption that $Nn^{-1} = 0.6$. Construct an unbiased estimator of the variance of the sample mean

assuming it is known that $\sigma_\alpha^2 = 0.02$ and $\sigma_\epsilon^2 = 0.15\sigma_e^2$, but σ_e^2 is unknown.

15. (Section 5.6) The jackknife was used to estimate variances in Example 5.6.2. In this exercise we use Taylor methods. Let

$$\begin{aligned}\mathbf{b}_i &= [Z_{1i}, Z_{2i}, Z_{3i}^2, \psi_n(Z_{1i} - \bar{Z}_1)^2, \psi_n(Z_{1i} - \bar{Z}_1)(Z_{2i} - \bar{Z}_2), \\ &\quad \psi_n(Z_{2i} - \bar{Z}_2)^2, \psi_n(Z_{3i} - \bar{Z}_3)^2],\end{aligned}$$

where $\psi_n = [n(n-1)^{-1}]^{1/2}$. Then $\bar{\mathbf{b}} = (\bar{Z}_1, \bar{Z}_2, m_{11}, m_{12}, m_{22}, m_{33})$ and an estimator of the variance of $\bar{\mathbf{b}}$ is

$$\hat{V}\{\bar{\mathbf{b}}\} = n^{-1}(n-1)^{-1} \sum_{i \in A} (\mathbf{b}_i - \bar{\mathbf{b}})'(\mathbf{b}_i - \bar{\mathbf{b}}).$$

The variance estimator is biased for the sample covariances, but is judged an adequate approximation in large samples. Then the estimated covariance matrix of

$$\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_x^2, \hat{\sigma}_{a,e}^2, \hat{\sigma}_{a,u}^2)'$$

is

$$\hat{V}\{\hat{\boldsymbol{\theta}}\} = \hat{\mathbf{H}}\hat{\mathbf{V}}\{\bar{\mathbf{b}}\}\hat{\mathbf{H}}',$$

where $\hat{\mathbf{H}}$ is the estimator of the partial derivative of $\hat{\boldsymbol{\theta}}$ with respect to $\bar{\mathbf{b}}$ evaluated at $\hat{\boldsymbol{\theta}}$. Show that the rows of $\hat{\mathbf{H}}$ are

$$\frac{\partial \hat{\theta}_1}{\partial \bar{\mathbf{b}}} = (1, -\hat{\beta}_1, 0, \bar{Z}_2 m_{12} \hat{\beta}_1, -\bar{Z}_2 m_{12}^{-1}, 0.5\bar{Z}_2 m_{12}^{-1}),$$

$$\frac{\partial \hat{\theta}_2}{\partial \bar{\mathbf{b}}} = (0, 0, 0, -m_{12}^{-1}\hat{\beta}_1, m_{12}^{-1}, -0.5m_{12}^{-1}),$$

$$\frac{\partial \hat{\theta}_3}{\partial \bar{\mathbf{b}}} = (0, 0, 0, 2m_{12}\hat{\sigma}_y^{-2}, -\hat{\sigma}_y^{-2}\hat{\sigma}_x^2, 0.5\hat{\sigma}_y^{-2}\hat{\sigma}_x^2),$$

$$\frac{\partial \hat{\theta}_4}{\partial \bar{\mathbf{b}}} = (0, 0, 0, 0, 0, 1),$$

$$\frac{\partial \hat{\theta}_5}{\partial \bar{\mathbf{b}}} = (0, 0, 1, -2m_{12}\hat{\sigma}_y^{-2}, \hat{\sigma}_y^{-2}\hat{\sigma}_x^2, -0.5\hat{\sigma}_y^2\hat{\sigma}_x^2),$$

where $\hat{\sigma}_y^2 = \hat{\beta}_1^2 \hat{\sigma}_x^2$. The estimated covariance matrix of $(m_{11}, m_{12}, m_{22}, m_{33})$ is

$$\begin{pmatrix} 3.4025 & 2.1599 & 1.4496 & 0 \\ 2.1599 & 1.6520 & 1.3556 & 0 \\ 1.4496 & 1.3556 & 1.4412 & 0 \\ 0 & 0 & 0 & 1.9254 \end{pmatrix} \times 10^{-6}$$

and the covariance matrix of $(\bar{Z}_1, \bar{Z}_2)$ is

$$\begin{pmatrix} 53.960 & 46.031 \\ 46.031 & 49.038 \end{pmatrix} \times 10^{-6}.$$

Compute the Taylor estimated variance of $\hat{\theta}$.

16. (Section 5.5) Let $a_g \sim NI(0, \sigma_u^2 + \sigma_{eg}^2)$ and let a sample $a_1, a_2, \ldots, a_m$ be given. Define an estimator of σ_u^2 to be the solution to the equation

$$m^{-1} \sum_{g=1}^{m} (\hat{\sigma}_u^2 + \sigma_{eg}^2)^{-1} a_g^2 = 1.00,$$

where the σ_{eg}^2, $0 < \sigma_{eg}^2 < C_\sigma$, $g = 1, 2, \ldots, m$, are known and C_σ is a positive constant. Let $\hat{\sigma}_u^2 = 0$ if $\Sigma_{g=1}^m \sigma_{eg}^{-2} a_g^2 < m$. Assume that $\sigma_u^2 > 0$ and that

$$\lim_{m \to \infty} m^{-1} \sum_{g=1}^{m} (\sigma_u^2 + \sigma_{eg}^2)^{-1} = \Phi,$$

where Φ is a positive constant. Obtain the limiting distribution of $\hat{\sigma}_u^2$ as $m \to \infty$.

Hint: The quantity $(\sigma_u^2 + \sigma_{eg}^2)^{-1} = \partial log(\sigma_u^2 + \sigma_{eg}^2)/\partial \sigma_u^2$. Also, $(\sigma_u^2 + \sigma_{eg}^2)^{-1}$ is monotone decreasing in σ_u^2 for positive σ_u^2. See Fuller (1996, Chapter 5) to prove that $\hat{\sigma}_u^2$ is consistent for σ_u^2.

This Page Intentionally Left Blank

CHAPTER 6

ANALYTIC STUDIES

6.1 INTRODUCTION

We introduced a model in which a finite population is generated from a superpopulation in Chapter 1. In Theorem 1.3.2 we considered an estimator for the superpopulation mean and for the finite population mean, but in subsequent chapters we concentrated on estimation for the finite population parameters.

The parameters estimated for the finite population are generally enumerative. For example, "How many people in the United States were unemployed on March 10, 2007?" or "What fraction of the persons employed in secondary education in Iowa in 2007 are female?" The questions pertain to a specific finite population and are *descriptive.* The questions asked about the infinite superpopulation are almost always *analytical.* For example, "If personal income (in the United States) increases 2%, how much will the consumption of beef increase?" or "Do people working in automobile production have a higher risk of lung cancer than production workers in farm tractor produc-

Sampling Statistics. By Wayne A. Fuller

tion?" Deming and Stephan (1941) described these two uses of survey data. See also Deming (1953).

The analytic questions often have a less well-specified population of interest than the descriptive questions. For example, with respect to beef consumption, is income to increase next year? or within two years? or is there no time limit on the applicability of the results? Ultimately, the user will judge whether or not the model assumptions are such that the results of the analysis are applicable in the user's situation of interest. Thus, the user must judge the degree to which the population sampled is similar to the population of interest.

To develop statistical estimators, the finite population is treated as a realization of some probabilistic mechanism, also called the *model*. The analysis is carried out under the assumptions of that model, and a part of the analysis involves checking those model assumptions for which checks are available, but not all model assumptions are subject to test. The sometimes implicit statement associated with the analysis is: "Given the model assumptions, it is estimated that... ."

Since the 1940s, statisticians have grappled with the problem of determining statistical procedures appropriate for analytic studies employing survey data. An early article on the use of census data for analytic purposes is that of Deming and Stephan (1941). Books containing discussions of the topic are Skinner, Holt, and Smith (1989), Korn and Graubard (1999), and Chambers and Skinner (2003). We concentrate on regression models and on procedures requiring modest model specification beyond that usually associated with regression models.

6.2 MODELS AND SIMPLE ESTIMATORS

For the formal study of analytic estimators, we require formal specification of the random process that is conceptually generating the finite population. It is reasonable to treat some design variables, such as stratification variables, as part of the vector generated by the random process if the design variables are assigned by the statistician on the basis of information about the finite population.

In one approach to estimating an analytic parameter, the finite population is considered to be a simple random sample from an infinite superpopulation. An equivalent description for this situation is the statement that the vector $(y_1, y_2, \ldots, y_N)$ is a realization of N independent identically distributed random variables. Then the finite population can be used to estimate the superpopulation parameter and inference proceeds in two steps; first a finite population quantity is estimated and then this estimate is used as an estimator of the parameter of interest.

To study the two-step approach, assume that the subject matter specialist has a model for a conceptual population with population parameter θ. Assume a sequence of finite populations that is a sequence of simple random samples from the conceptual population. Assume that a sample estimator $\hat{\theta}$ is unbiased, or nearly unbiased, for the finite population parameter θ_N with variance that is order n^{-1}. Also assume that $\hat{V}\{\hat{\theta} \mid \mathcal{F}\}$ is a nearly unbiased design-consistent variance estimator for $V\{\hat{\theta} \mid \mathcal{F}\}$. That is, assume that

$$E\{\hat{\theta} - \theta_N \mid \mathcal{F}_N\} = O_p(n^{-1}) \text{ a.s.}, \tag{6.2.1}$$

$$V\{\hat{\theta} \mid \mathcal{F}_N\} = O_p(n^{-1}) \text{ a.s.}, \tag{6.2.2}$$

$$E\{[V(\hat{\theta} \mid \mathcal{F}_N)]^{-2}[\hat{V}(\hat{\theta} \mid \mathcal{F}_N) - V(\hat{\theta} \mid \mathcal{F}_N)]^2 \mid \mathcal{F}_N\} = o_p(1) \text{ a.s.}, \tag{6.2.3}$$

$$E\{V(\hat{\theta} \mid \mathcal{F}_N)\} = O(n^{-1}), \tag{6.2.4}$$

and

$$V\{E(\hat{\theta} \mid \mathcal{F}_N)\} = V\{\theta_N\} + o(N^{-1}). \tag{6.2.5}$$

If the finite population is generated by a random process, one can usually define a function of the finite population, denoted by θ_N, that is an estimator of θ with the properties

$$E\{\theta_N\} = \theta + O(N^{-1}) \tag{6.2.6}$$

and

$$V\{\theta_N - \theta\} = O(N^{-1}). \tag{6.2.7}$$

Given (6.2.1) and (6.2.6), $\hat{\theta}$ is nearly unbiased for the superpopulation parameter θ,

$$E\{\hat{\theta}\} = E\{E(\hat{\theta} \mid \mathcal{F}_N)\} = \theta + O(n^{-1}) \tag{6.2.8}$$

Furthermore, given (6.2.1), (6.2.4), and (6.2.7), the variance of $\hat{\theta}$ as an estimator of θ is

$$V\{\hat{\theta} - \theta\} = E\{V(\hat{\theta} \mid \mathcal{F}_N)\} + V\{\theta_N\} + o(N^{-1}).$$

Thus, to estimate the variance of $\hat{\theta}$ as an estimator of θ, we can use

$$\hat{V}\{\hat{\theta} - \theta\} = \hat{V}\{\hat{\theta} \mid \mathcal{F}\} + \hat{V}\{\theta_N - \theta\}, \tag{6.2.9}$$

where $\hat{V}\{\theta_N - \theta\}$ is an estimator of $V\{\theta_N - \theta\}$.

We give some example model specifications.

Case 1. The finite population is a realization of $iid(\mu, \sigma_y^2)$ random variables, and a simple nonreplacement sample of size n is selected from the finite population. Because the simple random sample from the finite population is also a random sample from the infinite population, it follows that the sample mean is unbiased for the finite population mean and for the infinite population mean. See Theorems 1.3.1 and 1.3.2. Furthermore, from the results of Chapter 1,

$$E\{(\bar{y}_n - \bar{y}_N)^2 \mid \mathcal{F}\} = N^{-1}(N-n)n^{-1}S_{yN}^2 \tag{6.2.10}$$

and

$$E\{(\bar{y}_n - \mu)^2\} = n^{-1}\sigma_y^2, \tag{6.2.11}$$

where $y \sim (\mu, \sigma_y^2)$ in the infinite superpopulation. Estimators of the variances are

$$\hat{V}\{\bar{y}_n - \bar{y}_N \mid \mathcal{F}\} = N^{-1}(N-n)n^{-1}s_y^2 \tag{6.2.12}$$

and

$$\hat{V}\{\bar{y}_n - \mu\} = n^{-1}s_y^2, \tag{6.2.13}$$

where, as before,

$$s_y^2 = (n-1)^{-1}\sum_{i \in A}(y_i - \bar{y}_n)^2.$$

See Theorem 1.3.1. ■■

Case 1 is particularly simple because the estimator $\bar{y}_n$ is a "good" estimator for both parameters. If we add the assumption that the finite population is a realization of independent $N(\mu, \sigma_y^2)$ random variables, $\bar{y}_n$ is the best estimator, in that it minimizes the mean square error.

However, if one assumes, for example, that the finite population is a realization of independent lognormal random variables, there exist superior estimators for the mean under the model, because the maximum likelihood estimator of the mean of a sample from a lognormal distribution is not the simple mean.

Case 2. The sample is a simple nonreplacement sample of clusters from a finite population of clusters, where the finite population is a sample from

an infinite population of clusters. This is a modest extension of Case 1 obtained by replacing "elements" with "clusters of elements." The sample of clusters from the finite population of clusters is also a sample from the infinite population of clusters.

For a simple random sample of clusters selected from the finite population, let

$$\bar{y}_R = \left(\sum_{i\in A} M_i\right)^{-1} \sum_{i\in A}\sum_{j\in B_i} y_{ij}, \tag{6.2.14}$$

where M_i is the number of elements in cluster i, y_{ij} is the value for element j in cluster i, and B_i is the set of elements in cluster i. Then $\bar{y}_R$ is design consistent for the population element mean

$$\bar{y}_N = \left(\sum_{i\in U} M_i\right)^{-1} \sum_{i\in U}\sum_{j\in B_i} y_{ij}. \tag{6.2.15}$$

The extension from element sampling to cluster sampling may lead to considerable increase in the complexity of the analysis. Assume that the elements of the superpopulation satisfy the model

$$y_{ij} = \mu + \gamma_i + e_{ij}, \tag{6.2.16}$$

where the e_{ij} are $iid(0,\ \sigma_e^2)$, the γ_i are $iid(0,\ \sigma_\gamma^2)$, and e_{ij} is independent of γ_t for all ij and t. Assume that a random sample of m clusters is selected from the superpopulation to form the finite population, where the size of cluster i, denoted by M_i, is a random variable independent of e_j and γ_j for all i and j. Under the assumption that the M_i are independent of y_{ij}, the estimator (6.2.14) is consistent for μ, where $\mu = E\{y_{ij}\}$. However, if model (6.2.16) is true, the best linear unbiased estimator of μ is

$$\hat{\mu} = (\mathbf{J}_t'\boldsymbol{\Sigma}^{-1}\mathbf{J}_t)^{-1}\mathbf{J}_t'\boldsymbol{\Sigma}^{-1}\mathbf{Y}, \tag{6.2.17}$$

where

$$\mathbf{Y} = (y_{11},\ y_{12},\ \ldots,\ y_{1,M_1},\ y_{21},\ \ldots,\ y_{m1},\ \ldots,\ y_{m,M_m}),$$

$$\boldsymbol{\Sigma} = \text{blockdiag}(\boldsymbol{\Sigma}_{11},\ \boldsymbol{\Sigma}_{22},\ \ldots,\ \boldsymbol{\Sigma}_{mm}),$$

$$\boldsymbol{\Sigma}_{ii} = \mathbf{I}_{M_i}\sigma_e^2 + \mathbf{J}_{M_i}\mathbf{J}_{M_i}'\sigma_\gamma^2,$$

$\mathbf{J}_{M_i}$ is a column of M_i 1's and $\mathbf{J}_t$ is a column of $t = \boldsymbol{\Sigma}_{i\in A}M_i$ 1's. If $M_i = M$ for all i, estimator (6.2.17) reduces to (6.2.14). Also, the estimators are very similar for modest differences in cluster sizes. Practitioners will often choose

estimator (6.2.14) in such situations because it is design consistent for the finite population parameter. The assumption that M_i is independent of y_{ij} is required for estimator (6.2.17) to be consistent for μ. If, say, M_i is larger for clusters with large cluster means, the subject matter specialist must be sure that the μ of (6.2.16) is the parameter of interest. Note that the assumption that the cluster mean $\bar{y}_i$ and size M_i are uncorrelated can be tested using the sample correlation. ■■

Case 3. A stratified sample is selected from a finite population that is a realization of independent and identically distributed (μ, σ_y^2) random variables. The estimator

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h, \tag{6.2.18}$$

where $W_h = N^{-1}N_h$, is design unbiased for the finite population mean. Also, conditions (6.2.2) and (6.2.3) on the variance and estimated variance of the stratified mean are satisfied under the assumption of fixed W_h and finite population fourth moments. Thus, the variance of the stratified mean as an estimator of the superpopulation mean is

$$V\{\bar{y}_{st} - \mu\} = E\left\{\sum_{h=1}^{H} W_h^2 N_h^{-1}(N_h - n_h)n_h^{-1}S_h^2\right\} + N^{-1}\sigma_y^2. \tag{6.2.19}$$

An estimator of the population S_y^2 and hence of σ_y^2 is

$$\hat{S}_y^2 = \sum_{h=1}^{H} W_h n_h^{-1} \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_{st})^2. \tag{6.2.20}$$

It follows that an estimator of the variance of $\bar{y}_{st}$ as an estimator of the superpopulation mean μ is

$$\hat{V}_s\{\bar{y}_{st} - \mu\} = \sum_{h=1}^{H} W_h^2 N_h^{-1}(N_h - n_h)n_h^{-1}s_h^2 + N^{-1}\hat{S}_y^2. \tag{6.2.21}$$

Note the analogy between (6.2.21) and (3.3.11). The estimation of a superpopulation parameter using a sample from a finite population is analogous to estimating the mean of a finite population given a sample (phase 2 sample) selected from a (phase 1) sample. ■■

Case 4. To illustrate the importance of the assumptions made about the infinite population, consider a stratified sample and the model assumption that the superpopulation is a stratified population with the same fixed number of strata and the same relative stratum sizes as the finite population. Then the superpopulation mean is

$$\mu = \sum_{h=1}^{H} W_h \mu_h, \tag{6.2.22}$$

where the μ_h are the individual superpopulation stratum means. The finite population mean is

$$\bar{y}_N = \sum_{h=1}^{H} W_h \bar{y}_{hN}, \tag{6.2.23}$$

where $\bar{y}_{hN}$ is the finite population mean for stratum h. The variance of $\bar{y}_N$ as an estimator of μ is

$$V\{\bar{y}_N - \mu\} = \sum_{h=1}^{H} W_h^2 N_h^{-1} \sigma_h^2, \tag{6.2.24}$$

where the σ_h^2 are the individual superpopulation stratum variances. It follows that under the assumption of an infinite population with fixed strata, an estimator of the variance of $\bar{y}_{st}$ as an estimator of μ is

$$\hat{V}_{st}\{\bar{y}_{st} - \mu\} = \sum_{h=1}^{H} W_h^2 n_h^{-1} s_h^2. \tag{6.2.25}$$

If the stratified sample is a proportional stratified sample, the difference between the variance estimator (6.2.21) constructed under the assumption that the finite population is a simple random sample from a superpopulation and the variance estimator (6.2.25) constructed under the assumption of a stratified superpopulation is

$$\hat{V}_s\{\bar{y}_{st} - \mu\} - \hat{V}_{st}\{\bar{y}_{st} - \mu\} = N^{-1} \sum_{h=1}^{H} W_h^2 \left[(1 - n_h^{-1} - W_h) s_h^2 + (\bar{y}_h - \bar{y}_{st})^2\right].$$

The difference in the variances, and hence in the variance estimators, is due to the assumption made about the infinite population. One making the second assumption (6.2.22) is specifying that inferences are for infinite populations

with the same stratification structure as that of the finite population. In most situations, stratification is effective and (6.2.25) is smaller than (6.2.21). Of course, the difference goes to zero as the sampling rates go to zero. ■■

Case 5. The sample is a stratified sample from a superpopulation with an infinite number of strata. To create the finite population, a sample of H strata is selected from a population of strata. Then a sample of N_h, $h = 1, 2, \ldots, H$, elements is selected from the selected strata. It is assumed that the stratum means of the population of strata are iid random variables,

$$\mu_h \sim iid(\mu, \sigma_\alpha^2). \tag{6.2.26}$$

Conditional on the strata selected, denoted by $\mathbf{h}$, and the stratum fractions $\mathbf{W} = (W_1, W_2, \ldots, W_h)$,

$$E\{\bar{y}_{st} \mid (\mathbf{h}, \mathbf{W})\} = \sum_{h=1}^{H} W_h \mu_h. \tag{6.2.27}$$

Under assumption (6.2.26), the expectation of (6.2.27) is μ because the stratum means are assumed to be independent and identically distributed. However, given assumption (6.2.26), we can use additional assumptions to construct a superior estimator. The standard nested-error model adds the assumption of common stratum variances so that the observations satisfy

$$y_{hj} = \mu + \alpha_h + e_{hj}, \tag{6.2.28}$$

where $\alpha_h = \mu_h - \mu$, the e_{hi} are $iid(0, \sigma^2)$, the α_h are $iid(0, \sigma_\alpha^2)$, and e_{hj} is independent of α_r for all hj and r. Often, α_h and e_{hj} are assumed to be normally distributed, but the assumption of fourth moments is adequate for most purposes. Under model (6.2.28) with known σ_α^2 and σ^2, the best linear unbiased estimator of μ is

$$\hat{\mu} = (\mathbf{J}_n' \mathbf{V}^{-1} \mathbf{J}_n)^{-1} \mathbf{J}_n' \mathbf{V}^{-1} \mathbf{y}, \tag{6.2.29}$$

where $\mathbf{V} = \text{blockdiag}(\mathbf{V}_{11}, \ldots, \mathbf{V}_{hh}, \ldots, \mathbf{V}_{HH})$,

$$\mathbf{V}_{hh} = \mathbf{I}_{n_h} \sigma^2 + \mathbf{J}_{n_h} \mathbf{J}_{n_h}' \sigma_\alpha^2,$$

$\mathbf{J}_r$ is a column vector of r 1's, and $\mathbf{y}$ is the column vector of n observations. Simple estimators of σ^2 and σ_α^2 are

$$\hat{\sigma}^2 = (n - H)^{-1} \sum_{h=1}^{H} \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 \tag{6.2.30}$$

and

$$\hat{\sigma}_{\alpha}^2 = \max\left((H-1)^{-1}\sum_{h=1}^{H}\left[(\bar{y}_h - \bar{y}_{\cdot})^2 - n_h^{-1}\hat{\sigma}^2\right], 0\right), \quad (6.2.31)$$

where

$$\bar{y}_{\cdot} = H^{-1}\sum_{h=1}^{H}\bar{y}_h.$$

For alternative estimators, see texts on linear models such as Searle (1971). As with the analogous model for clusters, the use of estimator (6.2.29) requires the assumption that stratum sizes and stratum means are independent.

Very often the analyst will formulate a model such as (6.2.28) without consideration of the sample design. Our model (6.2.28) assumed that the random effects were sampling strata, but it is possible to define random effects with a different relationship to the design variables. See Pfeffermann et al. (1998). ■■

These examples demonstrate how the analysis depends on the superpopulation model. If all sampling rates are small, there are models where the design variance of the estimator of the finite population parameter is nearly equal to the variance of the estimator as an estimator of the superpopulation parameter. See Cases 1 and 3. Furthermore, if the postulated structure of the superpopulation is the structure of the sample, the estimated variance of the estimator for the superpopulation parameter is the estimator of the design variance with the finite population correction omitted. See Case 4. In other situations, such as Case 5, the superpopulation model leads to a different estimation procedure than that appropriate for a finite population parameter.

6.3 ESTIMATION OF REGRESSION COEFFICIENTS

6.3.1 Ordinary least squares and tests for bias

We now concentrate our discussion of model parameter estimation on regression estimation. Assume that the subject matter analyst specifies a regression model relating y_i to $\mathbf{x}_i$ as

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i, \quad (6.3.1)$$

where the e_i are independent $(0, \sigma^2)$ random variables independent of $\mathbf{x}_j$ for all i and j. Assume that the finite population can be treated as a set of vectors

satisfying (6.3.1). Then the model for the finite population can be written as

$$\begin{aligned} \mathbf{y}_N &= \mathbf{X}_N\boldsymbol{\beta} + \mathbf{e}_N, \\ \mathbf{e}_N &\sim (\mathbf{0},\ \mathbf{I}_N\sigma^2), \end{aligned} \tag{6.3.2}$$

where $\mathbf{y}_N$ is the N-dimensional vector of values for the dependent variable, $\mathbf{X}_N$ is the $N \times k$ matrix of values of the explanatory variables, and the error vector $\mathbf{e}_N$ is a vector of independent random variables independent of $\mathbf{X}_N$. We discuss the linear model (6.3.2) in this section, but the mean of y_i could be given by a function $g(\mathbf{x}_i,\ \boldsymbol{\beta})$, and the covariance matrix might be specified in a more general form. See Section 6.4.

Assume that a simple random sample of size n is selected from the finite population. Then the model structure also holds for the sample and we can write

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \\ \mathbf{e} &\sim (\mathbf{0},\ \mathbf{I}\sigma^2), \end{aligned} \tag{6.3.3}$$

and $\mathbf{e}$ is independent of $\mathbf{X}$, where $\mathbf{y}$ is the n-dimensional column vector of observations on y and $\mathbf{X}$ is the $n \times k$ matrix of observations on the explanatory variables. On the basis of the model, conditional on $\mathbf{X}$, the best linear unbiased estimator of $\boldsymbol{\beta}$ is the ordinary least squares estimator

$$\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{6.3.4}$$

with conditional variance

$$V\{\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}\} = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2. \tag{6.3.5}$$

Consider a sample selected with unequal probabilities π_i. If e_i is independent of π_i, the ordinary least squares estimator (6.3.4) remains unbiased with conditional variance (6.3.5). If π_i is correlated with e_i, the ordinary least squares estimator (6.3.4) is biased. For example, if π_i and e_i are positively correlated, large values of e_i have a greater probability of appearing in the sample and $E\{e_i \mid i \in A\}$ is positive.

If we assume that $\hat{\boldsymbol{\beta}}_{ols}$ has the requisite moments so that

$$E\{\hat{\boldsymbol{\beta}}_{ols} \mid \mathcal{F}_N\} = (\mathbf{X}_N'\mathbf{D}_{\pi,N}\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{D}_{\pi,N}\mathbf{y}_N + O(n^{-1}) \quad \text{a.s.,} \tag{6.3.6}$$

where $\mathbf{X}_N' = (\mathbf{x}_1',\ \mathbf{x}_2',\ \ldots,\ \mathbf{x}_N')$ and $\mathbf{D}_{\pi,N} = \text{diag}(\pi_1,\ \pi_2,\ \ldots,\ \pi_N)$. See (2.2.41) of Section 2.2. It follows that

$$E\{\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}\} = E\{(\mathbf{X}_N'\mathbf{D}_{\pi,N}\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{D}_{\pi,N}\mathbf{e}_N\} + O(n^{-1}), \tag{6.3.7}$$

where $\mathbf{e}_N = (e_1, e_2, \ldots, e_N)$ and the leading term in the bias of $\hat{\boldsymbol{\beta}}_{ols}$ as an estimator of $\boldsymbol{\beta}$ is a function of $E\{\mathbf{x}_i'\pi_i e_i\}$. For the simple model with $x_i \equiv 1$ and $\hat{\beta}_{ols} = \bar{y}_n$, the bias in $\bar{y}_n$ as an estimator of the superpopulation mean is

$$E\{\bar{y}_n - \mu\} = n^{-1}E\left\{\sum_{i\in U}\pi_i e_i\right\} + O(n^{-1}). \tag{6.3.8}$$

The probability weighted estimator of $\boldsymbol{\beta}$ for model (6.3.3) is

$$\hat{\boldsymbol{\beta}}_\pi = (\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{y}, \tag{6.3.9}$$

and by Theorem 2.2.1, $\hat{\boldsymbol{\beta}}_\pi$ is design consistent for

$$\boldsymbol{\beta}_N = (\mathbf{X}_N'\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{y}_N.$$

Under model (6.3.2), $E\{\boldsymbol{\beta}_N\} = \boldsymbol{\beta}$ and $V\{\boldsymbol{\beta}_N - \boldsymbol{\beta}\} = O(N^{-1})$. Therefore, $\hat{\boldsymbol{\beta}}_\pi$ is a consistent estimator of $\boldsymbol{\beta}$.

If $\mathbf{e}$ is independent of $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$, the conditional covariance matrix of $\hat{\boldsymbol{\beta}}_\pi - \boldsymbol{\beta}$ under model (6.3.3), conditional on $\mathbf{X}$, is

$$V\{\hat{\boldsymbol{\beta}}_\pi \mid \mathbf{X}\} = (\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{D}_\pi^{-1}\mathbf{X}(\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\sigma^2. \tag{6.3.10}$$

In most cases the variances in (6.3.10) are larger than the corresponding elements of (6.3.5), and one may be tempted to use the ordinary least squares estimator. However, if ordinary least squares is to be used, one should be comfortable with model (6.3.3) and the condition that $E\{\mathbf{x}_i'e_i \mid i \in A\} = \mathbf{0}$. Fortunately, it is possible to test for the condition.

A test of the hypothesis that the expected value of the coefficient in (6.3.4) is equal to the expected value of the coefficient in (6.3.9) can be performed as a test on coefficients in an expanded multiple regression model. The null hypothesis is

$$E\left\{(\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right\} = \mathbf{0}. \tag{6.3.11}$$

Multiplying (6.3.11) by $\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X}$, the hypothesis becomes

$$E\left\{\mathbf{X}'\mathbf{D}_\pi^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}\right\} = \mathbf{0}.$$

Recall that $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ is an idempotent matrix and the regression coefficient for $\mathbf{Z} = \mathbf{D}_\pi^{-1}\mathbf{X}$ in the multiple regression of $\mathbf{y}$ on $(\mathbf{X}, \mathbf{Z})$ is the regression coefficient for the regression of

$$(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \text{ on } (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{D}_\pi^{-1}\mathbf{X}.$$

Hence, if

$$E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\} = E\{(\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{y}\},$$

the expected value of the coefficient of $\mathbf{Z} = \mathbf{D}_\pi^{-1}\mathbf{X}$ in the multiple regression of $\mathbf{y}$ on $(\mathbf{X}, \mathbf{Z})$ is zero. That is, given hypothesis (6.3.11), $\gamma = \mathbf{0}$ in the expanded model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{a}, \tag{6.3.12}$$

where $\mathbf{a}$ is the error in the expanded model and $\mathbf{a} = \mathbf{e}$ under the null hypothesis.

Two options are available for testing $\gamma = \mathbf{0}$, the usual ordinary least squares test based on the null model with error assumptions of (6.3.3), and a test based on the sample design. See DuMouchel and Duncan (1983) and Fuller (1984). The use of the ordinary least squares test requires the assumption that the e_i are uncorrelated with common variance. If the e_i are correlated, as in a cluster design, or have unequal variances, the test based on the design is preferable.

One can compute a test using a subset of the x-variables if the dimension of x is large. In practice, the intercept is often the most biased of the coefficients. Hence, computing the test using a subset containing the intercept or adding only w_i to the regression can increase the power of the test. See Das Gupta and Perlman (1974). Estimation under the hypothesis that only some coefficients are biased is considered in Section 6.4.4.

If the test indicates that the probability-weighted estimator differs from the least squares estimator, the analyst must answer the question: Why? The usual first response will be a search for subject matter variables to add to the model. If the inclusion of such variables results in a nonsignificant test statistic, the expanded model can be accepted.

Example 6.3.1. The data in Table 6.1 are artificial data, generated to mimic the properties of data studied by Korn and Graubard (1999). In that study, y is gestational age of babies and x is birthweight. The sample in Table 6.1 is a stratified sample with the relative weights given in the second column. Age as a function of birthweight is of interest. Our initial model postulates age to be a linear function of birthweight. The ordinary least squares regression of age on birthweight (y on x) gives

$$\begin{aligned} \hat{y} &= 25.765 + 0.389x, \\ &\quad (0.370)\ (0.012) \end{aligned}$$

where $s^2 = 1.359$ and the ordinary least squares standard errors are in parentheses. The weighted estimator (6.3.9) gives

$$\begin{aligned} \hat{y} &= 28.974 + 0.297x, \\ &\quad (0.535)\ (0.016) \end{aligned}$$

Table 6.1 Birthweight and Age Data

Stratum	Weight	Birthwgt.	Gest. Age	Stratum	Weight	Birthwgt.	Gest. Age
1	0.05	10.5	29.4	10	0.84	31.9	38.3
	0.05	07.7	26.1		0.84	30.4	37.8
	0.05	07.9	26.4		0.84	32.0	39.6
	0.05	09.8	27.6		0.84	30.6	39.1
	0.05	09.1	28.2		0.84	30.1	39.0
2	0.06	11.5	29.1	11	0.90	33.5	38.8
	0.06	13.3	30.6		0.90	33.1	38.7
	0.06	12.0	30.4		0.90	33.6	40.4
	0.06	12.1	28.1		0.90	32.7	38.2
	0.06	11.6	29.9		0.90	32.5	38.8
3	0.07	15.9	31.3	12	0.95	33.9	38.9
	0.07	15.3	30.9		0.95	33.8	38.8
	0.07	16.1	31.7		0.95	34.8	39.1
	0.07	13.8	30.7		0.95	34.6	39.6
	0.07	15.1	31.2		0.95	34.3	39.3
4	0.10	19.0	34.2	13	1.00	35.4	38.9
	0.10	17.1	31.9		1.00	35.6	40.3
	0.10	19.6	32.9		1.00	34.9	40.0
	0.10	19.7	33.7		1.00	35.1	39.3
	0.10	19.1	32.9		1.00	36.3	39.5
5	0.20	21.1	33.9	14	1.00	37.1	40.3
	0.20	21.4	36.1		1.00	36.5	41.4
	0.20	22.1	36.0		1.00	37.5	40.3
	0.20	20.9	35.1		1.00	37.3	40.3
	0.20	22.0	34.9		1.00	36.5	40.4
6	0.22	23.1	34.5	15	1.00	39.1	41.6
	0.22	24.0	36.3		1.00	38.7	39.0
	0.22	22.5	36.5		1.00	38.1	39.5
	0.22	23.0	34.2		1.00	39.3	41.1
	0.22	23.1	36.4		1.00	38.5	40.7
7	0.80	25.1	35.3	16	1.00	40.3	40.2
	0.80	24.4	36.8		1.00	39.8	40.4
	0.80	25.9	37.5		1.00	39.9	40.7
	0.80	25.0	36.1		1.00	41.0	39.7
	0.80	25.1	36.1		1.00	39.5	39.5
8	0.80	27.9	38.0	17	1.00	41.4	40.2
	0.80	27.2	36.8		1.00	41.7	42.8
	0.80	26.1	37.0		1.00	42.4	41.1
	0.80	28.0	38.0		1.00	42.0	40.8
	0.80	27.8	36.6		1.00	41.2	42.7
9	0.82	29.2	37.4	18	1.00	43.2	40.0
	0.82	29.6	39.0		1.00	43.4	41.8
	0.82	29.6	39.6		1.00	42.6	41.0
	0.82	29.5	38.9		1.00	44.1	41.7
	0.82	28.6	38.7		1.00	43.5	40.9

where the standard errors in parentheses are obtained from

$$\hat{V}\{\hat{\boldsymbol{\beta}}_\pi\} = (\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\hat{\mathbf{D}}_{ee}\mathbf{D}_\pi^{-1}\mathbf{X}(\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1} \quad (6.3.13)$$

with $\hat{\mathbf{D}}_{ee} = \text{diag}(\hat{e}_1^2, \hat{e}_2^2, \ldots, \hat{e}_n^2)$ and $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_\pi$. The estimator (6.3.13) is the estimator $\hat{V}\{\hat{\boldsymbol{\beta}}\}$ in (2.2.39) with $\boldsymbol{\Phi} = \mathbf{D}_\pi$. Note that the variance estimator is appropriate for heterogeneous error variances.

The weighted estimate of the slope differs from the ordinary least squares estimate by several standard errors. Also, the calculated standard errors for the weighted procedure are much larger than those calculated for ordinary least squares. To construct a formal test of the hypothesis that the two procedures are estimating the same coefficients, we regress, using ordinary least squares, y on $(1, x, w, wx)$, where $w_i = \pi_i^{-1}$, to obtain

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0, \hat{\gamma}_1) &= (22.088, 0.583, 8.287, -0.326). \\ &\quad (0.532)(0.033)(0.861)\ (0.332)\end{aligned}$$

The residual mean square is 0.606 and the F test for the hypothesis that $(\gamma_1, \gamma_2) = (0, 0)$ is $F(2, 86) = 55.59$. The tabular 5% point for the F is 3.10 and the hypothesis is clearly rejected.

If the original model of a linear relationship is estimated, the ordinary least squares estimates are severely biased. However, the results lead us to ask if the model linear in birthweight is the appropriate model. A plot of the original data, or of the stratum means, shows a curvilinear relationship. Therefore, we fit the quadratic function by ordinary least squares to obtain

$$\begin{aligned}\hat{y} &= 28.355 + 0.331x - 0.887x_2, \\ &\quad (0.343)\ (0.010)\quad (0.082)\end{aligned}$$

where $x_2 = 0.01(x - 30)^2$. The residual mean square is 0.590. The quadratic function estimated by the weighted least squares of (6.3.9) is

$$\begin{aligned}\tilde{y} &= 28.458 + 0.327x - 0.864x_2. \\ &\quad (0.386)\ (0.011)\quad (0.108)\end{aligned}$$

The two estimated functions now appear similar. The ordinary least squares regression coefficients for the regression of y on $[1, x, x_2, w - \bar{w}, (w - \bar{w})(x - \bar{x}), (w - \bar{w})x_2]$ gives

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= (27.044,\ 0.384,\ -0.494) \\ &\quad (2.045)\ (0.019)\ (0.344)\end{aligned}$$

and

$$\begin{aligned}(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2) &= (-1.376, -0.203, -0.390). \\ &\quad (1.926)\ (0.178)\ (0.728)\end{aligned}$$

We define the added variables in terms of deviations from the means to improve numerical calculations and to make coefficients somewhat easier to interpret. The F test of the hypothesis that $(\gamma_0, \gamma_1, \gamma_2) = 0$, with 3 and 84 degrees of freedom, is 0.49. Hence, one easily accepts the null hypothesis that the two estimators of $(\beta_0,\ \beta_1,\ \beta_2)$ are estimating the same quantity. Although the two procedures may have the same expected value, the standard errors of the weighted procedure are larger than those of ordinary least squares. A plot of the residuals gives no reason to reject the model of homogeneous error variances. Therefore, we are comfortable with the quadratic model and the ordinary least squares estimates. ■■

6.3.2 Consistent estimators

It is quite possible that no variables leading to a nonsignificant test statistic for the hypothesis that $E\{\mathbf{x}_i\pi_i e_i\} = \mathbf{0}$ can be identified. If $E\{\mathbf{x}_i\pi_i e_i\} \neq \mathbf{0}$, it is sometimes said that the design is *informative* for the model. In such cases it becomes necessary to incorporate the sampling weights into the analysis. One approach is to add a specification for the weights themselves. See Pfeffermann (1993) and Pfeffermann and Sverchkov (1999).

Another approach is to specify an extended model that includes design variables. The extended model is then estimated and the subject matter parameters evaluated by taking an expectation with respect to the design variables. See Skinner (1994) and Holt, Smith, and Winter (1980).

It is possible that the weights have no subject matter content *per se* but are deemed by the analyst to be an artifact created by the sampler. Because the sampler has knowledge of the population, the probabilities are related to the errors in the model in a way that cannot be captured completely by the subject matter variables in the model. The probability weighted estimator (6.3.9) is consistent for $\boldsymbol{\beta}$ but can be inefficient if the diagonal matrix of weights differs from the inverse of the model covariance matrix.

In some situations, the properties of the population model can be used to construct consistent estimators that are more efficient under the model than the probability weighted estimator. Our objective is to obtain estimators for the subject matter coefficients that are not biased by the selection probabilities, but with a minimum of model specification for the weights themselves.

Assume that the finite population is a realization of a random process such that

$$\mathbf{y}_N = \mathbf{X}_N\boldsymbol{\beta} + \mathbf{e}_N \tag{6.3.14}$$

and

$$E\{\mathbf{e}_N \mid \mathbf{X}_N\} = \mathbf{0}.$$

Consider an estimator of the form

$$\hat{\boldsymbol{\beta}}_{W\Psi} = (\mathbf{X}'\mathbf{D}_\pi^{-1}\boldsymbol{\Psi}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\boldsymbol{\Psi}\mathbf{y}, \tag{6.3.15}$$

where $\boldsymbol{\Psi}$ is a diagonal matrix with diagonal elements ψ_i, and ψ_i is defined for all $i \in U$. Assume that

$$E\{\mathbf{e}_N \mid \boldsymbol{\Psi}_N, \mathbf{X}_N\} = \mathbf{0},$$

where $\boldsymbol{\Psi}_N$ is the $N \times N$ population matrix with diagonal elements ψ_i. Then, given regularity conditions,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{W\Psi} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{D}_\pi^{-1}\boldsymbol{\Psi}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\boldsymbol{\Psi}\mathbf{e} \\ &= \mathbf{X}_N'\boldsymbol{\Psi}_N\mathbf{X}_N)^{-1}\mathbf{X}_N'\boldsymbol{\Psi}_N\mathbf{e}_N + O_p(n^{-1/2}) \end{aligned} \tag{6.3.16}$$

and $\hat{\boldsymbol{\beta}}_{W\Psi}$ is consistent for $\boldsymbol{\beta}$ because, by assumption,

$$E\left\{\mathbf{X}_N'\boldsymbol{\Psi}_N\mathbf{e}_N\right\} = \mathbf{0}. \tag{6.3.17}$$

We give the limiting distribution of estimator (6.3.15) in Theorem 6.3.1. There are a number of assumptions that can be summarized in two basic conditions. First, the finite population is a sample from a superpopulation for which a central limit theorem holds and with enough moments for consistent variance estimation. Second, the sample design is such that a central limit theorem holds for almost all sequences of finite populations and such that a design consistent variance estimator is available. The assumption that the e_i are independent can be relaxed and that assumption does not preclude correlation in the finite population. For example, the finite population may contain clusters of related individuals.

Theorem 6.3.1. Let $\{(y_i,\ \mathbf{x}_i,\ \psi_i,\ \sigma_i^2)\}$ be a sequence of independent random vectors of dimension $k+4$ with bounded eighth moments, where y_i is related to $\mathbf{x}_i$ through the model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i,$$

$$e_i \sim ind(0, \sigma_i^2),$$

and $E\{e_i \mid \mathbf{x}_i, \psi_i\} = 0$ for all i. Let $\{\mathcal{F}_N\}, N = k+4,\ k+5,\ \ldots,$ be a sequence of finite populations, where $\mathcal{F}_N$ is composed of the first N elements of $\{(y_i,\ \mathbf{x}_i,\ \psi_i,\ \sigma_i^2)\}$. Let $\mathbf{z}_i = (y_i,\ \mathbf{x}_i)$, let

$$(\mathbf{M}_{Z\Psi Z,N}, \mathbf{M}_{ZZ,N}) = N^{-1}(\mathbf{Z}_N'\boldsymbol{\Psi}_N\mathbf{Z}_N,\ \mathbf{Z}_N'\mathbf{Z}_N), \tag{6.3.18}$$

and let

$$\hat{\mathbf{M}}_{Z\Psi Z} = N^{-1}\mathbf{Z}'\mathbf{D}_{\pi}^{-1}\boldsymbol{\Psi}\mathbf{Z}, \tag{6.3.19}$$

where $\mathbf{Z}$ is the $n_N \times (k+2)$ matrix of sample observations on $\mathbf{z}_i$ and $\mathbf{Z}_N$ is the $N \times (k+2)$ matrix of population values of $\mathbf{z}_i$. Assume that:

(i) The ψ_i satisfy $0 < \psi_i < K_h$ for some positive K_h and all i.

(ii) The matrices of (6.3.18) satisfy

$$\lim_{N\to\infty} (\mathbf{M}_{Z\Psi Z,N},\ \mathbf{M}_{ZZ,N}) = (\mathbf{M}_{Z\Psi Z}, \mathbf{M}_{ZZ}) \quad \text{a.s.}, \tag{6.3.20}$$

where $\mathbf{M}_{ZZ}$ and $\mathbf{M}_{Z\Psi Z}$ are positive definite.

(iii) The sequence of sample designs is such that for any $\mathbf{z}$ with bounded fourth moments,

$$\lim_{N\to\infty} n_N V\{\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N \mid \mathcal{F}_N\} = \mathbf{V}_{\infty,\bar{z}\bar{z}} \quad \text{a.s.}, \tag{6.3.21}$$

where $\bar{\mathbf{z}}_{HT} = N^{-1}\Sigma_{i\in A}\pi_i^{-1}\mathbf{z}_i$, and $\bar{\mathbf{z}}_N$ is the finite population mean of $\mathbf{z}$.

(iv) The variance $V\{\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N \mid \mathcal{F}_N\}$ is positive definite almost surely, and

$$[V\{\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N \mid \mathcal{F}_N\}]^{-1/2}(\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}) \quad \text{a.s.} \tag{6.3.22}$$

(v) The sampling rates satisfy

$$\lim_{N\to\infty} N^{-1}n_N = f_\infty$$

almost surely, where $0 \le f_\infty < 1$ and n_N is the expected sample size for a sample selected from the Nth population.

(vi) The estimator $\hat{V}\{\bar{\mathbf{z}}_{HT} \mid \mathcal{F}_N\}$ is a quadratic estimator of $V\{\bar{\mathbf{z}}_{HT} \mid \mathcal{F}_N\}$ such that

$$\hat{V}\{\bar{\mathbf{z}}_{HT} \mid \mathcal{F}_N\} - V\{\bar{\mathbf{z}}_{HT} \mid \mathcal{F}_N\} = o_p(n_N^{-1}) \tag{6.3.23}$$

for any $\mathbf{z}$ with bounded fourth moments.

(vii) The variances of $\mathbf{b}_i = \mathbf{x}_i'\psi_i e_i$ satisfy

$$\lim_{N\to\infty} N^{-1}\sum_{i\in U} \boldsymbol{\Sigma}_{bb,ii} = \boldsymbol{\Sigma}_{bb} \quad \text{a.s.},$$

where $\mathbf{\Sigma}_{bb,ii} = E\{\mathbf{b}_i\mathbf{b}_i' \mid \mathbf{x}_i, \psi_i\}$.

Let $\hat{\mathbf{\Sigma}}_{bb}$ be an estimator of $\mathbf{\Sigma}_{bb}$, where $\hat{\mathbf{\Sigma}}_{bb} = \mathbf{\Sigma}_{bb,N} + o_p(1)$ and $\mathbf{\Sigma}_{bb,N} = N^{-1}\Sigma_{i=1}^{N}\mathbf{\Sigma}_{bb,ii}$. Let $\hat{\boldsymbol{\beta}}_{W\Psi}$ be defined by (6.3.15).

Then

$$[\hat{V}\{\hat{\boldsymbol{\beta}}_{W\Psi}\}]^{-1/2}(\hat{\boldsymbol{\beta}}_{W\Psi} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{I}) \tag{6.3.24}$$

as $N \to \infty$, where

$$\hat{V}\{\hat{\boldsymbol{\beta}}_{W\Psi}\} \quad = \quad \hat{\mathbf{M}}_{X\Psi X}^{-1}[\hat{V}\{\bar{\mathbf{b}}_{HT} \mid \mathcal{F}_N\} + N^{-1}\hat{\mathbf{\Sigma}}_{bb}]\hat{\mathbf{M}}_{X\Psi X}^{-1}, \tag{6.3.25}$$

$\hat{V}\{\bar{\mathbf{b}}_{HT} \mid \mathcal{F}_N\}$ is the estimator of the design variance of $\bar{\mathbf{b}}_{HT}$ calculated with $\hat{\mathbf{b}}_i = \mathbf{x}_i'\psi_i\hat{e}_i$, and $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{W\Psi}$.

Proof. By the design and moment assumptions,

$$\hat{\mathbf{M}}_{X\Psi X}^{-1} - \mathbf{M}_{X\Psi X,N}^{-1} \quad = \quad O_p(n_N^{-1/2}). \tag{6.3.26}$$

By the assumption that $E\{\psi_i e_i \mid x_i\} = 0$,

$$E\{\mathbf{M}_{X\Psi e,N}\} \quad = \quad \mathbf{0}, \tag{6.3.27}$$

and by the moment assumptions, $\hat{\mathbf{M}}_{X\Psi e} = O_p(n_N^{-1/2})$. Thus,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{W\Psi} - \boldsymbol{\beta} \quad &= \quad \hat{\mathbf{M}}_{X\Psi X}^{-1}\hat{\mathbf{M}}_{X\Psi e} \\ &=: \quad \mathbf{M}_{X\Psi X}^{-1}\bar{\mathbf{b}}_{HT} + O_p(n_N^{-1}), \end{aligned} \tag{6.3.28}$$

where $(\hat{\mathbf{M}}_{X\Psi e}, \mathbf{M}_{X\Psi e}) =: (\bar{\mathbf{b}}_{HT}, \bar{\mathbf{b}}_N)$. Similarly,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{W\Psi} - \boldsymbol{\beta}_N \quad &= \quad \mathbf{M}_{X\Psi X,N}^{-1}(\bar{\mathbf{b}}_{HT} - \bar{\mathbf{b}}_N) + O_p(n_N^{-1}) \\ &= \quad \mathbf{M}_{X\Psi X}^{-1}(\bar{\mathbf{b}}_{HT} - \bar{\mathbf{b}}_N) + O_p(n_N^{-1}). \end{aligned} \tag{6.3.29}$$

By (6.3.21), and (6.3.22),

$$n_N^{1/2}(\bar{\mathbf{b}}_{HT} - \bar{\mathbf{b}}_N) \mid \mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{V}_{\infty,\bar{b}\bar{b}}) \quad \text{a.s.}, \tag{6.3.30}$$

where $\mathbf{V}_{\infty,\bar{b}\bar{b}}$ is well defined by (6.3.21). By the moment assumptions,

$$N^{1/2}\bar{\mathbf{b}}_N \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{\Sigma}_{bb}). \tag{6.3.31}$$

It follows by Theorem 1.3.6 that

$$n_N^{1/2}\bar{\mathbf{b}}_{HT} \xrightarrow{\mathcal{L}} N(\mathbf{0},\ \mathbf{V}_{\infty,\bar{b}\bar{b}} + f_\infty\mathbf{\Sigma}_{bb}). \tag{6.3.32}$$

By (6.3.23), the estimator of variance

$$\hat{V}\{\bar{\mathbf{b}}_{HT} \mid \mathcal{F}_N\} = V\{\bar{\mathbf{b}}_{HT} \mid \mathcal{F}_N\} + o_p(n_N^{-1}),$$

and by the arguments in the proof of Theorem 2.2.1, the result still holds when e_i is replaced with $\hat{e}_i$. Therefore,

$$[\hat{V}\{\hat{\boldsymbol{\beta}}_{W\Psi}\}]^{-1/2}(\hat{\boldsymbol{\beta}}_{W\Psi} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}), \tag{6.3.33}$$

where $\hat{V}\{\hat{\boldsymbol{\beta}}_{W\Psi}\}$ is as defined in (6.3.25). ■

Theorem 6.3.1 is closely related to Theorem 2.2.1. For example, result (6.3.29) of the proof could be obtained from Theorem 2.2.1 by setting $\boldsymbol{\Phi} = \text{diag}(w_i\psi_i)$. Theorem 2.2.1 gives the limiting distribution of $n_N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)$, while Theorem 6.3.1 gives the limiting distribution of $n_N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is defined by a superpopulation model. Also, result (6.3.24) for the estimator of the vector of superpopulation regression coefficients is closely related to result (3.3.19) for the estimator of the finite population mean with a two-phase sample.

Under the regularity conditions of Theorem 6.3.1, $\hat{\boldsymbol{\beta}}_{W\Psi}$ is consistent for $\boldsymbol{\beta}$ for any $\boldsymbol{\Psi}$ such that $E\{e_i \mid \mathbf{x}_i, \psi_i\} = 0$ and such that the moment requirements are satisfied. Therefore, we search for a $\boldsymbol{\Psi}$ that minimizes the variance of the approximate distribution, where

$$V\{\hat{\boldsymbol{\beta}}_{W\Psi}\} = n_N^{-1}\mathbf{M}_{X\Psi X}^{-1}(\mathbf{V}_{\infty,\bar{b}\bar{b}} + f_\infty\boldsymbol{\Sigma}_{bb})\mathbf{M}_{X\Psi X}^{-1}. \tag{6.3.34}$$

See (6.3.32). The variance expression is relatively simple for Poisson sampling from a finite population generated as independent random variables because $\pi_i^{-1}e_i$ is then independent of $\pi_j^{-1}e_j$ for all i and j. It follows that for Poisson sampling,

$$V\{\hat{\boldsymbol{\beta}}_{W\Psi} \mid \mathbf{X}_N, \boldsymbol{\Psi}_N\}$$

$$= E\left\{N^{-2}\mathbf{M}_{X\Psi X}^{-1}\mathbf{X}_N'\mathbf{D}_{\pi,N}^{-1}\boldsymbol{\Psi}_N^2\mathbf{D}_{ee,N}\mathbf{X}_N\mathbf{M}_{X\Psi X}^{-1} \mid \mathbf{X}_N, \boldsymbol{\Psi}_N\right\}, \tag{6.3.35}$$

where $\mathbf{D}_{ee,N} = \text{diag}(e_1^2, e_2^2, \ldots, e_N^2)$. To further simplify the search for $\boldsymbol{\Psi}$, consider the univariate model

$$y_i = x_{1i}\beta_1 + e_i, \tag{6.3.36}$$

$$e_i \sim ind(0, \sigma_i^2). \tag{6.3.37}$$

For Poisson sampling, the variance of the approximate distribution of $\hat{\beta}_{1,W\Psi} - \beta_1$ for model (6.3.36) is

$$E\left\{\left(\sum_{i\in U} x_{1i}^2\psi_i\right)^{-2}\sum_{i\in U} x_{1i}^2\psi_i^2a_i^2\right\}, \tag{6.3.38}$$

where $a_i^2 = \pi_i^{-1} e_i^2$. Because π_i can be a function of (x_{1i}, e_i), $E\{x_{1i}^2 \psi_i^2 a_i^2\}$ need not be $E\{x_{1i}^2 \psi_i^2\} E\{a_i^2\}$. Nevertheless, a ψ_i that is a good approximation for σ_{ai}^{-2}, where $\sigma_{ai}^2 = E\{a_i^2\}$, should give good efficiency for $\hat{\boldsymbol{\beta}}_{W\Psi}$.

Typically, σ_{ai}^2 is not known and we consider functions of x_i as approximations for σ_{ai}^2. A general representation for $\pi_i^{-1} e_i^2 = a_i^2$ is

$$a_i^2 = q_a(\mathbf{x}_i, \boldsymbol{\gamma}_a) + r_{ai}, \tag{6.3.39}$$

where r_{ai} includes the difference $a_i^2 - E\{a_i^2\}$ and the difference $E\{a_i^2\} - \psi_a(\mathbf{x}_i, \boldsymbol{\gamma}_a)$. Sometimes it is possible to specify a $q_a(\mathbf{x}_i, \boldsymbol{\gamma}_a)$ that is linear in the parameters. For example, we might write

$$a_i^2 = \mathbf{q}_i \boldsymbol{\gamma}_a + r_{ai}, \tag{6.3.40}$$

where the $\mathbf{q}_i$ are known functions of the $\mathbf{x}_i$. The $\mathbf{q}_i$ can contain variables other than functions of the $\mathbf{x}_i$, provided that the additional variables are such that $E\{e_i \mid \mathbf{x}_i, \mathbf{q}_i\} = 0$. To estimate $\boldsymbol{\gamma}_a$, we use a consistent estimator of $\boldsymbol{\beta}$ to construct estimators of the e_i, replace the e_i in (6.3.40) with the $\hat{e}_i$, and compute the regression of $\pi_i^{-1} \hat{e}_i^2$ on $\mathbf{q}_i$. One can compute a weighted regression using π_i^{-1} as weights or one can compute the ordinary least squares regression. The two approaches can give different limiting values for the estimated $\boldsymbol{\gamma}_a$, but the use of either $\hat{\boldsymbol{\gamma}}_a$ to construct the $\hat{\psi}_i$ will give a consistent estimator for $\boldsymbol{\beta}$.

A natural preliminary estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_\pi$ of (6.3.9). The work of Pfeffermann and Sverchkov (1999) provides an alternative preliminary estimator of $\boldsymbol{\beta}$. Pfeffermann and Sverchkov (1999) considered the regression model (6.3.1) with constant error variance and a design in which the π_i may be related to $\mathbf{x}_i$ and to e_i. They presented estimators based on detailed parametric models and also argued that it is reasonable to estimate $\boldsymbol{\beta}$ by minimizing

$$\sum_{i=1}^{n} w_i \tilde{w}_i^{-1} (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 \tag{6.3.41}$$

with respect to $\boldsymbol{\beta}$, where $w_i = \pi_i^{-1}$ and $\tilde{w}_i$ is an estimator of $E\{w_i | \mathbf{x}_i, \ i \in A\}$. The resulting estimator,

$$\hat{\boldsymbol{\beta}}_{PS} = (\mathbf{X}' \tilde{\mathbf{W}}^{-1} \mathbf{D}_\pi^{-1} \mathbf{X})^{-1} \mathbf{X}' \tilde{\mathbf{W}}^{-1} \mathbf{D}_\pi^{-1} \mathbf{y}, \tag{6.3.42}$$

is of the form (6.3.15) with $\boldsymbol{\Psi} = \tilde{\mathbf{W}}^{-1} = \text{diag}(\tilde{w}_1^{-1}, \ \tilde{w}_2^{-1}, \ \ldots, \ \tilde{w}_n^{-1})$. Given a linear model for w_i, the parameters of a functional representation of w_i can be estimated by ordinary least squares. Letting

$$w_i = \mathbf{q}_i \boldsymbol{\gamma}_w + r_{wi}, \tag{6.3.43}$$

where $\mathbf{q}_i$ is a function of $\mathbf{x}_i$, an estimator of γ_w is

$$\tilde{\gamma}_w = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{w}, \tag{6.3.44}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_n)'$ and the ith row of $\mathbf{Q}$ is $\mathbf{q}_i$. The estimator $\tilde{w}_i = \mathbf{q}_i\hat{\gamma}_w$ is an estimator of the expected value of w_i given x_i and given that element i is in the sample. In general, $\tilde{w}_i$ is not a consistent estimator of the superpopulation expected value of w_i given $\mathbf{x}_i$.

If the original model has constant error variances, and if the correlation between w_i and e_i^2 is modest, w_i will be correlated with σ_{ai}^2 to the degree that w_i is correlated with a function of $\mathbf{x}_i$. Hence, in situations with strong correlation, the estimator (6.3.42) will perform well. If we believe the error variances differ or that there is a correlation between w_i and e_i, we can use $\hat{\boldsymbol{\beta}}_{PS}$ or $\hat{\boldsymbol{\beta}}_{\pi}$ to construct $\hat{e}_i$ and use the $\hat{e}_i$ in (6.3.40) to estimate γ_a.

With an estimator of γ_a of (6.3.40), we define an estimator of $\boldsymbol{\beta}$ by

$$\tilde{\boldsymbol{\beta}}_{W\Psi} = (\mathbf{X}'\mathbf{D}_{\pi}^{-1}\hat{\boldsymbol{\Psi}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_{\pi}^{-1}\hat{\boldsymbol{\Psi}}\mathbf{y}, \tag{6.3.45}$$

where $\hat{\boldsymbol{\Psi}} = \text{diag}(\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_n)$ and $\hat{\psi}_i = \hat{\psi}_a(\mathbf{x}_i, \hat{\gamma}_a) = [q_a(\mathbf{x}_i, \hat{\gamma}_a)]^{-1}$. Under mild assumptions on $q_a(\mathbf{x}_i, \gamma_a)$, the population, and the design,

$$\hat{\gamma}_a = \gamma_a + O_p(n^{-1/2}). \tag{6.3.46}$$

Assume that $\psi_a(\mathbf{x}, \gamma_a)$ is continuous and twice differentiable, and let (6.3.46) be satisfied. Then

$$\psi_a(\mathbf{x}_i, \hat{\gamma}_a) = \psi_a(\mathbf{x}_i, \gamma_a) + \xi_{\psi}(\mathbf{x}_i, \gamma_a)(\hat{\gamma}_a - \gamma_a) + O_p(n^{-1}), \tag{6.3.47}$$

where $\xi_{\psi}(\mathbf{x}_i, \gamma) = \partial\psi_a(\mathbf{x}_i, \gamma)/\partial\gamma'$. It follows that

$$\begin{aligned}
&N^{-1}\sum_{i\in A}\mathbf{x}_i'\psi_a(\mathbf{x}_i, \hat{\gamma}_a)w_ie_i \\
&\quad = N^{-1}\sum_{i\in A}\mathbf{x}_i'\psi_a(\mathbf{x}_i, \gamma_a)w_ie_i \\
&\qquad + \left(N^{-1}\sum_{i\in A}\mathbf{x}_i'\xi_{\psi}(\mathbf{x}_i, \gamma_a)w_ie_i\right)(\hat{\gamma}_a - \gamma_a) + O_p(n^{-1}) \\
&\quad = N^{-1}\sum_{i\in A}\mathbf{x}_i'\psi_a(\mathbf{x}_i, \gamma_a)w_ie_i + O_p(n^{-1}),
\end{aligned} \tag{6.3.48}$$

because $E\{\mathbf{x}_i'\xi_\psi(\mathbf{x}_i, \gamma_a)e_i\} = \mathbf{0}$ and the multiplier for $(\hat{\gamma}_a - \gamma_a)$ in (6.3.48) is $O_p(n^{-1/2})$. Thus,

$$\tilde{\boldsymbol{\beta}}_{W\Psi} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{D}_\pi^{-1}\boldsymbol{\Psi}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}\boldsymbol{\Psi}\mathbf{e} + O_p(n^{-1}) \quad (6.3.49)$$

and the results of Theorem 6.3.1 hold for the estimator constructed with $\hat{\boldsymbol{\Psi}}$.

To apply the results of Theorem 6.3.1, we require a consistent estimator of the variance of $\hat{\boldsymbol{\beta}}_{W\Psi} - \boldsymbol{\beta}$. For Poisson sampling, variance estimation is relatively straightforward. The conditional variance of $\hat{\mathbf{M}}_{X\Psi e}$ for Poisson sampling from a finite population generated as independent random variables is

$$V\left\{\sum_{i\in A}\psi_i\mathbf{x}'\pi_i^{-1}e_i \mid \mathcal{F}\right\} = \sum_{i\in U}\psi_i^2\pi_i^{-1}(1-\pi_i)\mathbf{x}_i'e_i^2\mathbf{x}_i \quad (6.3.50)$$

and

$$V\left\{N^{-1}\sum_{i\in U}\psi_i\mathbf{x}_i'e_i \mid \mathbf{X}_N\right\} = N^{-2}\sum_{i\in U}\psi_i^2\mathbf{x}_i'\mathbf{x}_i\sigma_e^2.$$

Therefore, a consistent estimator of $V\{\hat{\boldsymbol{\beta}}_{W\Psi}\}$ for Poisson sampling is

$$\hat{V}\{\hat{\boldsymbol{\beta}}_{W\Psi}\} = n(n-k)^{-1}(\mathbf{X}'\hat{\boldsymbol{\Psi}}\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{D}}_{rr}\mathbf{X}(\mathbf{X}'\hat{\boldsymbol{\Psi}}\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}, \quad (6.3.51)$$

where k is the dimension of $\mathbf{x}_i$, $\hat{\mathbf{D}}_{rr} = \text{diag}(\hat{r}_{e1}^2, \hat{r}_{e2}^2, \ldots, \hat{r}_{en}^2)$, $\hat{r}_{ei} = \hat{\psi}_i^2\pi_i^{-2}\hat{e}_i^2$, and $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{W\Psi}$.

For a general design and a finite population correction that can be ignored, a variance estimator is

$$\hat{V}\{\hat{\boldsymbol{\beta}}_{W\Psi} \mid \mathcal{F}\} = (\mathbf{X}'\hat{\boldsymbol{\Psi}}\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\hat{V}\{N\hat{\mathbf{M}}_{X\Psi e}\}(\mathbf{X}'\hat{\boldsymbol{\Psi}}\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}, \quad (6.3.52)$$

where $\hat{V}\{N\hat{\mathbf{M}}_{X\Psi e}\} = \hat{V}\{\sum_{i\in A}\mathbf{x}_i'\psi_i\pi_i^{-1}e_i\}$ is the Horvitz–Thompson estimator, or other consistent estimator, of the variance of the sum calculated with $\mathbf{x}_i'\hat{\psi}_i\hat{e}_i$. See the proof of Theorem 2.2.1 for a proof that (6.3.52) is consistent for $V\{\hat{\boldsymbol{\beta}}_{W\Psi} - \boldsymbol{\beta}_N|\mathcal{F}\}$.

Example calculations for stratified sampling are given in Example 6.3.2.

Example 6.3.2. The data in Table 6.2 are a stratified sample with three strata, where the vector of sample sizes $(n_1, n_2, n_3) = (20, 10, 20)$. The sample is from a population of size 9000 composed of three types of elements. The parameter of interest is the mean of y. The type of observation, j, is known

for the sample elements, but the number of each type is not known. Similarly, there is no population information on x available for use in estimation. The population model has a common mean, but a variance that depends on type. Thus, the population model is

$$y_{ji} = \mu + e_{ji}, \tag{6.3.53}$$

where y_{ji} is observation i on type j and $e_{ji} \sim ind(0, \sigma^2_{ej})$. The regression estimator (6.3.45) for the μ of model (6.3.53) reduces to the ratio estimator

$$\tilde{\mu}_{W\psi} = \left(\sum_{h=1}^{3} w_h \sum_{ji \in A_h} \psi_j \zeta_{ji} \right)^{-1} \sum_{h=1}^{3} w_h \sum_{ji \in A_h} \psi_j \zeta_{ji} y_{hji}, \tag{6.3.54}$$

where we add the subscript h for stratum, $w_h = N^{-1} N_h n_h^{-1}$, A_h is the set of indexes in stratum h, and

$$\begin{array}{rcll} \zeta_{ji} & = & 1 & \text{if element } i \text{ is in group } j \\ & = & 0 & \text{otherwise.} \end{array}$$

The variance of the approximate distribution of estimator (6.3.54) is the

Table 6.2 Stratified Sample

Stratum	Type	y	Stratum	Type	y	Stratum	Type	y
1	1	12.90		2	14.86		3	11.58
	2	13.05		3	3.75		1	15.34
	2	8.19		1	12.72		3	24.29
	2	14.72	2	1	16.47		2	16.83
	3	8.04		3	15.45		1	16.92
	3	12.87		1	15.39		2	17.46
	2	8.80		2	15.69		2	19.37
	3	13.95		2	12.42		1	14.16
	3	5.92		3	12.97		2	17.28
	3	12.33		2	2.12		3	16.23
	3	2.78		1	13.02		3	15.77
	2	12.67		1	15.15		1	15.40
	1	13.37		1	13.38		3	18.18
	2	11.32	3	2	17.75		1	17.11
	1	13.59		3	27.20		3	20.26
	3	12.33		2	19.82		2	15.35
	3	16.15		2	17.66			

stratified variance for a ratio,

$$V\{\tilde{\mu}_{W\Psi}\} \doteq \left(N^{-1}\sum_{ji\in U}\psi_j\zeta_{ji}\right)^{-2}\sum_h N^{-2}N_h^2 n_h^{-1}\left(M_{2,\psi e,h}-M_{1,\psi e,h}^2\right), \tag{6.3.55}$$

where

$$M_{2,\psi e,h} = (N_h-1)^{-1}\sum_{ji\in U_h}\psi_j^2 e_{hji}^2,$$

$$M_{1,\psi e,h} = (N_h-1)^{-1}N_h^{-1}\left(\sum_{ji\in U_h}\psi_j e_{hji}\right)^2,$$

$e_{hji} = y_{hji}-\mu$, and U_h is the population set of indexes for stratum h. The sampling rate is small, so we ignore the Σ_{bb} part of the variance (6.3.32). Our objective is to find $\psi_j, j=1,2,3$, so that $\tilde{\mu}_{W\Psi}$ is more efficient than $\bar{y}_{st}$ for μ.

Attempting to minimize an estimator of expression (6.3.55) with respect to the ψ_j would result in ψ_j that are functions of the components of $M_{1,\psi e,h}$, where those components are functions of the means of the e_{hji} within the strata. Weights that are functions of means may produce bias in the estimator of μ. Therefore, we prefer $\hat{\psi}_j$ that depend as little as possible on means of the e_{hji} and we consider only the portion of the variance (6.3.55) that is a function of $M_{2,\psi e,h}$. That portion of the variance is of the same form as (6.3.51), and we choose $\hat{\psi}_j^{-1}$ to estimate the expected value of $w_h e_{hji}^2$. An estimator of the expected value of $w_h e_{hji}^2$ for elements of type j is

$$\hat{\psi}_j^{-1} = \left(\sum_{h=1}^{3}\sum_{ji\in A} w_h\zeta_{ji}\right)^{-1}\sum_{h=1}^{3}\sum_{ji\in A} w_h^2\zeta_{ji}\hat{e}_{hji}^2,$$

where $\hat{e}_{hji} = y_{hji}-\bar{y}_{st}$. The three values are $475.4, 2253.8$, and 5307.6 for $j=1,2$, and 3, respectively. The estimator (6.3.54) computed with $\hat{\psi}_j$ replacing ψ_j is $\hat{\mu}_{W\Psi} = 14.533$.

An estimator of the variance of $\hat{\mu}_{W\Psi}$ is the estimated variance for a ratio,

$$\hat{V}\{\hat{\mu}_{W\Psi}\} = \left(\sum_{h=1}^{3} w_h\sum_{ji\in A_h}\hat{\psi}_j\zeta_{ji}\right)^{-2}\hat{V}\{\bar{e}_{W\Psi}\},$$

where

$$\hat{V}\{\bar{e}_{W\Psi}\} = \sum_{h=1}^{3} K_{1h}\sum_{ji\in A_h} w_h^2\hat{\psi}_j^2\hat{e}_{hji}^2 - \sum_{h=1}^{3} K_{2h}\left(\sum_{ji\in A_h} w_h\hat{\psi}_j\hat{e}_{hji}\right)^2,$$

$K_{1h} = n_h(n_h - 1)^{-1}$, $K_{2h} = (n_h - 1)^{-1}$, and $\hat{e}_{hji} = y_{hji} - \hat{\mu}_{W\Psi}$. For our sample,

$$\hat{V}\{\hat{\mu}_{W\Psi}\} = 0.0981.$$

The stratified estimate of the mean is

$$\bar{y}_{st} = (11.219 + 13.906 + 17.702)/3 = 14.276$$

and the estimated variance of $\bar{y}_{st}$, ignoring the finite population correction, is

$$\hat{V}\{\bar{y}_{st}\} = (13.829/20 + 4.771/10 + 11.676/20)/9 = 0.1947.$$

There is an estimated gain of about 40% in efficiency from using the ψ-estimator relative to the stratified estimator. For additional analyses of the illustrative data, see Exercises 8 and 9. ■■

In Example 6.3.3 we analyze a data set that has structure similar to that of Example 6.3.2.

Example 6.3.3. The Canadian Workplace and Employee Survey conducted by Statistics Canada was introduced in Example 3.1.1. The sample is a stratified sample with simple random sampling of workplaces in strata. The population of workplaces in Canada was placed in categories based on industrial activity and region. Then workplaces in each category were further divided into three strata on the basis of a function of 1998 tax records, where the function correlates highly with payroll. The sample allocation to strata was made on the basis of the variance of the measure of size. The sampling rates were quite different for the three strata. The original weights are about 2200 for the small workplace stratum, about 750 for the medium-sized workplace stratum, and about 35 for the large workplace stratum. The original weights were adjusted, primarily for nonresponse. The survey is described in Patak, Hidiroglou, and Lavallée (1998). The data in Table 6.3 were generated to approximate data collected in the 1999 survey of workplaces and are representative of an activity-region category. The original analysis was performed by Zdenek Patak at Statistics Canada.

Table 6.3: **Canadian Workplace Data**

ID	Weight	Employ-ment	Payroll (×1000)	ID	Weight	Employ-ment	Payroll (×1000)
1	786	17	260	2	32	661	6873
3	36	3	366	4	38	58	1336
5	37	55	1868	6	30	39	1251
7	2557	3	65	8	39	86	1276
9	35	39	482	10	39	53	897
11	2198	3	39	12	2198	7	78
13	29	6	197	14	35	151	1861
15	34	71	615	16	37	200	4520
17	26	127	1416	18	37	118	826
19	30	136	1943	20	34	46	732
21	36	157	1916	22	34	244	4193
23	28	73	1269	24	30	3	71
25	743	29	403	26	743	37	374
27	38	73	990	28	29	7	441
29	1929	4	97	30	33	2	86
31	845	22	233	32	495	3	129
33	38	3	139	34	612	28	448
35	141	23	618	36	2378	4	132
37	743	40	851	38	2512	3	38
39	2198	6	123	40	2109	1	24
41	597	5	161	42	32	60	1462
43	43	68	2008	44	39	107	916
45	34	65	4394	46	25	48	697
47	31	96	1544	48	40	94	373
49	29	173	599	50	32	79	1689
51	34	37	1295	52	31	71	1595
53	41	4	112	54	743	15	513
55	39	102	1027	56	699	43	699
57	31	136	702	58	36	99	1420
59	28	143	2567	60	2288	7	186
61	2647	2	14	62	1884	3	71
63	37	55	794	64	39	180	5413
65	2378	7	166	66	25	149	2373
67	1750	1	27	68	39	93	883
69	32	99	804	70	43	82	909
71	36	156	731	72	36	108	5508

Continued

ID	Weight	Employ-ment	Payroll (×1000)	ID	Weight	Employ-ment	Payroll (×1000)
73	36	75	1685	74	1974	5	63
75	50	223	2436	76	32	75	1549
77	39	62	1590	78	34	75	641
79	44	73	491	80	2153	1	20
81	684	18	300	82	1839	7	113
83	24	213	2445	84	27	145	5185
85	29	81	622	86	30	5	106
87	39	66	1524	88	40	26	314
89	32	77	1933	90	76	58	1036
91	38	37	506	92	35	194	3989
93	40	63	767	94	1884	9	65
95	23	56	1257	96	2423	6	74
97	655	50	786	98	2826	3	17
99	524	22	186	100	2737	1	21
101	2109	3	47	102	129	93	2247
103	37	87	1234	104	26	96	911
105	1929	6	410	106	1660	7	63
107	17	45	967	108	39	96	1103
109	34	76	1956	110	38	92	1624
111	32	68	1481	112	568	23	285
113	1884	9	188	114	2557	5	49
115	12	55	737	116	34	74	1358
117	33	187	4097	118	25	61	1601
119	2109	1	11	120	31	31	394
121	2423	7	162	122	2333	6	75
123	32	190	3068	124	699	24	321
125	2109	3	47	126	80	127	1565
127	42	301	2781	128	33	22	718
129	39	16	131	130	30	82	938
131	32	50	1492	132	62	17	351
133	15	358	4310	134	28	99	1373
135	1884	14	54	136	34	10	310
137	28	131	1734	138	582	39	582
139	35	22	301	140	2243	8	48
141	2423	3	46	142	1839	14	161

The initial subject matter regression model for the population is

$$y_i = \beta_0 + x_i\beta_1 + e_i, \tag{6.3.56}$$

where the e_i are $iid(0,\ \sigma^2)$ random variables, $E\{e_i \mid x_i\} = 0$, y_i is the logarithm of 1999 payroll, and x_i is the logarithm of 1999 total employment. The simple sample mean vector is $(\bar{y}_n,\ \bar{x}_n) = (13.082, 3.376)$ and the weighted mean vector is $(\bar{y}_\pi,\ \bar{x}_\pi) = (11.3222,\ 1.693)$. The large difference between the two estimates is due to the wide range in the weights and the fact that the weights are strongly correlated with size. Fitting model (6.3.56) by ordinary least squares (OLS), we obtain

$$\begin{aligned} \bar{y}_i &= 13.082 + 0.907(x_i - \bar{x}_n) \\ &\quad\ (0.048)\ \ (0.032) \\ s^2 &= 0.320, \end{aligned} \tag{6.3.57}$$

where $\bar{x}_n = 3.376$ is the simple sample mean, and the numbers in parentheses are the standard errors from the OLS calculations. We centered x_i at the sample mean to make it easier to identify some of the effects.

The probability weighted regression of (6.3.9) is

$$\begin{aligned} \hat{y}_i &= 12.889 + 0.931(x_i - \bar{x}_n), \\ &\quad\ (0.113)\ \ (0.053) \end{aligned} \tag{6.3.58}$$

where the numbers in parentheses are the standard errors calculated from the covariance matrix (6.3.52) for a stratified sample. If we use the residuals described in Section 2.2.4 to calculate the standard errors, we obtain estimated biases in the variance estimators due to estimation of e_i of about 2%.

The standard error of the weighted estimator of the intercept is more than twice that of OLS. Equations (6.3.57) and (6.3.58) are not greatly different, but the intercepts do differ by more than three OLS standard errors.

To test the hypothesis that the two estimators are estimating the same quantity, we compute the OLS regression of y_i on $[1,\ x_i - \bar{x}_n,\ w_i^*,\ w_i^*(x_i - \bar{x}_n)]$, where $w_i^* = 0.001(w_i - \bar{w}_n)$ and $\pi_i = w_i^{-1}$ is the selection probability. The estimated equation is

$$\begin{aligned} \hat{y}_i &= 13.099 + 0.732(x_i - \bar{x}_n) - 0.372w_i^* + 0.016w_i^*(x_i - \bar{x}_n) \\ &\quad\ (0.070)\ \ (0.047) \qquad\qquad (0.117) \qquad (0.053) \end{aligned} \tag{6.3.59}$$

and $s^2 = 0.263$, where the numbers in parentheses are the standard errors from the OLS calculations. The F test for the hypothesis of common expectation for $\hat{\boldsymbol{\beta}}_\pi$ and $\hat{\boldsymbol{\beta}}_{ols}$ is $F(2, 138) = 15.20$. The 1% point for F with 2 and 138 degrees of freedom is 4.76, and we conclude that the OLS estimator is biased for $\boldsymbol{\beta}$. The significant test is consistent with the sample design. The stratification was based on an index of payroll constructed from tax records. Thus, a correlation between the error in the payroll model for 1999 and the index for 1998 is plausible.

In an attempt to find an estimator superior to the probability weighted estimator, we begin with the Pfeffermann–Sverchkov estimator. To construct the estimator we first regress w_i on functions of x_i. If the sample weights for the payroll data are plotted against x_i, they fall into three groups with respect to x. The groups are defined by the intervals (0, 2.71), [2.71, 4.09), and [4.09, ∞). The means of w_i for the three groups are 1655.5, 279.2, and 36.2, respectively. Using these estimated weights, we calculate the Pfeffermann–Sverchkov estimator,

$$\hat{\boldsymbol{\beta}}_{PS} = (\mathbf{X}'\mathbf{D}_{\pi}^{-1}\tilde{\mathbf{W}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_{\pi}^{-1}\tilde{\mathbf{W}}^{-1}\mathbf{y},$$

where $\tilde{\mathbf{W}} = \text{diag}(\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n)$ and $\tilde{w}_i$ are the predicted values (group means) for the regression of w_i on dummy variables for the intervals of x, and the ith row of $\mathbf{X}$ is $(1, x_i - \bar{x})$. The estimated equation is

$$\begin{aligned} \hat{y}_{PS,i} &= 12.975 + 0.969(x_i - \bar{x}), \\ &\quad (0.048)\ (0.032) \end{aligned} \tag{6.3.60}$$

where the standard errors were calculated using (6.3.52) with $\hat{\boldsymbol{\Psi}} = \tilde{\mathbf{W}}^{-1}$. The standard errors of (6.3.60) are essentially equal to those of OLS. The estimated covariance matrix is

$$\hat{V}\{\hat{\boldsymbol{\beta}}_{PS}\} = \begin{pmatrix} 2.286 & -0.005 \\ -0.039 & 1.011 \end{pmatrix} \times 10^{-3}.$$

To construct an estimator with the estimated value for $\boldsymbol{\Psi}$ based on σ_{ai}^2, where $\sigma_{ai}^2 = E\{\pi_i^{-1}a_i^2\}$, let $w_i\tilde{e}_i^2 = \tilde{a}_{1i}^2$, where $\tilde{e}_i = y_i - (1, x_i - \bar{x})\hat{\boldsymbol{\beta}}_{PS}$. The regression equation estimated for the regression of $\tilde{a}_{1i}^2$ on dummy variables defined for the intervals of x is

$$\hat{a}_{1i}^2 = 11.19 + 536.02z_{1i} + 9.62z_{2i}, \tag{6.3.61}$$

where

$$\begin{aligned} z_{1i} &= 1 \quad \text{if } x_i < 2.71 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

and

$$\begin{aligned} z_{2i} &= 1 \quad \text{if } 2.71 \leq x_i < 4.09 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Using (6.3.61), the $\hat{\psi}_i^{-1}$, scaled to be comparable to $\tilde{w}_i$, are

$$\begin{aligned} \hat{\psi}_i^{-1} &= 1770.3 \quad \text{if } x_i < 2.71 \\ &= 65.8 \quad \text{if } 2.71 \leq x_i < 4.09 \\ &= 36.2 \quad \text{if } x_i \geq 4.09. \end{aligned} \tag{6.3.62}$$

The estimated equation calculated with $\tilde{\beta}_{W\Psi}$ of (6.3.45) is

$$\begin{aligned} \hat{y}_{w\psi,i} &= 12.963 + 0.979x_i, \\ &\quad (0.044)\ (0.032) \end{aligned} \tag{6.3.63}$$

where the standard errors were calculated using the estimated variance matrix (6.3.52).

The estimated variance for the intercept for the Ψ-estimator is about 15% less than the estimated variance for the Pfeffermann–Sverchkov estimator and comparable to that for OLS. The estimated superiority relative to Pfeffermann–Sverchkov is most likely due to differing error variances. If we regress $\tilde{w}_i^{-1}\tilde{a}_{1i}^2$ on the dummy variables, we obtain the estimated equation

$$\begin{aligned} \tilde{w}_i^{-1}\tilde{a}_{1i}^2 &= 0.309 + 0.021z_{1i} - 0.235z_{2i}, \\ &\quad (0.058)\ (0.089)\quad (0.096) \end{aligned} \tag{6.3.64}$$

where the numbers in parentheses are the OLS standard errors. The regression (6.3.64) suggests that σ_{ai}^2 is not a constant multiple of $\tilde{w}_i$.

In comparing OLS and the Ψ-estimator one should remember that the estimated variances for OLS and for the Ψ-estimator are biased, although the relative bias is $O(n^{-1})$. Second, if the error variances are not constant, the OLS estimator is not the minimum variance estimator.

The model (6.3.56) specifies $E\{e_i \mid x_i\} = 0$, and this assumption is required for consistency. The assumption is only partially subject to test, but it is possible to test the hypothesis that e is uncorrelated with the indicators used to estimate Ψ, given model (6.3.56). We construct a test by adding the two indicators of equation (6.3.64) to the regression model for y. Using the probability weighted estimator, the 2×2 portion of the covariance matrix associated with (z_{1i}, z_{2i}) is

$$\hat{V}_{22} = \begin{pmatrix} 0.1898 & 0.0716 \\ 0.0716 & 0.0348 \end{pmatrix}$$

and the test statistic for the hypothesis that the two coefficients are zero is

$$F(2, 136) = 0.5(-0.365, -0.183)\hat{\mathbf{V}}_{22}^{-1}(-0.365, -0.183)' = 0.49.$$

One easily accepts the hypothesis that the mean of the e's is zero for each of the three groups used to define Ψ. ■■

6.4 INSTRUMENTAL VARIABLES

6.4.1 Introduction

In estimation for the model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i, \tag{6.4.1}$$

where $E\{e_i\} = 0$, there may be some members of $\mathbf{x}_i$ that cannot be treated as independent of e_i. This occurs in econometric models when some of the variables are mutually determined under a simultaneous equation model. The variables correlated with e_i are called *endogenous* and denoted by y in econometrics. In a second situation some members of $\mathbf{x}_i$ are measured with error. In both the measurement error and endogenous variable cases, OLS estimators are biased.

Assume that some additional variables, denoted by $\mathbf{u}_i$, are available with the superpopulation properties

$$E\{\mathbf{u}_i' e_i\} = \mathbf{0} \tag{6.4.2}$$

and

$$E\{\mathbf{x}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{x}_i\} \neq 0. \tag{6.4.3}$$

Variables satisfying (6.4.2) and (6.4.3) are called *instrumental variables* or *instruments*. It is sometimes said that $\mathbf{u}$ is an instrument for $\mathbf{x}$.

6.4.2 Weighted instrumental variable estimator

We first construct the probability weighted version of the standard instrumental variable estimator. Assume that we have a sample selected from a finite population, where the finite population is generated as a simple random sample from a superpopulation in which (6.4.1) holds. Let $(y_i, \mathbf{x}_i, \mathbf{u}_i, w_i)$, where $N^{-1}\pi_i^{-1} = w_i$, be the vector of observations. If we multiply (6.4.1) by $\mathbf{u}_i' w_i$, and sum, we obtain

$$\sum_{i\in A} \mathbf{u}_i' w_i y_i = \sum_{i\in A} \mathbf{u}_i' w_i \mathbf{x}_i\boldsymbol{\beta} + \sum_{i\in A} \mathbf{u}_i' w_i e_i. \tag{6.4.4}$$

If (6.4.1), (6.4.2), and (6.4.3) hold in the superpopulation, then

$$E\left\{\sum_{i\in A} w_i e_i\right\} = E\left\{\sum_{i\in U} e_i\right\} = 0, \tag{6.4.5}$$

$$E\left\{\sum_{i\in A} w_i \mathbf{u}_i' e_i\right\} = E\left\{\sum_{i\in U} \mathbf{u}_i' e_i\right\} = \mathbf{0}, \tag{6.4.6}$$

and

$$E\left\{\sum_{i\in A} w_i \mathbf{u}_i' \mathbf{x}_i \mathbf{x}_i' \mathbf{u}_i\right\} = E\left\{\sum_{i\in U} \mathbf{u}_i' \mathbf{x}_i \mathbf{x}_i' \mathbf{u}_i\right\} \neq 0. \tag{6.4.7}$$

The $\mathbf{u}_i$ can contain some elements of $\mathbf{x}_i$ provided that the model specifies those elements to be independent of e_i. Letting $\mathbf{z}_i = w_i \mathbf{u}_i$, equation (6.4.4) in matrix notation is

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{b}}, \tag{6.4.8}$$

where $\hat{\mathbf{b}} = \mathbf{Z}'\mathbf{e}$, $\mathbf{Z}$ is an $n \times k_2$ matrix, $\mathbf{X}$ is an $n \times k_1$ matrix, and $k_2 \geq k_1$. Equation (6.4.8) has the appearance of a regression problem, where $\mathbf{Z}'\mathbf{y}$ is the vector of "dependent" variables, $\mathbf{Z}'\mathbf{X}$ is the matrix of "explanatory" variables, and $\hat{\mathbf{b}}$ is the vector of errors in the equation. Because the elements of $\hat{\mathbf{b}}$ are correlated with unequal variance, it would not be appropriate to apply ordinary least squares to (6.4.8) unless $k_2 = k_1$. To obtain an estimator of $\boldsymbol{\beta}$ superior to ordinary least squares, we require an approximation to the covariance matrix of $\hat{\mathbf{b}}$. Letting such an estimator be denoted by $\tilde{\mathbf{V}}_{bb}$, we define an instrumental variable estimator by

$$\hat{\boldsymbol{\beta}}_{IV} = [(\mathbf{Z}'\mathbf{X})'\tilde{\mathbf{V}}_{bb}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{Z}'\mathbf{X})'\tilde{\mathbf{V}}_{bb}^{-1}\mathbf{Z}'\mathbf{y}. \tag{6.4.9}$$

If e_i^2 has a small correlation with w_i and if the original e_i have common variances, a first approximation to a multiple of the covariance matrix of $\hat{\mathbf{b}}$ is

$$\tilde{\mathbf{V}}_{bb} = \mathbf{Z}'\mathbf{Z}. \tag{6.4.10}$$

Estimator (6.4.9) with $\tilde{\mathbf{V}}_{bb}$ of (6.4.10) is called the *two-stage least squares estimator*. See, for example, Wooldridge (2006, Chapter 15). The estimator (6.4.9) with $\mathbf{z}_i$ proportional to $w_i \mathbf{u}_i$ is the instrumental variable analog of the design weighted regression estimator (6.3.9).

6.4.3 Instrumental variables for weighted samples

Instrumental variable estimation is a method suitable for models in which the error in the equation is correlated with the explanatory variables. If π_i and e_i are correlated, then

$$E\{x_i e_i \mid i \in A\} \neq 0$$

and the basic problem is the same as that associated with the presence of endogenous variables or measurement error in the explanatory variables.

In the sampling situation, there are some ready-made instrumental variables. Under the assumption that $E\{e_i \mid \mathbf{x}_i\} = 0$ for all i in the superpopulation, $\mathbf{x}_i w_i$ is a possible vector of instrumental variables because

$$E\left\{\sum_{i\in A} \mathbf{x}_i w_i e_i\right\} = E\left\{N^{-1}\sum_{i\in U} \mathbf{x}_i e_i\right\} = \mathbf{0}.$$

In fact, any function of $\mathbf{x}_i$ multiplied by w_i is a potential instrument under the assumption that $E\{e_i \mid \mathbf{x}_i\} = 0$.

When $\mathbf{z}_i = w_i\mathbf{x}_i$ and $\tilde{\mathbf{V}}_{bb} = \mathbf{Z}'\mathbf{Z}$, estimator (6.4.9) reduces to the probability-weighted estimator (6.3.9). Thus, the estimator $\hat{\boldsymbol{\beta}}_\pi$ of (6.3.9) is a particular instrumental variable estimator. Similarly, the Pfeffermann–Sverchkov estimator (6.3.42) is an instrumental variable estimator with $\mathbf{z}_i = \tilde{w}_i^{-1} w_i \mathbf{x}_i$.

The error in estimator (6.4.9) is

$$\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta} = \hat{\mathbf{L}}'\hat{\mathbf{b}}, \tag{6.4.11}$$

where

$$\hat{\mathbf{L}}' = [(\mathbf{Z}'\mathbf{X})'\tilde{\mathbf{V}}_{bb}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{Z}'\mathbf{X})'\tilde{\mathbf{V}}_{bb}^{-1}$$

and $\hat{\mathbf{b}}$ is as defined in (6.4.8). Under the assumption that the finite population is a sample of independent random variables, the proof of Theorem 6.3.1 can be mimicked to show that the estimator has a normal distribution in the limit. The large-sample variance of $\hat{\boldsymbol{\beta}}_{IV}$ is

$$V_{LS}\{n^{1/2}(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta})\} = \mathbf{L}_N'[V_\infty\{n^{1/2}\hat{\mathbf{b}} \mid \mathcal{F}_N\} + V\{n^{1/2}\mathbf{b}_N\}]\mathbf{L}_N, \tag{6.4.12}$$

where $\mathbf{b}_N = \mathbf{Z}_N'\mathbf{D}_{\pi,N}\mathbf{e}_N$, and $V_\infty\{n^{1/2}\hat{\mathbf{b}} \mid \mathcal{F}_N\}$ is the variance of the limiting distribution of $n^{1/2}\hat{\mathbf{b}}$,

$$\mathbf{L}_N' = E\{[(\mathbf{Z}_N'\mathbf{D}_{\pi,N}\mathbf{X}_N)'\mathbf{V}_{\infty,bb}^{-1}(\mathbf{Z}_N'\mathbf{D}_{\pi,N}\mathbf{X}_N)]^{-1}(\mathbf{Z}_N'\mathbf{D}_{\pi,N}\mathbf{X}_N)'\mathbf{V}_{\infty,bb}^{-1}\},$$

$\mathbf{V}_{\infty,bb} = p\lim n\tilde{\mathbf{V}}_{bb}$, and $\mathbf{D}_{\pi,N} = \operatorname{diag}(\pi_1, \pi_2, \ldots, \pi_N)$.

The variance of $\hat{\boldsymbol{\beta}}_{IV}$ can be estimated with

$$\hat{V}\{\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}\} = \hat{\mathbf{L}}'[\hat{V}\{\hat{\mathbf{b}} \mid \mathcal{F}\} + \mathbf{Z}'\mathbf{D}_\pi\hat{\mathbf{D}}_{ee}\mathbf{Z}]\hat{\mathbf{L}}, \tag{6.4.13}$$

where $\hat{\mathbf{D}}_{ee} = \operatorname{diag}(\hat{e}_i^2)$, $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{IV}$, and $\hat{V}\{\hat{\mathbf{b}} \mid \mathcal{F}\}$ is a consistent quadratic estimator of the variance of $\hat{\mathbf{b}}$ calculated with $\hat{e}_i$ replacing e_i.

If e_i^2 is strongly correlated with w_i, it is possible to improve the estimator using an estimator of the covariance matrix of $\hat{\mathbf{b}}$. Thus, a second round estimator is (6.4.9) with $\hat{V}\{\hat{\mathbf{b}} \mid \mathcal{F}\}$ of (6.4.13) replacing the $\tilde{\mathbf{V}}_{bb}$ of (6.4.10).

6.4.4 Instrumental variable pretest estimators

In Example 6.3.3 we used a global test of the hypothesis that the unweighted estimator is unbiased. We now investigate tests for more specific hypotheses under the superpopulation model

$$y_i = \beta_0 + \mathbf{x}_{1,i}\boldsymbol{\beta}_1 + e_i, \tag{6.4.14}$$

where $E\{e_i \mid \mathbf{x}_i\} = 0$ for all i. One situation that leads to a test for a reduced set of explanatory variables is that in which the selection probability is correlated with the error e_i, but

$$E\left\{\sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'e_i\right\} = \mathbf{0}. \tag{6.4.15}$$

Model (6.4.15) will hold if the e_i are $iid(0, \sigma^2)$ random variables independent of $\mathbf{x}_i$ and the selection probabilities have the representation

$$\pi_i = g_1(\mathbf{x}_i) + g_2(e_i) + u_i, \tag{6.4.16}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are continuous differentiable functions and u_i is independent of $(\mathbf{x}_i,\ e_i)$. Then

$$E\left\{\sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'e_i\right\} = E\left\{\sum_{i\in U}\pi_i(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'e_i\right\} = \mathbf{0} \tag{6.4.17}$$

because $E\{g_2(e_i)e_i\}$ is a constant. It follows that the estimator defined by

$$\left(\sum_{i\in A}\left[\pi_i^{-1},\ (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})\right]'(1,\ \mathbf{x}_{1,i})\right)\hat{\boldsymbol{\beta}}_{IV} = \sum_{i\in A}\left[\pi_i^{-1},\ (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})\right]'y_i$$

is consistent for $\boldsymbol{\beta}$. We can replace $\bar{\mathbf{x}}_{1,N}$ with $\bar{\mathbf{x}}_{1,\pi}$ if $\bar{\mathbf{x}}_{1,N}$ is unknown. Then, under (6.4.17) and our usual moment assumptions, the estimator defined by

$$\sum_{i\in A}[\pi_i^{-1},\ (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})]'(1,\ \mathbf{x}_{1,i})\hat{\boldsymbol{\beta}}_{IV} = \sum_{i\in A}\left[\pi_i^{-1},\ (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})\right]'y_i \tag{6.4.18}$$

is consistent for $\boldsymbol{\beta}$.

We now develop a test of the hypothesis that a set of variables can be used as instruments, given an initial set known to satisfy the requirements for instruments. We consider the general problem, but we will often be

interested in testing the set $(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})$, given the set $(w_i,\ w_i\mathbf{x}_{1,i})$. Let $\mathbf{z}_i = (\mathbf{z}_{1i}, \mathbf{z}_{2i})$ be the vector potential instruments, where $\mathbf{z}_{1i}$ is known to satisfy the requirements of instruments. Thus, it is assumed that

$$E\left\{\sum_{i\in A}\mathbf{z}_{1,i}e_i\right\} = \mathbf{0} \tag{6.4.19}$$

and we wish to test

$$E\left\{\sum_{i\in A}\mathbf{z}_{2,i}e_i\right\} = \mathbf{0}. \tag{6.4.20}$$

We write the subject matter equation in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{6.4.21}$$

The two-stage least squares estimator constructed using the entire $\mathbf{z}$ vector can be written as

$$\hat{\boldsymbol{\beta}}_{IV} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \tag{6.4.22}$$

where

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

From (6.4.9), the estimated generalized least squares version of the instrumental variable estimator is

$$\hat{\boldsymbol{\beta}}_{IV,EGLS} = [(\mathbf{Z}'\mathbf{X})'\hat{\mathbf{V}}_{bb}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{Z}'\mathbf{X})'\hat{\mathbf{V}}_{bb}^{-1}\mathbf{Z}'\mathbf{y}, \tag{6.4.23}$$

where $\hat{\mathbf{V}}_{bb}$ is an estimator of the variance of $\mathbf{Z}'\mathbf{e}$.

To test that $E\{\mathbf{Z}_2'\mathbf{e}\} = \mathbf{0}$, using (6.4.22) as our basic estimator, we compute

$$(\hat{\gamma}_1', \hat{\gamma}_2')' = \hat{\gamma} = [(\hat{\mathbf{X}}, \mathbf{R}_2)'(\hat{\mathbf{X}}, \mathbf{R}_2)]^{-1}(\hat{\mathbf{X}}, \mathbf{R}_2)'\mathbf{y}, \tag{6.4.24}$$

where $\mathbf{R}_2 = \mathbf{Z}_2 - \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Z}_2$ and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$. If the finite population correction can be ignored, an estimated covariance matrix for $\hat{\gamma}$ is

$$\hat{V}\{\hat{\gamma}\} = [(\hat{\mathbf{X}}, \mathbf{R}_2)'(\hat{\mathbf{X}}, \mathbf{R}_2)]^{-1}\hat{V}\{(\hat{\mathbf{X}}, \mathbf{R}_2)'\mathbf{e}\}[(\hat{\mathbf{X}}, \mathbf{R}_2)'(\hat{\mathbf{X}}, \mathbf{R}_2)]^{-1}. \tag{6.4.25}$$

Continuing to ignore the finite population correction, an estimator of $V\{(\hat{\mathbf{X}}, \mathbf{R}_2)'\mathbf{e}\}$ is the Horvitz–Thompson estimator calculated with $(\hat{\mathbf{X}}, \mathbf{R}_2)'\hat{\mathbf{e}}$, where $\hat{e}_i = y_i - (\hat{\mathbf{x}}_i, \mathbf{r}_{2i})\hat{\gamma}$ and $\mathbf{r}_{2i}$ is the ith row of $\mathbf{R}_2$. Under the null hypothesis that $E\{\mathbf{Z}_2'\mathbf{e}\} = \mathbf{0}$, $\hat{\gamma}_1$, the coefficient for $\hat{\mathbf{X}}$, is estimating $\boldsymbol{\beta}$, and $\hat{\gamma}_2$,

the coefficient for $\mathbf{r}_{2i}$, is estimating the zero vector. Therefore, a test statistic is

$$F(k_{22}, d) = k_{22}^{-1}\hat{\boldsymbol{\gamma}}_2'\hat{\mathbf{V}}_{\gamma\gamma 22}^{-1}\hat{\boldsymbol{\gamma}}_2, \tag{6.4.26}$$

where k_{22} is the dimension of $\mathbf{r}_{2i}$, $\hat{\mathbf{V}}_{\gamma\gamma 22}$ is the lower right $k_{22} \times k_{22}$ block of $\hat{V}\{\hat{\boldsymbol{\gamma}}\}$, and d is the number of primary sampling units less the number of strata. Under the null hypothesis that $E\{\mathbf{Z}_2'\mathbf{e}\} = \mathbf{0}$, the test statistic is approximately distributed as F with k_{22} and d_F degrees of freedom, where, for cluster-stratified sampling, a reasonable value for d_F is the number of primary sampling units less the number of strata.

Example 6.4.1. We continue study of the Canadian workplace data. On the basis of the analysis of Example 6.3.3, $\mathbf{z}_{1i} = [w_i, w_i(x_i - \bar{x}_n), \hat{\Psi}_i w_i, \hat{\Psi}_i w_i(x_i - \bar{x}_n)]$, where $\hat{\Psi}_i$ is defined in (6.3.62), can be used as a vector of instrumental variables. The estimated equation using the four variables as instruments in $\hat{\boldsymbol{\beta}}_{IV}$ of (6.4.22) is

$$\begin{aligned} \hat{y}_i &= 12.976 + 0.977(x_i - \bar{x}_n), \\ &\quad (0.050)\ (0.031) \end{aligned} \tag{6.4.27}$$

where the standard errors are the square roots of the diagonal elements of

$$(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\hat{\mathbf{D}}_{ee}\hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} = \begin{pmatrix} 2.333 & -0.174 \\ -0.174 & 0.988 \end{pmatrix} \times 10^{-3}, \tag{6.4.28}$$

$\hat{\mathbf{D}}_{ee} = \text{diag}(\hat{e}_1^2, \hat{e}_2^2, \ldots, \hat{e}_n^2)$, and $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{IV}$. This estimated variance is very similar to the estimated variance that recognizes the stratification.

The estimated generalized least squares (EGLS) estimator of (6.4.23) is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{IV,EGLS} &= [(\mathbf{Z}_1'\mathbf{X})'(\mathbf{Z}_1'\hat{\mathbf{D}}_{ee}\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{X}]^{-1}(\mathbf{Z}_1'\mathbf{X})'(\mathbf{Z}_1'\hat{\mathbf{D}}_{ee}\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{y} \\ &= (12.974,\ 0.975)', \\ &\quad (0.042)\ (0.030) \end{aligned} \tag{6.4.29}$$

where $\hat{\mathbf{V}}_{bb} = \mathbf{Z}_1'\hat{\mathbf{D}}_{ee}\mathbf{Z}_1$ and $\hat{\mathbf{D}}_{ee}$ is defined in (6.4.28). The standard errors in (6.4.29) are the diagonal elements of

$$\hat{V}\{\hat{\boldsymbol{\beta}}_{IV,EGLS}\} = [(\mathbf{Z}_1'\mathbf{X})'(\mathbf{Z}_1'\hat{\mathbf{D}}_{ee}\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{X}]^{-1}. \tag{6.4.30}$$

The estimated variances for EGLS are usually underestimates, but the fairly large reduction in the standard error for the intercept suggests that EGLS represents a real improvement relative to estimator (6.4.22).

In the ordinary least squares regression of y_i on $[1,\ x_i - \bar{x}_n,\ w_i^*,\ w_i^*(x_i - \bar{x}_n)]$ displayed in (6.3.59), the coefficient for $w_i^*(x_i - \bar{x}_n)$ is less than the

standard error. The nonsignificant coefficient is consistent with the way in which the selection probabilities were determined. Because the selection probabilities are related to the payrolls of the previous year and assuming a similar subject matter model for the previous year, (6.4.16) is a reasonable specification for the probabilities. This suggests that the instrumental variable estimator with instrument vector containing $x_i - \bar{x}_\pi$ is an appropriate procedure. To check if $x_i - \bar{x}_\pi$ can be used as an instrument, the vector $\hat{\gamma}$ of (6.4.24) is computed,

$$\begin{aligned} \hat{\gamma}' &= (12.976,\ 0.977,\ 0.008), \\ &\quad\ (0.049)\ (0.032)\ (0.022) \end{aligned}$$

where the coefficients are for $(1, x_i - \bar{x}_n, r_{2i})$, r_{2i} is the deviation from the regression of $x_i - \bar{x}_\pi$ on $\mathbf{z}_{1i}$, and the standard errors are the square roots of the diagonal elements of (6.4.25). Because the test statistic for γ_3 is 0.341, we easily accept the hypothesis that $x_i - \bar{x}_\pi$ can be used as an instrument. The EGLS instrumental variable estimator with

$$\mathbf{z}_i = [w_i, w_i(x_i - \bar{x}_n), \hat{h}_i w_i, \hat{h}_i w_i(x_i - \bar{x}_n), x_i - \bar{x}_\pi] \tag{6.4.31}$$

is

$$\begin{aligned} \hat{y}_i &= 12.977 + 0.976(x_i - \bar{x}_n), \\ &\quad\ (0.041)\ (0.030) \end{aligned}$$

where the standard errors are from an estimated covariance matrix of the form (6.4.30) using $\mathbf{z}_i$ of (6.4.31).

In this example there is a very strong correlation between x and w. Hence, the addition of $x_i - \bar{x}_\pi$ to the set of instruments results in a modest reduction in the standard error. ■■

6.5 NONLINEAR MODELS

Many of the procedures of the preceding sections can be extended to nonlinear models and to other estimation procedures. Consider a problem for which maximum likelihood estimation is appropriate for simple random samples. Assume that the subject matter analyst specifies a population density function $f(y, \boldsymbol{\theta})$, where the parameter $\boldsymbol{\theta}$ is of interest. Given a random sample from $f(y, \boldsymbol{\theta})$, the maximum likelihood estimator of $\boldsymbol{\theta}$ is obtained by maximizing

$$L(\boldsymbol{\theta}) = \sum_{i \in A} \log f(y_i, \boldsymbol{\theta}) \tag{6.5.1}$$

with respect to $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_S$ be the value of $\boldsymbol{\theta}$ that maximizes (6.5.1).

A direct extension of (6.5.1) to an unequal probability sample is obtained by weighting the elements of the likelihood function by the inverses of the selection probabilities. Assume that a sample is selected with probabilities π_i from a finite population, where the finite population is a sample of independent identically distributed observations generated from $f(y, \boldsymbol{\theta})$. Then the weighted log-likelihood is

$$L_\pi(\boldsymbol{\theta}) = \sum_{i \in A} \pi_i^{-1} \log f(y_i, \boldsymbol{\theta}). \tag{6.5.2}$$

If the sample design is such that selection of y_i is independent of the selection of y_j for all $i \neq j$, (6.5.2) is proportional to the probability of observing the set of sample y values. One might choose to construct an estimator of $\boldsymbol{\theta}$ by maximizing (6.5.2) whether or not the independence assumption holds. Therefore, we call the function (6.5.2) a weighted log-likelihood whether or not the sample design produces independent observations. Let $\hat{\boldsymbol{\theta}}_\pi$ be the value of $\boldsymbol{\theta}$ that maximizes (6.5.2), and let $\boldsymbol{\theta}_N$ be the value of $\boldsymbol{\theta}$ that maximizes

$$\sum_{i \in U} \log f(y_i, \boldsymbol{\theta}). \tag{6.5.3}$$

One can view $\hat{\boldsymbol{\theta}}_\pi$ as an estimator of $\boldsymbol{\theta}_N$, where $\boldsymbol{\theta}_N$ is the estimator that one would obtain if one applied maximum likelihood to the finite population. To formalize this view, assume that $\log f(y_i, \boldsymbol{\theta})$ has continuous first and second derivatives, let $\mathbf{b}(y_i, \hat{\boldsymbol{\theta}}) = \partial \log f(y_i, \hat{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}$ be the vector of partial derivatives evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and let

$$\mathbf{H}(y_i, \hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \log f(y_i, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}$$

be the negative of the matrix of second partial derivatives evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. The $\hat{\boldsymbol{\theta}}_S$ that maximizes the unweighted likelihood of (6.5.1) satisfies

$$\sum_{i \in A} \mathbf{b}(y_i, \hat{\boldsymbol{\theta}}_S) = \mathbf{0}, \tag{6.5.4}$$

the $\boldsymbol{\theta}_N$ that maximizes (6.5.3) satisfies

$$\sum_{i \in U} \mathbf{b}(y_i, \boldsymbol{\theta}_N) = \mathbf{0}, \tag{6.5.5}$$

and the $\hat{\boldsymbol{\theta}}_\pi$ that maximizes the weighted likelihood of (6.5.2) satisfies

$$\sum_{i \in A} \pi_i^{-1} \mathbf{b}(y_i, \hat{\boldsymbol{\theta}}_\pi) = \mathbf{0}. \tag{6.5.6}$$

Equations such as (6.5.4) and (6.5.6) are estimating equations of the type studied in Section 1.3.4.

Assume that the sampling design is such that Horvitz–Thompson estimators are design consistent. Then the error in $\hat{\boldsymbol{\theta}}_\pi$ as an estimator of the true value $\boldsymbol{\theta}^\circ$ is

$$\begin{aligned} \hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}^\circ &= \left(\sum_{i \in A} \pi_i^{-1} \mathbf{H}(y_i, \hat{\boldsymbol{\theta}}_\pi) \right)^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}(y_i, \hat{\boldsymbol{\theta}}_\pi) + O_p(n^{-1}) \\ &= \left(\sum_{i \in U} \mathbf{H}(y_i, \boldsymbol{\theta}^\circ) \right)^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}(y_i, \boldsymbol{\theta}^\circ) + O_p(n^{-1}). \end{aligned} \tag{6.5.7}$$

See Theorem 1.3.9. The finite population vector $\boldsymbol{\theta}_N$ satisfies

$$\boldsymbol{\theta}_N - \boldsymbol{\theta}^\circ = \left(\sum_{i \in U} \mathbf{H}(y_i, \boldsymbol{\theta}^\circ) \right)^{-1} \sum_{i \in U} \mathbf{b}(y_i, \boldsymbol{\theta}^\circ) + O_p(N^{-1}) \tag{6.5.8}$$

and from (6.5.7) and (6.5.8),

$$\begin{aligned} \hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_N &= \left(\sum_{i \in U} \mathbf{H}(y_i, \boldsymbol{\theta}^\circ) \right)^{-1} \left(\sum_{i \in A} \pi_i^{-1} \mathbf{b}(y_i, \boldsymbol{\theta}^\circ) - \sum_{i \in U} \mathbf{b}(y_i, \boldsymbol{\theta}^\circ) \right) \\ &\quad + O_p(n^{-1}). \end{aligned} \tag{6.5.9}$$

Let the covariance matrix of the approximate distribution of $\hat{\boldsymbol{\theta}}_S$ for a simple random sample of size m be $V_m\{\hat{\boldsymbol{\theta}}_S\}$, and let $V\{\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_N \mid \mathcal{F}\}$ be the conditional covariance matrix of the approximate distribution of $\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_N$. Then, under the assumption that the finite population is a simple random sample from the infinite population and that the variances are defined, the variance of $\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}^\circ$ is

$$V\{\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}^\circ\} = V_N\{\boldsymbol{\theta}_N\} + E\{V(\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_N \mid \mathcal{F})\}. \tag{6.5.10}$$

The variance of the approximate conditional distribution of $\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_N$ can be estimated by

$$\hat{V}\{\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_N \mid \mathcal{F}\} = \hat{\mathbf{T}}_H^{-1} \hat{V} \left\{ \sum_{i \in A} \pi_i^{-1} \mathbf{b}_i \mid \mathcal{F} \right\} \hat{\mathbf{T}}_H^{-1}, \tag{6.5.11}$$

where

$$\hat{\mathbf{T}}_H = \sum_{i \in A} \pi_i^{-1} \mathbf{H}(y_i, \hat{\boldsymbol{\theta}}_\pi)$$

and $\hat{V}\{\Sigma_{i \in A} \pi_i^{-1} \mathbf{b} \mid \mathcal{F}\}$ is the Horvitz–Thompson variance estimator calculated with $\hat{\mathbf{b}}_i = \mathbf{b}(y_i, \hat{\boldsymbol{\theta}}_\pi)$ replacing $\mathbf{b}_i = \mathbf{b}(y_i, \boldsymbol{\theta}_N)$. An estimator of the variance of the approximate distribution of $\hat{\boldsymbol{\theta}}_\pi$ as an estimator of $\boldsymbol{\theta}^\circ$ is

$$\begin{aligned} \hat{V}\{\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}^\circ\} &= \hat{V}\{\boldsymbol{\theta}_N\} + \hat{V}\{\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_N \mid \mathcal{F}\} \\ &= N \hat{\mathbf{T}}_H^{-1} \hat{\boldsymbol{\Sigma}}_{bb} \hat{\mathbf{T}}_H^{-1} + \hat{V}\{\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_N \mid \mathcal{F}\}, \quad (6.5.12) \end{aligned}$$

where

$$\hat{\boldsymbol{\Sigma}}_{bb} = N^{-1} \sum_{i \in A} \pi_i^{-1} \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i'.$$

A question addressed in Section 6.3.1 is: "Do I need to include the weights?" or alternatively, "Is the estimator that maximizes (6.5.1) estimating the same quantity as the estimator that maximizes (6.5.2)?" A probabilistic answer to the question can be obtained by constructing a test similar to that developed in Section 6.3.1.

The Newton–Raphson iterative procedure for constructing the estimator defined by (6.5.4) uses a trial value and then constructs an estimated change as

$$\tilde{\boldsymbol{\delta}} = \left(\sum_{i \in A} \tilde{\mathbf{H}}_i \right)^{-1} \sum_{i \in A} \tilde{\mathbf{b}}_i, \quad (6.5.13)$$

where $\tilde{\mathbf{b}}_i = \mathbf{b}(y_i, \tilde{\boldsymbol{\theta}})$, $\tilde{\mathbf{H}}_i = \mathbf{H}(y_i, \tilde{\boldsymbol{\theta}})$, $\tilde{\boldsymbol{\theta}}$ is the trial value, and $\tilde{\boldsymbol{\theta}} + \tilde{\boldsymbol{\delta}}$ is the value for the next iteration. If $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_s$, then $\tilde{\boldsymbol{\delta}} = \mathbf{0}$. We build on this approach to construct a test. Let $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\delta}}_1', \hat{\boldsymbol{\delta}}_2')'$ be defined by

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\delta}}_1 \\ \hat{\boldsymbol{\delta}}_2 \end{pmatrix} &= \begin{pmatrix} \sum_{i \in A} \tilde{\mathbf{H}}_i & \sum_{i \in A} w_i \tilde{\mathbf{H}}_i \\ \sum_{i \in A} w_i \tilde{\mathbf{H}}_i & \sum_{i \in A} w_i^2 \tilde{\mathbf{H}}_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i \in A} \tilde{\mathbf{b}}_i \\ \sum_{i \in A} w_i \tilde{\mathbf{b}}_i \end{pmatrix} \\ &=: \tilde{\mathbf{M}}_{HH}^{-1} \sum_{i \in A} \tilde{\mathbf{q}}_i, \quad (6.5.14) \end{aligned}$$

where $w_i = \pi_i^{-1}$ and $\tilde{\mathbf{q}}_i = (\tilde{\mathbf{b}}_i', w_i \tilde{\mathbf{b}}_i')'$. An estimator of the variance of $\hat{\boldsymbol{\delta}}$ is

$$\hat{\mathbf{V}}_{\delta\delta} = \hat{V}\{\hat{\boldsymbol{\delta}}\} = \tilde{\mathbf{M}}_{HH}^{-1} \hat{\mathbf{V}}_{qq} \tilde{\mathbf{M}}_{HH}^{-1}, \quad (6.5.15)$$

where $\hat{\mathbf{V}}_{qq}$ is the estimated variance of $\Sigma_{i\in A}\mathbf{q}_i$. If the design is such that we can treat y_i, $i \in A$, as independent of y_j, $j \in A$, for $i \neq j$,

$$\hat{\mathbf{V}}_{qq} = \hat{V}\left\{\sum_{i\in A}\mathbf{q}_i\right\} = n(n-2r)^{-1}\sum_{i\in A}\tilde{\mathbf{q}}_i\tilde{\mathbf{q}}_i',$$

where r is the dimension of $\boldsymbol{\theta}$, is a possible estimator. For a general design, $\hat{\mathbf{V}}_{qq}$ will be the estimator appropriate for the design.

A test of the hypothesis that $E(\hat{\boldsymbol{\theta}}_S) = E(\hat{\boldsymbol{\theta}}_\pi)$ is

$$F(r,c) = k^{-1}\hat{\boldsymbol{\delta}}_2'\hat{\mathbf{V}}_{\delta\delta,22}^{-1}\hat{\boldsymbol{\delta}}_2, \tag{6.5.16}$$

where $\hat{\mathbf{V}}_{\delta\delta,22}$ is the lower right $r \times r$ submatrix of $\hat{\mathbf{V}}_{\delta\delta}$ of (6.5.15), $c = n - 2r$ for a simple random sample, and c is the number of primary sampling units less the number of strata for a stratified sample. Under the null model, the test statistic is approximately distributed as F with r and c degrees of freedom. If r is large, a test can be constructed using a subset of variables.

Instrumental variables can also be used for nonlinear models. An important model is that in which the expected value of y is an explicit nonlinear function. Let

$$\begin{aligned} y_i &= g(\mathbf{x}_i, \boldsymbol{\theta}) + e_i, \\ e_i &\sim ind(0, \sigma^2), \end{aligned} \tag{6.5.17}$$

where $g(\mathbf{x}_i, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ with continuous first and second derivatives and e_i is independent of $\mathbf{x}_j$ for all i and j, $i \neq j$. The instrumental variable estimator is developed as an extension to the equation associated with least squares estimation, where the least squares estimator for $\boldsymbol{\theta}$ is the $\boldsymbol{\theta}$ that minimizes

$$Q(\boldsymbol{\theta}) = \sum_{i\in A}[y_i - g(\mathbf{x}_i, \boldsymbol{\theta})]^2. \tag{6.5.18}$$

In the Gauss–Newton method for finding the minimum of $Q(\boldsymbol{\theta})$, the function $g(\mathbf{x}, \boldsymbol{\theta})$ is expanded about a trial value, $\tilde{\boldsymbol{\theta}}$, to obtain the approximation

$$Q(\boldsymbol{\theta}) \doteq \sum_{i\in A}[y_i - g(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}) - k(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})\boldsymbol{\delta}]^2, \tag{6.5.19}$$

where $k(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}) = \partial g(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}'$ and $\boldsymbol{\delta} = \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}$. The improved estimator of $\boldsymbol{\theta}$ is $\tilde{\boldsymbol{\theta}} + \hat{\boldsymbol{\delta}}$, where

$$\hat{\boldsymbol{\delta}} = \left(\sum_{i\in A}\tilde{\mathbf{k}}_i'\tilde{\mathbf{k}}_i\right)^{-1}\sum_{i\in A}\tilde{\mathbf{k}}_i'[y_i - g(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})], \tag{6.5.20}$$

and $\tilde{\mathbf{k}}_i =: k(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})$. The procedure can be iterated to obtain the $\tilde{\boldsymbol{\theta}}$ that minimizes $Q(\boldsymbol{\theta})$. A modification may be required to assure convergence. See Fuller (1996, p. 272).

One step of an instrumental variable estimator is obtained by replacing $\tilde{\mathbf{k}}_i'$ with a vector of instrumental variables, $\mathbf{z}_i$, to obtain

$$\sum_{i \in A} \mathbf{z}_i' \tilde{\mathbf{k}}_i \hat{\tilde{\boldsymbol{\delta}}}_{IV} = \sum_{i \in A} \mathbf{z}_i' [y_i - g(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})]. \tag{6.5.21}$$

If the dimension of $\mathbf{z}_i$ is greater than that of $\tilde{\mathbf{k}}_i$, the estimator of $\boldsymbol{\delta}$ analogous to (6.4.9) is

$$\hat{\tilde{\boldsymbol{\delta}}}_{IV} = [(\mathbf{Z}'\tilde{\mathbf{K}})'\tilde{\mathbf{V}}_{bb}^{-1}(\mathbf{Z}'\tilde{\mathbf{K}})]^{-1}(\mathbf{Z}'\tilde{\mathbf{K}})'\tilde{\mathbf{V}}_{bb}^{-1}\mathbf{Z}'\tilde{\mathbf{e}}, \tag{6.5.22}$$

where $\tilde{\mathbf{V}}_{bb}$ is an estimator of the variance of $\tilde{\mathbf{b}} = \mathbf{Z}'\mathbf{e}$, the ith row of $\mathbf{Z}$ is $\mathbf{z}_i$, the ith row of $\tilde{\mathbf{K}}$ is $\tilde{\mathbf{k}}_i$ and $\tilde{\mathbf{e}} = (\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_n)'$. If the trial value $\tilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$, a single step estimator, $\tilde{\boldsymbol{\theta}} + \hat{\tilde{\boldsymbol{\delta}}}_{IV}$ will have the same asymptotic properties as a multiple-step estimator. If desired, the procedure can be iterated until $\hat{\tilde{\boldsymbol{\delta}}}_{IV}$ of (6.5.21) is the zero vector. The variance of $\hat{\boldsymbol{\theta}}_{IV}$ can be estimated with

$$\hat{V}\{\hat{\boldsymbol{\theta}}_{IV}\} = \hat{V}\{\boldsymbol{\theta}_N\} + \hat{V}\{\hat{\boldsymbol{\theta}}_{IV} - \boldsymbol{\theta}_N \mid \mathcal{F}\},$$

where

$$\hat{V}\{\hat{\boldsymbol{\theta}}_{IV} - \boldsymbol{\theta}_N \mid \mathcal{F}\} = [(\mathbf{Z}'\tilde{\mathbf{K}})'\tilde{\mathbf{V}}_{bb}^{-1}(\mathbf{Z}'\tilde{\mathbf{K}})]^{-1}.$$

6.6 CLUSTER AND MULTISTAGE SAMPLES

The correlation associated with clustering of sample elements is a dimension of design that has a universal impact on estimation properties. The elements within a primary sampling unit are almost always positively correlated, and this correlation must be recognized in variance estimation.

A potential regression model for cluster samples is given in (2.6.6). In practice, the model might need to be extended to permit different slopes in different clusters and (or) different variances in different clusters. Because of the difficulty in constructing models for survey clustered data, a common procedure is to employ a consistent estimator and then estimate the variance recognizing the cluster effects. Let y_{ij} denote the jth element in the ith primary sampling unit, and let $\pi_{(ij)}$ be the probability that second-stage unit ij is selected for the sample. Assume that it is desired to estimate the $\boldsymbol{\beta}$ of the linear model,

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + e_{ij}. \tag{6.6.1}$$

Assume that an estimator of the form

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y} \tag{6.6.2}$$

is a consistent estimator, where $\mathbf{y}$ is the vector of observations and $\boldsymbol{\Phi}$ is a diagonal matrix. For example, $\boldsymbol{\Phi}^{-1}$ might be $\mathbf{D}_{\pi}^{-1}\hat{\boldsymbol{\Psi}}$ of (6.3.45). If the finite population is a sample of primary sampling units from an infinite population of primary sampling units, the error in $\hat{\boldsymbol{\beta}}$ of (6.6.2) is

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{e} \\ &= \sum_{ij\in A} \mathbf{b}_{ij},\end{aligned}$$

where $\mathbf{b}_{ij} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{x}'_{ij}\phi_{ij}^{-1}\hat{e}_{ij}$ and ϕ_{ij} is the ijth diagonal element of $\boldsymbol{\Phi}$. Then an estimator of the design variance of $\hat{\boldsymbol{\beta}}$ is a variance estimator of the sum, such as the Horvitz–Thompson variance estimator, computed with $\hat{\mathbf{b}}_{ij} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{x}'_{ij}\phi_{ij}^{-1}\hat{e}_{ij}$ and $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}$.

Example 6.6.1. To illustrate the effect of clustering, we use the analysis of a sample of Canadian workplaces conducted by Zdenek Patak. The sample is composed of 5781 employees in 1389 workplaces interviewed in 1999. A logistic function was fit to estimate the probability that a person received classroom training in the last 12 months. The variables are:

$y = 1$ if employee received training
$= 0$ otherwise,
$x_1 = 1$ if gender is male
$= 0$ if gender is female,
$x_2 = 1$ if full-time employee
$= 0$ if part-time employee,
$x_3 = 1$ if employee has some university education
$= 0$ otherwise,
$x_4 = 1$ if workplace adopted an innovation in last 12 months
$= 0$ otherwise,
$x_5 = 1$ if size of workplace is $<= 15$ employees
$= 0$ otherwise,
$x_6 = 1$ if size of workplace is $<= 60$ employees
$= 0$ otherwise.

The conditional expected value of y under the logistic model is

$$E\{y_i \mid \mathbf{x}_i\} = (1 + exp\{\mathbf{x}_i\boldsymbol{\beta}\})^{-1}exp\{\mathbf{x}_i\boldsymbol{\beta}\},$$

where $\mathbf{x}_i = (1, x_{1i}, x_{2i}, \ldots, x_{6i})$. The parameter vector was estimated by maximizing the weighted likelihood, where the weighted likelihood is the likelihood constructed as if the observations are independent. The estimated vector is

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_6) = (-0.971, -0.142, 0.497, 0.627, 0.282, -1.096, -0.572).$$

Although the sample is an unequal probability sample, the estimation procedure assigned equal weight to all observations. The standard errors were estimated by two procedures. The first set of estimated standard errors are those constructed as if the sampling units are the individual employees and are $(0.123, 0.056, 0.107, 0.060, 0.060, 0.079, 0.069)$. The second estimation procedure used the correct design specification, in which the workplaces are the primary sampling units. The estimated standard errors using the workplace as the primary sampling unit are (0.150, 0.061, 0.124, 0.063, 0.079, 0.096, 0.091).

The differences between the two sets of estimated standard errors are fairly typical for cluster samples. Although it is possible for standard errors for a cluster sample to be smaller than those for an element sample, in the majority of cases the standard errors for a cluster sample are larger than those for an element sample. This example is typical in that the correctly estimated standard errors are larger than the biased estimates based on elements. Also, the bias in the standard error of the intercept is generally larger than the bias in coefficients measured on elements. In this case the bias in the standard error of the intercept is estimated to be about 22%, while the average bias in the standard errors for the three element variables is about 9%. The bias in the standard errors is often largest for the coefficients of variables that are measured on large units, in our case on the primary sampling units. The average bias for the three workplace variables is about 29%, corresponding to a 66% bias in the estimated variance.

The estimated vector constructed using the sampling weights is

$$\begin{aligned} \hat{\boldsymbol{\beta}}' \;=\; & (-1.351, -0.159, \;\; 0.690, \;\; 0.785, \;\; 0.376, -0.980, -0.639), \\ & \;\; (0.242) \;\; (0.120) \; (0.175) \; (0.123) \; (0.155) \;\; (0.164) \;\; (0.177) \end{aligned}$$

where the standard errors are computed under the correct design specification. The estimated standard errors are much larger for the weighted fit than for the unweighted fit. The test of the hypothesis that the weighted and unweighted estimators are estimating the same quantity is $F(7, 342) = 1.19$. Thus, the null hypothesis is easily accepted and the unweighted estimators are chosen as the final estimates. ■■

6.7 PRETEST PROCEDURES

In Section 6.3 we discussed procedures in which a test of model assumptions was used to decide between two estimation procedures. In practice, the preliminary test may have a subject matter context. For example, the subject matter specialist may have begun the analysis with a theory that stated that the mean of stratum 1 should be equal to the mean of stratum 2. If that is so, the test that the two stratum means are equal is a test of the theory as well as a part of the estimation procedure.

Given that the model is accepted by the test, an estimator is constructed under the model assumptions. The test is often called a *pretest*. Procedures based on a pretest were studied by Bancroft (1944) and Huntsberger (1955). See also Brown (1967), Gregoire, Arabatzis, and Reynolds (1992), Albers, Boon, and Kallenberg (2000), and Saleh (2006). Because of the frequent use of procedures of this type, we illustrate the properties of the procedure using a simulation experiment conducted by Wu (2006).

Samples were created using 400 trials of the following selection procedure. A vector (e_i, x_i, u_i) is selected, where e_i is distributed as the part of the normal distribution on $[-2,\ 2]$, x_i is a uniform $(0, 1)$ random variable, and u_i is a uniform $(0, 1)$ random variable. The e_i, x_i, and u_i are mutually independent. Let

$$\pi_i = (0.01 + 0.99x_i)(1 - \xi) + \xi(2.0 + e_i), \tag{6.7.1}$$

where ξ is a parameter that is varied in the experiment. If $u_i \leq \pi_i$, the vector (e_i, x_i, u_i) is accepted and y_i is defined by

$$y_i = 4.0 + e_i. \tag{6.7.2}$$

If $u_i > \pi_i$, the vector is rejected. The size of the sample is random with an expected size of 200 elements.

Estimators of the population mean were constructed using a pretest procedure to decide between the simple mean $\bar{y}$ and the weighted mean $\bar{y}_\pi$. First, the y_i values were regressed on the vector $(1, \pi_i^{-1})$ to obtain the coefficient vector

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)' = \left(\sum_{i \in A}(1, \pi_i^{-1})'(1, \pi_i^{-1})\right)^{-1} \sum_{i \in A}(1, \pi_i^{-1})' y_i. \tag{6.7.3}$$

The vector $(1, \pi_i^{-1})$ corresponds to $(\mathbf{x}_i, \mathbf{z}_i)$ of (6.3.12). Based on the regression, the test statistic is

$$t_{\beta_1} = [\hat{V}\{\hat{\beta}_1\}]^{-1/2}\hat{\beta}_1, \tag{6.7.4}$$

where

$$\hat{V}\{\hat{\beta}\} = \left(\sum_{i\in A}(1,\pi_i^{-1})'(1,\pi_i^{-1})\right)^{-1} s^2$$

and

$$s^2 = (n-2)^{-1}\sum_{i\in A}[y_i - (1,\pi_i^{-1})\hat{\beta}]^2.$$

Then the pretest estimator of μ is

$$\begin{aligned}\tilde{\mu} &= \bar{y} \quad \text{if } \mid t_{\beta_1} \mid < t_\gamma \\ &= \bar{y}_\pi \quad \text{if } \mid t_{\beta_1} \mid \geq t_\gamma,\end{aligned} \tag{6.7.5}$$

where

$$\bar{y} = n^{-1}\sum_{i\in A} y_i,$$

$$\bar{y}_\pi = \left(\sum_{i\in A}\pi_i^{-1}\right)^{-1}\sum_{i\in A}\pi_i^{-1}y_i,$$

and t_γ is the γ quantile of Student's t. The value $\gamma = 0.05$ was used in the simulation.

Table 6.4 Monte Carlo Results for Size 0.05 Pretest Estimator (10,000 Samples)

ξ	$\frac{Bias(\bar{y})}{S.E.(\bar{y})}$	$\frac{M.S.E.(\bar{y})}{V(\bar{y})}$	$\frac{V(\bar{y}_\pi)}{M.S.E.(\bar{y})}$	$\frac{Bias(\tilde{\mu})}{S.E.(\bar{y})}$	$\frac{M.S.E.(\tilde{\mu})}{M.S.E.(\bar{y}_\pi)}$
0.00	0.000	1.000	2.210	0.000	0.594
0.02	0.112	1.013	1.948	0.078	0.638
0.04	0.247	1.061	1.750	0.186	0.706
0.06	0.376	1.141	1.579	0.284	0.758
0.10	0.619	1.383	1.178	0.423	0.970
0.14	0.861	1.741	0.896	0.484	1.175
0.17	1.043	2.087	0.750	0.456	1.298
0.20	1.229	2.510	0.600	0.386	1.376
0.23	1.437	3.065	0.483	0.302	1.332
0.27	1.683	3.831	0.384	0.125	1.172
0.30	1.838	4.380	0.327	0.041	1.063
0.40	2.472	7.112	0.203	0.000	1.000
0.50	3.171	11.055	0.135	0.000	1.000

Results for 10,000 samples for the size 0.05 pretest are given in Table 6.4. If the ξ of (6.7.1) is zero, $\bar{y}$ is unbiased for the mean because π_i is independent

of $\bar{y}$. For any positive ξ, π_i and e_i are correlated and the simple mean of the observed sample is biased for the population mean. The weighted sample mean $\bar{y}_\pi$ is a consistent estimator of the population mean and is essentially unbiased for all ξ. For $\xi = 0$, the weights are independent of y_i, and the simple mean is twice as efficient as $\bar{y}_\pi$ for the population mean. As ξ increases, the relative efficiency of $\bar{y}$ declines rapidly. At $\xi = 0.14$ the bias in $\bar{y}$ is 86% of the standard deviation, and the mean square error of $\bar{y}$ is 10% larger than the variance of $\bar{y}_\pi$.

The variance of the pretest estimator $\tilde{\mu}$ is about 30% greater than that of $\bar{y}$ at $\xi = 0$ because $\bar{y}_\pi$ is being used as the estimator about 5% of the time, and those samples are often samples with large deviations. On the other hand, $\tilde{\mu}$ is about 68% more efficient than $\bar{y}_\pi$ when $\xi = 0$. As ξ increases, the efficiency of $\tilde{\mu}$ declines relative to $\bar{y}_\pi$, reaching a minimum of about 72% for ξ near 0.20. As ξ increases further, the pretest rejects the null model more frequently, and $\tilde{\mu}$ is essentially equal to $\bar{y}_\pi$ for $\xi \geq 0.40$. The shape of the relative mean square error of $\tilde{\mu}$ relative to $\bar{y}_\pi$ as a function of ξ is typical of pretest procedures. If the null condition ($\xi = 0$) holds, the pretest procedure is better than the alternative ($\bar{y}_\pi$) but not as good as the estimator constructed under the null model. As the parameter ξ moves away from the null model, the pretest procedure improves relative to the null model estimator and declines relative to the (nearly) unbiased estimator $\bar{y}_\pi$. For very large ξ, the null model is almost always rejected and the pretest procedure is essentially equivalent to $\bar{y}_\pi$.

The standard error for $\tilde{\mu}$ computed using the variance estimation procedure appropriate for the estimator chosen is

$$\begin{aligned} \hat{V}\{\tilde{\mu}\} &= n^{-1}s^2 \qquad \text{if } \mid t_{\beta_1} \mid < t_\gamma \\ &= \hat{V}\{\bar{y}_\pi\} \qquad \text{if } \mid t_{\beta_1} \mid \geq t_\gamma, \end{aligned}$$

where

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

and

$$\hat{V}\{\bar{y}_\pi\} = n(n-1)^{-1} \left(\sum_{i=1}^{n} \pi_i^{-1} \right)^{-2} \sum_{i=1}^{n} \pi_i^{-2} (y_i - \bar{y}_\pi)^2.$$

We call the statistic

$$\tilde{t}_\mu = [\hat{V}\{\tilde{\mu}\}]^{-1/2} (\tilde{\mu} - \mu)$$

the t-statistic for $\tilde{\mu}$, recognizing that the distribution is not that of Student's t. As the simulation results of Table 6.5 illustrate, the statistic has a variance greater than 1 for all ξ values studied. Hence, the statistic exceeds the

Table 6.5 Monte Carlo Properties of t-Statistic for Pretest Estimator

ξ	Mean	Variance	$P\{\tilde{t}_\mu > t_{0.05}\}$
0.00	0.00	1.09	0.060
0.02	0.09	1.08	0.060
0.04	0.21	1.11	0.068
0.06	0.31	1.10	0.071
0.10	0.46	1.17	0.093
0.14	0.54	1.33	0.117
0.17	0.53	1.49	0.140
0.20	0.46	1.56	0.137
0.23	0.36	1.50	0.124
0.27	0.18	1.30	0.093
0.30	0.09	1.14	0.065
0.40	0.07	1.06	0.055
0.50	0.09	1.06	0.056

tabular value for Student's t by more than the nominal fraction. The poorest performance occurs near $\xi = 0.20$, where the ratio of the mean square error of $\tilde{\mu}$ to the mean square error of $\bar{y}_\pi$ is largest.

6.8 REFERENCES

Section 6.1. Deming (1953), Deming and Stephan (1941).

Section 6.2. Chambers and Skinner (2003), Hartley and Sielken (1975), Korn and Graubard (1995a, 1999).

Section 6.3. Beaumont (2008), Binder and Roberts (2003), Breckling et al. (1994), DuMouchel and Duncan (1983), Fuller (1984), Godambe and Thompson (1986), Graubard and Korn (2002), Holt, Smith, and Winter (1980), Huang and Hidiroglou (2003), Korn and Graubard (1995a, 1995b, 1999), Magee (1998), Pfeffermann (1993), Pfeffermann and Sverchkov (1999), Pfeffermann et al. (1998), Skinner (1994).

Section 6.4. Wooldridge (2006), Wu and Fuller (2006).

Section 6.5. Binder (1983), Chambers and Skinner (2003), Fuller (1996), Godambe (1991), Godambe and Thompson (1986).

Section 6.6. Bancroft (1944), Bancroft and Han (1977), Gregoire, Arabatzes, and Zieschang (1992), Huntsberger (1955), Saleh (2006), Wu (2006).

Section 6.7. Wu (2006).

6.9 EXERCISES

1. (Section 6.2) What is the bias of estimator (6.2.20) as an estimator of S_y^2?

2. (Section 6.3) Consider the regression equation (6.3.12) used to test the hypothesis that the ordinary least squares estimator is unbiased for $\boldsymbol{\beta}$. Let $\mathbf{x}_i = (1, x_{1i})$. Assume that π_i is a multiple of $(a + e_i)^{-1}$, where a is a positive constant, e_i is distributed on the interval (b_1, b_2), and $a + b_1 > 0$. What would be the residual mean square for regression (6.3.12)?

3. (Section 6.2) Show that the error in the estimator S_y^2 of (6.2.20) as an estimator of S_y^2 is $O_p(n^{-1/2})$ for populations with a fourth moment.

4. (Section 6.4) Show that the instrumental variable estimator of $\boldsymbol{\beta}_1$ of (6.4.18) can be written as the ordinary least squares regression of $y_i - \bar{y}_\pi$ on $\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi}$, where $\boldsymbol{\beta}' = (\beta_0, \boldsymbol{\beta}_1')$.

5. (Section 6.3) Let a Poisson sample be selected from the finite population $\{\mathcal{F}_N\} = \{(x_{11}, y_1), (x_{12}, y_2), \ldots, (x_{1N}, y_N)\}$, where the (x_{1i}, y_i) are iid random variables with sixth moments. Let model (6.3.36) hold, where e_i is independent of x_{1j} for all i and j. Show that

$$V\left\{\sum_{i\in A} \pi_i^{-1} x_{1i}\psi_i e_i\right\} = NE\{\pi_i^{-1}\psi_i^2 x_{1i}^2 e_i^2\},$$

given that (π_i, ψ_i) is a function of x_{1i} with the requisite moments. It is to be understood that $E\{\pi_i^{-1}\psi_i^2 x_{1i}^2 e_i^2\}$ is the superpopulation mean of $\pi_i^{-1}\psi_i^2 x_{1i}^2 e_i^2$.

6. (Section 6.3) Let $\mathcal{F}_N = (z_1, z_2, \ldots, z_N)$ be a realization of $iid(\mu_z, \sigma_z^2)$ random variables. Let a Poisson sample be selected with probabilities π_i and let

$$\bar{z}_\pi = \left(\sum_{i\in A} \pi_i^{-1}\right)^{-1} \sum_{i\in A} \pi_i^{-1} z_i.$$

Using the Taylor approximation for the variance of a ratio, show that

$$V\{\bar{z}_\pi - \mu_z\} \doteq N^{-2} E\left\{\sum_{i=1}^{N} \pi_i^{-1}(z_i - \mu_z)^2\right\}.$$

7. (Section 6.3) Assume the superpopulation model

$$y_{hi} = \mathbf{x}_{hi}\boldsymbol{\beta} + e_{hi},$$

$$e_{hi} \sim ind(0, \sigma_i^2),$$

where e_{hi} is independent of $\mathbf{x}_{hi}$ and there is a fixed fraction W_h of the population in stratum h, $h = 1, 2, \ldots, H$. Assume that the population has eighth moments. Let a sequence of stratified samples be selected from a sequence of stratified finite populations satisfying the model. Let the sampling fraction $N_h^{-1} n_h$ in stratum h be fixed. Let

$$\tilde{\boldsymbol{\beta}}_{W\Psi} = \left(\sum_{hi\in A} \mathbf{x}'_{hi}\psi_{hi}\pi_h^{-1}\mathbf{x}_{hi}\right)^{-1} \sum_{hi\in A} \mathbf{x}'_{hi}\psi_{hi}\pi_h^{-1} y_{hi},$$

where ψ_{hi} is a known bounded function of $\mathbf{x}_{hi}$. Show that

$$\hat{\boldsymbol{\Sigma}}_{bb} = N^{-1} \sum_h N_h n_h^{-1} \sum_{i\in A_h} \mathbf{x}'_i \psi_i^2 \hat{e}_i^2 \mathbf{x}_i,$$

where $\hat{e}_i = y_i - \mathbf{x}_i\tilde{\boldsymbol{\beta}}_{W\Psi}$, is consistent for $\boldsymbol{\Sigma}_{bb}$ of (6.3.34).

8. (Section 6.3) In Example 6.3.2 the three types were assumed to have a common mean. Using the stratified means for the three types, test the hypothesis of common means.

9. (Section 6.3) Estimate the $\sigma_{ej}^2, j = 1, 2, 3$, of Example 6.3.2 by computing the domain means of $(y_{hji} - \hat{\mu}_{st})^2$. Estimate the variances of your estimators.

10. (Section 6.3) Construct the Pfeffermann–Sverchkov estimator for the sample of Example 6.3.2. Estimate the variance of the estimator. Why is the estimated efficiency relative to the stratified estimator poor?

REFERENCES

Albers, W., Boon, P.C. and Kallenberg, W.C.M. (2000). Size and power of pretest procedures. *The Annals of Statistics* **28**, 195–214.

Andersson, P.G. and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology* **31**, 95–99.

Bahadur, R.R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics* **37**, 577–580.

Bailar, B.A. (1968). Recent research in reinterview procedures. *Journal of the American Statistical Association* **63**, 41–63.

Bailar, B.A. and Bailar, J.C. (1983). Comparison of the biases of the hot-deck imputation procedure with an "equal-weights" imputation procedure. In: W.G. Madow and I. Olkin (eds.), *Incomplete Data in Sample Surveys*, Vol. 3. Academic Press, New York, pp. 299–311.

Bailar, B.A. and Dalenius, T. (1969). Estimating the response variance components of the US Bureau of the Census' Survey Model. *Sankhya B* **31**, 341–360.

Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* **15**, 190–204.

Bancroft, T.A. and Han, C.P. (1977). Inference based on conditional specification: a note and a bibliography. *International Statistical Review* **51**, 117–127.

Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician* **42**, 174–177.

Bardsley, P. and Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics* **33**, 290–299.

Basu, D. (1958). On sampling with and without replacement. *Sankhya* **20**, 287–294.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1. In: V.P. Godambe and D.A. Sprott (eds.), *Foundations of Statistical Inference,* Holt, Rinehart and Winston, Toronto, Ontario, Canada, pp. 203–242.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28–36.

Battese, G.E., Hasabelnaby, N.A. and Fuller, W.A. (1989). Estimation of livestock inventories using several area- and multiple-frame estimators. *Survey Methodology* **15**, 13–15.

Beale, E.M.L. (1962). Some uses of computers in operational research. *Industrielle Organisation* **31**, 51–52.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika* **95**, 539–553.

Bellhouse, D.R. (1977). Optimal designs for sampling in two dimensions. *Biometrika* **64**, 605–611.

Bellhouse, D.R. (1981). Spatial sampling in the presence of a trend. *Journal of Statistical Planning and Inference* **5**, 365–375.

Bellhouse, D.R. (1984). A review of optimal designs in survey sampling. *Journal of Statistics* **12**, 53–65.

Bellhouse, D.R. (1985). Computing methods for variance estimation in complex surveys. *Journal of Official Statistics* **1**, 323–329.

Bellhouse, D.R. (1988). Systematic sampling. In: P.R. Krishnaiah and C.R. Rao (eds.), *Handbook of Statistics*, Vol. 6. North-Holland, Amsterdam, pp. 125–145.

Bellhouse, D.R. (2000). Survey sampling theory over the twentieth century and its relation to computing technology. *Survey Methodology* **26**, 11–20.

Bellhouse, D.R. and Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica* **9**, 407–424.

Bellhouse, D.R. and Stafford, J.E. (2001). Local polynomial regression techniques in complex surveys. *Survey Methodology* **27**, 197–2003.

Berger, Y.G. and Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B* **67**, 79–89.

Bethel, J. (1989a). Sample allocation in multivariate surveys. *Survey Methodology* **15**, 47–57.

Bethel, J. (1989b). Minimum variance estimation in stratified sampling. *Journal of the American Statistical Association* **84**, 260–265.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* **4**, 251–260.

Bethlehem, J.G. and Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics* **3**, 141–153.

Bethlehem, J.G. and Kersten, H.M.P. (1985). On the treatment of nonresponse in sample surveys. *Journal of Official Statistics* **1**, 287–300.

Bickel, P.J. and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics* **12**, 470–482.

Bickel, P.J. and Krieger, A.M. (1989). Confidence bands for a distribution function. *Journal of the American Statistical Association* **84**, 95–100.

Biemer, P.P. (2004). An analysis of classification error for the revised Current Population Survey employment questions. *Survey Methodology* **30**, 127–140.

Biemer, P.P., Groves, R.M., Lyberg, L.L., Mathiowetz, N.A., and Sudman, S. (eds.) (1991). *Measurement Errors in Surveys*. Wiley, New York.

Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society B* **44**, 388–393.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279–292.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139–147.

Binder, D.A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology* **22**, 17–22.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M., and Jocelyn, W. (1997). Variance estimation for two-phase stratified sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 267–272.

Binder, D.A. and Hidiroglou, M.A. (1988). Sampling in time. In: P.R. Krishnaiah and C.R. Rao (eds.), *Handbook of Statistics*, Vol. 6. North-Holland, Amsterdam, pp. 187–211.

Binder, D.A. and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. In: R.L. Chambers and C.J. Skinner (eds.), *Analysis of Survey Data*. Wiley, New York, pp. 29–48.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Blight, B.J.N. (1973). Sampling from an autocorrelated finite population. *Biometrika* **60**, 375–385.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum liklihood inference from sample survey data. *International Statistical Review* **62**, 349–363.

Breidt, F.J. (1995). Markov chain designs for one-per-stratum sampling. *Survey Methodology* **2**, 63–70.

Breidt, F.J. and Fuller, W.A. (1993). Regression weighting for multiphase samples. *Sankhya B* **55**, 297–309.

Breidt, F.J. and Fuller, W.A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological and Environmental Statistics* **4**, 391–403.

Breidt, F.J., McVey, A. and Fuller, W.A. (1996). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics* (Golden Jubilee Number) **49**, 79–90.

Breidt, F.J. and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics* **28**, 1026–1053.

Breunig, R.V. (2001). Density estimation for clustered data. *Econometric Reviews* **20**, 353–367.

Brewer, K.R.W. (1963a). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics* **5**, 5–13.

Brewer, K.R.W. (1963b). Ratio estimation and finite population: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* **5**, 93–105.

Brewer, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association* **74**, 911–915.

Brewer, K.R.W. (1981). The analytical use of unequal probability samples: a case study. *Bulletin of the International Statistical Institute* **49**, 685–698.

Brewer, K.R.W. (1999). Design-based or prediction-based inference? Stratified random vs. stratified balanced sampling. *International Statistical Review* **67**, 35–47.

Brewer, K.R.W. (2002). *Combined Survey Sampling Inference: Weighing Basu's Elephants*. Oxford University Press, New York.

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer-Verlag, New York.

Brewer, K.R.W., Hanif, M. and Tam, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association* **83**, 128–132.

Brewer, K.R.W. and Mellor, R.W. (1973). The effect of sample structure on analytical surveys. *Australian Journal of Statistics* **15**, 145–152.

Brick, J.M., Kalton, G. and Kim, J.K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology* **30**, 57–66.

Brier, S.E. (1980). Analysis of contingency tables under cluster sampling. *Biometrika* **67**, 591–596.

Brown, L. (1967). The conditional level of Student's t test. *Annals of Mathematical Statistics* **38**, 1068–1071.

Bryant, E.C., Hartley, H.O. and Jessen, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association* **55**, 105–124.

Cannell, C., Fowler, F.J., Kalton, G., Oksenberg, L. and Bischoping, K. (1989). New quantitative techniques for presenting survey questions. *Bulletin of the International Statistical Institute* **53(2)**, 481–495.

Carroll, J.L. and Hartley, H.O. (1964). The symmetric method of unequal probability sampling without replacement. Abstract in *Biometrics* **20**, 908–909.

Carroll, R.J., Ruppert, D. and Stephanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–620.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. Wiley, New York.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1979). Prediction theory for finite populations when model-based and design-based principles are combined. *Scandinavian Journal of Statistics* **6**, 97–106.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In: W.G. Madow and I. Olkin (eds.), *Incomplete Data in Sample Surveys*, Vol. 3. Academic Press, New York, pp. 143–160.

Chakrabarti, M.C. (1963). On the use of incidence matrices in sampling from a finite population. *Journal of the Indian Statistical Association* **1**, 78–85.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics* **12**, 3–32.

Chambers, R.L., Dorfman, A.H. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society B* **60**, 397–412.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* **88**, 268–277.

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73**, 597–604.

Chambers, R.L. and Skinner, C.J. (eds.) (2003). *Analysis of Survey Data.* Wiley, Chichester, UK.

Chambless, L.E. and Boyle, K.E. (1985). Maximum likelihood methods for complex survey data: logistic regression and discrete proportional hazards models. *Communicatons in Statistics: Theory and Methods* **14**, 1377–1392.

Chang, T. and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95**, 555–571.

Chaudhuri, A. (1981). Non-negative unbiased variance estimators. In: D. Krewski, R. Platek, and J.N.K. Rao (eds.), *Current Topics in Survey Sampling*. Academic Press, New York, pp. 317–328.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. Marcel Dekker, New York.

Chaudhuri, A. and Vos, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. North-Holland, Amsterdam.

Chen, C., Fuller, W.A. and Breidt, F.J. (1999). A spline estimator of the distribution function of a variable measured with error. *Communicatons in Statistics: Theory and Methods* **29**, 1293–1310.

Chen, C., Fuller, W.A. and Breidt, F.J. (2002). Spline estimators of the density function of a variable measured with error. *Communications in Statistics: Simulation and Computation* **32**, 73–86.

Chen, J. and Rao, J.N.K. (2006). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*. To appear.

Chen, J. and Rao, J.N.K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica* **17**, 1047–1064.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics* **16**, 113–132.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association* **96**, 260–269.

Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica* **10**, 1153–1169.

Chen, S.X. (1998). Weighted polynomial models and weighted sampling schemes for finite population. *Annals of Statistics* **26**, 1894–1915.

Chen, S.X. (2000). General properties and estimation of conditional Bernoulli models. *Journal of Multivariate Analysis* **74**, 67–87.

Chen, S.X., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81**, 457–469.

Chua, T.C. and Fuller, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association* **82**, 46–51.

Chung, K.L. (1968). *A Course in Probability Theory*. Harcourt, Brace, and World. New York.

Cochran, W.G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association* **34**, 492–510.

Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association* **37**, 199–212.

Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics* **17**, 164–177.

Cochran, W.G. (1951). Modern methods in the sampling of human populations. *American Journal of Public Health* **41**, 647–653.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

Cohen, J.E. (1976). The distribution of the chi-squared statistic under clustered sampling from contingency tables. *Journal of the American Statistical Association* **71**, 665–670.

Cohen, S.B., Burt, V.L. and Jones, G.K. (1986). Efficiencies in variance estimation for complex survey data. *The American Statistician* **40**, 157–164.

Cohen, S.B., Xanthopoulos, J.A. and Jones, G.K. (1988). An evaluation of statistical software procedures appropriate for the regression analysis of complex survey data. *Journal of Official Statistics* **4**, 17–34.

Cook, J. and Stefanski, L.A. (1994). A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association* **89**, 1314–1328.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman & Hall, New York.

Cordy, C.B. and Thomas, D.R. (1997). Deconvolution of a distribution function. *Journal of the American Statistical Association* **92**, 1459–1465.

Cox, L.H. and Boruch, R.F. (1988). Record linkage, privacy and statistical policy. *Journal of Official Statistics* **4**, 3–16.

Cressie, N.A.C. (1991). *Statistics for Spatial Data.* Wiley, New York.

Dalén, J. (1986). Sampling from finite populations: actual coverage probabilities for confidence intervals on the population mean. *Journal of Official Statistics* **2**, 13–24.

Dalenius, T. (1983). Some reflections on the problem of missing data. In: W.G. Madow and I. Olkin (eds.), *Incomplete Data in Sample Surveys*, Vol. 3. Academic Press, New York, pp. 411–413.

Dalenius, T. and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association* **54**, 88–101.

Das Gupta, S. and Perlman, M.D. (1974). Power of the noncentral F-test: effect of additional variates on Hotelling's T^2-test. *Journal of the American Statistical Association* **69**, 174–180.

DaSilva, D.N. and Opsomer, J.D. (2004). Properties of the weighting cell estimator under a nonparametric response mechanism. *Survey Methodology* **30**, 45–66.

David, H.A. (1981). *Order Statistics*, 2nd ed. Wiley, New York.

David, H.A. and Nagaraja, H.N. (2003). *Order Statistics*, 3rd ed. Wiley, New York.

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *Annals of Statistics* **36**, 665–685.

Deming, W.E. (1943). *Statistical Adjustment of Data*. Wiley, New York.

Deming, W.E. (1950). *Some Theory of Sampling*. Wiley, New York.

Deming, W.E. (1953). On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias of non-response. *Journal of the American Statistical Association* **48**, 743–772.

Deming, W.E. (1956). On simplifications of sampling design through replication with equal probabilities and without stages. *Journal of the American Statistical Association* **51**, 24–53.

Deming, W.E. (1960). *Sample Design in Business Research*. Wiley, New York.

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* **11**, 427–444.

Deming, W.E. and Stephan, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association* **36**, 45–49.

Demnati, A. and Rao, J.N.K. (2004). Linearization variance estimators for survey data. (with discussion). *Survey Methodology* **30**, 17–34.

Deng, L.Y. and Wu, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association* **82**, 568–576.

DeStavola, B.L. and Cox, D.R. (2008). On the consequences of overstratificiation. *Biometrika* **95**, 992–996.

Deville, J.C., Grosbras, J.M. and Roth, N. (1988). Efficient sampling algorithms and balanced samples. *COMPSTAT 1988: Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg, Germany, pp. 255–256.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Deville, J.C., Särndal, C.E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Assocation* **88**, 1013–1020.

Deville, J.C. and Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika* **91**, 893–912.

Deville, J.C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference* **128**, 411-425.

Diana, G. and Tommasi, C. (2003). Optimal estimation for finite population mean in two-phase sampling. *Statistical Methods and Applications* **12**, 41–43.

Dippo, C.S. and Wolter, K.M. (1984). A comparison of variance estimators using the Taylor series approximation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 113–121.

Dorfman, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics* **35**, 29–41.

Dorfman, A.H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics* **21**, 1452–1475.

Doss, D.C., Hartley, H.O. and Somayajulu, G.R. (1979). An exact small sample theory for post-stratification. *Journal of Statistical Planning and Inference* **3**, 235–248.

Drew, J.D., Singh, M.P. and Choudhry, G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology* **8**, 17–47.

Duchesne, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics* **16**, 133–138.

DuMouchel, W.H. and Duncan, G.J. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association* **78**, 535–543.

Duncan, G.J. and Kalton, G. (1988). Issues of design and analysis of surveys across time. *International Statistical Review* **55**, 97–117.

Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society B* **15**, 262–269.

Durbin, J. (1958). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute* **36**, 113–119.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika* **46**, 477–480.

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics* **16**, 152–164.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics* **9**, 139–172.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. Monograph 38. SIAM, Philadelphia.

Efron, B. and Tibshirani, R.J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54–77.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Eltinge, J.L. and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology* **23**, 33–40.

Elvers, E., Särndal, C.E., Wretman, J.H. and Örnberg, G. (1985). Regression analysis and ratio analysis for domains: a randomization theory approach. *Canadian Journal of Statistics* **9**, 139–172.

Erdös, P., and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy* **4**, 49–61.

Ericson, W.A. (1965). Optimum stratified sampling using prior information. *Journal of the American Statistical Association* **60**, 750–771.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society B* **31**, 195–224.

Ericson, W.A. (1988). Bayesian inference in finite populations. In: P.R. Krishnaiah and C.R. Rao (eds.), *Handbook of Statistics*, Vol. 6. North-Holland, Amsterdam, pp. 213–246.

Estevao, V., Hidiroglou, M.A. and Särndal, C.E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics* **11**, 181–204.

Estevao, V.M. and Särndal, C.E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics* **18**, 233–255.

Fan, C.T., Muller, M.E. and Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) techniques and digital computers. *Journal of the American Statistical Association* **57**, 387–402.

Fay, R.E. (1984). Application of linear and log linear models to data from complex samples. *Survey Methodology* **10**, 82–96.

Fay, R.E. (1985). A jackknifed chi-square test for complex samples. *Journal of the American Statistical Association* **80**, 148–157.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of Bureau of Census Annual Research Conference.* American Statistical Association, pp. 429–440.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 227–232.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* **91**, 490–498.

Fay, R.E. (1999). Theory and application of nearest neighbor imputation in Census 2000. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 112–121.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places. An application of James–Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.

Fellegi, I.P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Assocation* **58**, 183–201.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association* **75**, 261–168.

Fienberg, S.E. and Tanur, J.M. (1987). Experimental and sampling structures: parallels diverging and meeting. *International Statistical Review* **55**, 75–96.

Firth, D. and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society B* **60**, 3–21.

Fisk, A. (1995). *How to Sample in Surveys.* Sage, Thousand Oaks, CA.

Ford, B.M. (1983). An overview of hot-deck procedures. In: W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. Academic Press, New York, pp. 185–207.

Foreman, E.K. (1991). *Survey Sampling Principles*. Marcel Dekker, New York.

Francis, I. (1981). *Statistical Software: A Comparative Review*. North-Holland, Amsterdam.

Francisco, C.A. (1987). Estimation of quantiles and the interquartile range in complex surveys. Ph.D. dissertation. Iowa State University, Ames, IA.

Francisco, C.A. and Fuller, W.A. (1991). Estimation of quantiles with survey data. *Annals of Statistics* **19**, 454–469.

Francisco, C.A., Fuller, W.A. and Fecso, R. (1987). Statistical properties of crop production estimates. *Survey Methodology* **13**, 45–62.

Frankel, M.R. (1971). *Inference from Survey Samples: An Empirical Investigation*. Institute for Social Research, University of Michigan, Ann Arbor, MI.

Fuller, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association* **61**, 1172–1183.

Fuller, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society B* **32**, 209–226.

Fuller, W.A. (1971). A procedure for selecting nonreplacement unequal probability samples. Unpublished manuscript. Ames, IA.

Fuller, W.A. (1973). Regression for sample surveys. Paper presented at a meeting of the International Statistical Institute, Vienna, Austria, August.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhya C* **37**, 117–132.

Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology* **10**, 97–118.

Fuller, W.A. (1987a). Estimators of the factor model for survey data. In: I.B. MacNeill and G.J. Umphrey (eds.), *Applied Probability, Statistics and Sampling Theory*. D. Reidel, Boston, pp. 265–284.

Fuller, W.A. (1987b). *Measurement Error Models*. Wiley, New York.

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology* **16**, 167–180.

Fuller, W.A. (1991a). Simple estimators for the mean of skewed populations. *Statistica Sinica* **1**, 137–158.

Fuller, W.A. (1991b). Regression estimation in the presence of measurement error. In: P.P. Biemer et al. (eds.), *Measurement Errors in Surveys*. Wiley, New York, pp. 617–635.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406.

Fuller, W.A. (1995). Estimation in the presence of measurement error. (With discussion.) *International Statistical Review* **63**, 121–141.

Fuller, W.A. (1996). *Introduction to Statistical Time Series*, 2nd ed. Wiley, New York.

Fuller, W.A. (1998). Replication variance estimation for two phase samples. *Statistica Sinica* **8**, 1153–1164.

Fuller, W.A. (1999a). Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics* **4**, 331–345.

Fuller, W.A. (1999b). Estimation procedures for the United States national resources inventory. *Proceedings of the Survey Methods Section of the Statistical Society of Canada*, pp. 39–44.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology* **28**, 5–23.

Fuller, W.A. (2003). Estimation for multiple phase samples. In: R.L. Chambers and C.J. Skinner (eds.), *Analysis of Survey Data*. Wiley, Chichester, UK, pp. 307–322.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*. To appear.

Fuller, W.A. and An, A.B. (1998). Regression adjustments for nonresponse. *Journal of the Indian Society of Agricultural Statistics* **51**, 331–342.

Fuller, W.A. and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association* **68**, 626–632.

Fuller, W.A. and Breidt, F.J. (1999). Estimation for supplemented panels. *Sankhya, Special Series B* **61**, 58–70.

Fuller, W.A. and Harter, R.M. (1987). The multivariate components of variance model for small area estimation. In: R. Platek, J.N.K. Rao, C.E. Särndal, and M.R. Singh (eds.), *Small Area Statistics*. Wiley, New York, pp. 103–123.

Fuller, W.A. and Hidiroglou, M.A. (1978). Regression estimation after correcting for attenuation. *Journal of the American Statistical Association* **73**, 99–104.

Fuller, W.A. and Isaki, C.T. (1981). Survey design under superpopulation models. In: D. Krewski, J.N.K. Rao, and R. Platek (eds.), *Current Topics in Survey Sampling*. Academic Press, New York, pp. 199–226.

Fuller, W.A. and Kim, J.K. (2005). Hot deck imputation for the response model. *Survey Methodology* **31**, 139–149.

Fuller, W.A., Kennedy, W., Schnell, D., Sullivan, G. and Park, H.J. (1986). PC CARP. Statistical Laboratory, Iowa State University, Ames, IA.

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting for the 1987–1988 nationwide food consumption survey. *Survey Methodology* **20**, 75–85.

Fuller, W.A. and Rao, J.N.K. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *Annals of Statistics* **6**, 1149–1158.

Fuller, W.A. and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian labour force survey. *Survey Methodology* **27**, 45–52.

Fuller, W.A. and Wang, J. (2000). Geographic information in small area estimation for a national survey. *Statistics in Transition* **4**, 587–596.

Gallant, A.R. (1987). *Nonlinear Statistical Models*. Wiley, New York.

Gambino, J., Kennedy, B. and Singh, M.P. (2001). Regression composite estimation for the Canadian labour force survey: evaluation and implementation. *Survey Methodology* **27**, 65–74.

Gerow, K. and McCulloch, C.E. (2000). Simultaneously model unbiased, design-unbiased estimation. *Biometrics* **56**, 873–878.

Ghosh, J.K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Annals of Mathematical Statistics* **42**, 1957–1961.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science* **9**, 55–76.

Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York.

Giles, P. (1988). A model for generalized edit and imputation of survey data. *Canadian Journal of Statistics* **16**(suppl.), 57–73.

Giommi, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology* **13**, 127–134.

Glasser, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute* **30**, 28–32.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B* **17**, 269–278.

Godambe, V.P. (ed.) (1991). *Estimating Functions*. Oxford University Press, Oxford, UK.

Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations, 1. *Annals of Mathematical Statistics* **36**, 1707–1722.

Godambe, V.P. and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review* **54**, 127–138.

Godfrey, J., Roshwalb, A. and Wright, R.L. (1984). Model-based stratification in inventory cost estimation. *Journal of Business and Economic Statistics* **2**, 1–9.

Goldberger, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association* **57**, 369–375.

Golder, P.A. and Yeomans, K.A. (1973). The use of cluster analysis for stratification. *Applied Statistics* **22**, 213–219.

Gonzalez, M.E., Ogus, J.L., Shapiro, G. and Tepping, B.T. (1975). Standards for discussion and presentation of errors in survey and census data. *Journal of the American Statistical Association* **73**, 7–15.

Goodman, R. and Kish, L. (1950). Controlled selection: a technique in probability sampling. *Journal of the American Statistical Association* **45**, 350–372.

Graubard, B.I. and Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science* **17**, 73–96.

Gray, G.B. (1975). Components of variance model in multistage stratified samples. *Survey Methodology* **1**, 27–43.

Gray, G.B. and Platek, R. (1976). Analysis of design effects and variance components in multi-stage sample surveys. *Survey Methodology* **2**, 1–30.

Graybill, F.A. (1976). *Theory and Application of the Linear Model.* Wadsworth, Belmont, CA.

Graybill, F.A. (1983). *Matrices with Applications in Statistics*, 2nd ed. Wadsworth, Belmont, CA.

Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* **77**, 251–261.

Gregoire, T.G., Arabatzis, A.A. and Reynolds, M.R. (1992). Mean squared error performance of simple linear regression conditional upon the outcome of pretesting the intercept. *The American Statistician* **46**, 89–93.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**, 489–504.

Gross, S.T. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research*, American Statistical Association, pp. 181–184.

Groves, R.M. (1987). Research on survey data quality. *Public Opinion Quarterly* **51**, S156–S172.

Hájek, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pro Pestovani Matematiky* **84**, 387–423.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy* **5**, 361–374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* **35**, 1491–1523.

Hájek, J. (1971). Comment on a paper by D. Basu. In: V.P. Godambe and D.A. Sprott (eds.), *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto, Ontario, Canada, p. 236.

Hájek, J. (1981). *Sampling from a Finite Population.* Marcel Dekker, New York.

Hall, P. and Heyde, C.C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.

Hannan, E.J. (1962). Systematic sampling. *Biometrika*, **49**, 281–283.

Hansen, M.H., Dalenius, T. and Tepping, B.J. (1985). The development of sample surveys of finite populations. In: A.C. Atkinson and S.E. Fienberg (eds.), *A Celebration of Statistics: The ISI Centenary Volume.* Springer-Verlag, New York, pp. 327–354.

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* **14**, 333–362.

Hansen, M.H. and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association* **41**, 517–529.

Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute* **38**, 359–374.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vols. I and II. Wiley, New York.

Hansen, M.H., Hurwitz, W.N., Marks, E.S. and Mauldin, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association* **46**, 147–190.

Hansen, M.H., Hurwitz, W.N. and Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance. In: C.R. Rao (ed.), *Contributions to Statistics.* Pergamon Press, Calcutta, India, pp. 111–136.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association* **78**, 776–793.

Hanurav, T.V. (1966). Some aspects of unified sampling theory. *Sankhya A* **28**, 175–204.

Hanurav, T.V. (1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *Journal of the Royal Statistical Society B* **29**, 374–391.

Hartley, H.O. (1965). Multiple purpose optimum allocation in stratified sampling. *Proceedings of Social Statistics Section*, American Statistical Association, pp. 258–261.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics* **33**, 350–374.

Hartley, H.O. and Rao, J.N.K. (1968). A new estimation theory for survey samples. *Biometrika* **55**, 547–557.

Hartley, H.O. and Rao, J.N.K. (1978). Estimation of nonsampling variance components in sample surveys. In: N.K. Namboodiri (ed.), *Survey Sampling and Measurement.* Academic Press, New York, pp. 35–43.

Hartley, H.O. and Ross, A. (1954). Unbiased ratio estimators. *Nature* **174**, 270–271.

Hartley, H.O. and Sielken, R.L. (1975). A super population viewpoint for finite population sampling. *Biometrics* **32**, 411–422.

Harville, D.A. (1976). Extension of the Gauss–Markov theorem to include estimation of random effects. *Annals of Statistics* **4**, 384–395.

Haslett, S. (1985). The linear non-homogeneous estimator in sample surveys. *Sankhya B* **47**, 101–117.

Hedayat, A. (1979). Sampling designs with reduced support size. In: J.S. Rustagi (ed.), *Optimizing Methods in Statistics: Proceedings of an International Conference.* Academic Press, New York, pp. 273–288.

Hedayat, A.S. and Sinha, B.K. (1991). *Design and Inference in Finite Population Sampling.* Wiley, New York.

Henderson, C.R. (1963). Selection index and expected genetic advance. In: *Statistical Genetics and Plant Breeding*. National Academy of Sciences–National Research Council, Washington, DC, pp. 141–163.

Herson, J. (1976). An investigation of relative efficiency of least squares prediction to conventional probability sampling plans. *Journal of the American Statistical Association* **71**, 700–703.

Hess, I. (1985). *Sampling for Social Research Surveys, 1947–1980.* Institute for Social Research, Ann Arbor, MI.

Hidiroglou, M.A. (1974). Estimation of regression parameters for finite populations. Ph.D. dissertation. Iowa State University, Ames, IA.

Hidiroglou, M.A. (1986a). The construction of a self-representing stratum of large units in survey design. *The American Statistician* **40**, 27–31.

Hidiroglou, M.A. (1986b). Estimation of regression parameters for finite populations: a Monte-Carlo study. *Journal of Official Statistics* **2**, 3–11.

Hidiroglou, M.A. (1994). Sampling and estimation for establishment surveys: stumbling blocks and progress. *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 153–162.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology* **27**, 143–154.

Hidiroglou, M.A. and Berthelot, J.M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology* **12**, 73–83.

Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1980). *SUPERCARP–Sixth Edition.* Statistical Laboratory, Survey Section, Iowa State University, Ames, IA.

Hidiroglou, M.A. and Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics* **29**, 107–112.

Hidiroglou, M.A. and Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology* **30**, 67–78.

Hidiroglou, M.A. and Rao, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys, I and II. *Journal of Official Statistics* **3**, 117–132, 133–140.

Hidiroglou, M.A. and Särndal, C.E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology* **24**, 11–20.

Hidiroglou, M.A., Särndal, C.E. and Binder, D.A. (1995). Weighting and estimation in business surveys. In: B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (eds.), *Business Survey Methods*. Wiley, New York, pp. 477–502.

Hidiroglou, M.H. and Srinath, K.P. (1981). Some estimators of the population totals from a simple random sample containing large units. *Journal of the American Statistical Association* **76**, 690–695.

Hilgard, E.R. and Payne, S.L. (1944). Those not at home: riddle for pollsters. *Public Opinion Quarterly* **8**, 254–261.

Holt, D., Scott, A.J. and Ewings, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society A* **143**, 302–330.

Holt, D. and Smith, T.M.F. (1976). The design of surveys for planning purposes. *The Australian Journal of Statistics* **18**, 37–44.

Holt, D. and Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society A* **142**, 33–46.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society A* **143**, 474–487.

Holt, M.M. (1982). *SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data. Research Triangle Institute*, Research Triangle Park, NC.

Horgan, J.M. (2006). Stratification of skewed populations: a review. *International Statistical Review* **74**, 67–76.

Horn, S.D., Horn, R.A. and Duncan, D.B. (1975). Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association* **70**, 380–385.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Huang, E.T. and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 300–305.

Huang, R. and Hidiroglou, M.A. (2003). Design consistent estimators for a mixed linear model on survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Assocation, pp. 1897–1904.

Huddleston, H.F., Claypool, P.L. and Hocking, R.R. (1970). Optimal sample allocation to strata using convex programming. *Applied Statistics* **19**, 273–278.

Hughes, E. and Rao, J.N.K. (1979). Some problems of optimal allocation in sample surveys involving inequality constraints. *Communications in Statistics A* **8**, 1551–1574.

Hung, H.M. and Fuller, W.A. (1987). Regression estimation of crop acreages with transformed Landsat data as auxiliary variables. *Journal of Business and Economics Statistics* **5**, 475–482.

Huntsberger, D.V. (1955). A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics* **26**, 734–743.

Husain, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis. Iowa State University, Ames, IA.

Ireland, C.T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* **55**, 179–188.

Isaki, C.T. (1970). Survey designs utilizing prior information. Unpublished Ph.D. dissertation. Iowa State University, Ames, IA.

Isaki, C.T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association* **78**, 117–123.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* **77**, 89–96.

Isaki, C.T., Ikeda, M.M., Tsay, J.H. and Fuller, W.A. (2000). An estimation file that incorporates auxiliary information. *Journal of Official Statistics* **16**, 155–172.

Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2000). Estimation of census adjustment factors. *Survey Methodology* **26**, 31–42.

Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2004). Weighting sample data subject to independent controls. *Survey Methodology* **30**, 35–44.

Jagers, P. (1986). Post-stratification against bias in sampling. *International Statistical Review* **54**, 159–167.

Jagers, P., Odén, A. and Trulsson, L. (1985). Post-stratification and ratio estimation. *International Statistical Review* **53**, 221–238.

Jarque, C.M. (1981). A solution to the problem of optimum stratification in multivariate sampling. *Applied Statistics* **30**, 163–169.

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agriculture Experiment Station Research Bulletin 304.*

Jessen, R.J. (1969). Some methods of probability nonreplacement sampling. *Journal of the American Statistical Association* **64**, 175-193.

Jessen, R.J. (1978). *Statistical Survey Techniques.* Wiley, New York.

Jönrup, H. and Rennermalm, B. (1976). Regression analysis in samples from finite populations. *Scandinavian Journal of Statistics* **3**, 33–37.

Johnson, N.L. and Smith, H. (1969). *New Developments in Survey Sampling*. Wiley, New York.

Johnson, N.R. (1967). A Monte Carlo investigation on some regression and ratio estimators. Unpublished M.S. thesis. Iowa State University, Ames, IA.

Kackar, R.N. and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853–862.

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review* **51**, 175–188.

Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics* **18**, 129–154.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology* **12**, 1–16.

Kalton, G. and Kish, L. (1981). Two efficient random imputation procedures. *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 146–151.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics A*, **13**, 1919–1939.

Kalton, G. and Maligalig, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, Washington, DC, pp. 409–428.

Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (eds.) (1989). *Panel Surveys*. Wiley, New York.

Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance estimation. *Journal of the American Statistical Association* **96**, 1387–1396.

Kendall, M. and Stuart, A. (1976). *The Advanced Theory of Statistics*, 3rd ed, Vol. 3. Griffin, London.

Kim, J.K., Brick, M.J., Fuller, W.A., and Kalton, G. (2006). On the bias of the multiple imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society B* **68**, 509–521.

Kim, J.K. and Fuller, W.A. (1999). Jackknife variance estimation after hot deck imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 825–830.

Kim, J.K. and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika* **91**, 559–578.

Kim, J.K., Fuller, W.A. and Bell, W.R. (2009). Variance estimation for nearest neighbor imputation for U.S. Census long form data. Unpublished manuscript. Iowa State University, Ames, IA.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association* **101**, 312–320.

Kim, J.K. and Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica* **13**, 641–653.

Kish, L. (1965). *Survey Sampling*. Wiley, New York.

Kish, L. and Anderson, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association* **73**, 24–34.

Kish, L. and Frankel, M.R. (1970). Balanced repeated replication for standard errors. *Journal of the American Statistical Association* **65**, 1071–1094.

Kish, L. and Frankel, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society B* **36**, 1–37.

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. Springer-Verlag, New York.

Koch, G.G., Freeman, D.H., Jr. and Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review* **43**, 59–78.

Koehler, K.J. and Wilson, J.R. (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. *Communication in Statistics A* **15**, 2977–2990.

Konijn, H.S. (1962). Regression analysis for sample surveys. *Journal of the American Statistical Association* **57**, 590–606.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni t-statistics. *The American Statistician* **44**, 270–276.

Korn, E.L. and Graubard, B.I. (1995a). Analysis of large health surveys: accounting for the sampling designs. *Journal of the Royal Statistical Society A* **158**, 263–295.

Korn, E.L. and Graubard, B.I. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician* **49**, 291–295.

Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys.* Wiley, New York.

Kott, P.S. (1985). A note on model-based stratification. *Journal of Business and Economic Statistics* **3**, 284–286.

Kott, P.S. (1990a). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference* **24**, 287–296.

Kott, P.S. (1990b). Variance estimation when a first-phase area sample is restratified. *Survey Methodology* **16**, 99–103.

Kott, P.S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association* **89**, 693–696.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics* **17**, 521–526.

Kott, P.S. (2006a). Delete-a-group variance estimation for the general regression estimator under Poisson sampling. *Journal of Official Statistics* **22**, 759–767.

Kott, P.S. (2006b). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology* **32**, 133–142.

Kott, P.S. and Bailey, J.T. (2000). The theory and practice of Brewer selection with Poisson PRN sampling. *Proceedings of the International Conference on Establishment Surveys II,* Buffalo, NY.

Kott, P.S. and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology* **23**, 81–89.

Kovar, J., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics* **16**(suppl.), 25–45.

Krewski, D. and Rao, J.N,K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* **9**, 1010–1019.

Krieger, A.M. and Pfeffermann, D. (1992). Maximum likelihood from complex sample surveys. *Survey Methodology* **18**, 225–239.

Krieger, A.M. and Pfeffermann, D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics* **13**, 123–142.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika* **75**, 97–103.

Kumar, S. and Rao, J.N.K. (1984). Logistic regression analysis of labour force survey data. *Survey Methodology* **10**, 62–81.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute* **33**, 133–140.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* **18**, 199–210.

Lanke, J. (1975). On UMV-estimators in survey sampling. *Metrika* **20**, 196–202.

Lanke, J. (1983). Hot deck imputation techniques that permit standard methods for assessing precision of estimates. (Essays in honor of Tore E. Dalenius.) *Statistical Review* **21**, 105–110.

Lavallée, P. and Hidiroglou, M.A. (1987). On the stratification of skewed populations. *Survey Methodology* **14**, 33–43.

Lazzeroni, L.C. and Little, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics* **14**, 61–78.

Legg, J.C. and Fuller, W.A. (2009). Two-phase sampling. In: D. Pfeffermann and C.R. Rao (eds.), *Sample Surveys: Theory, Methods and Inference*. Wiley, New York.

Legg, J.C. and Yu, C.L. (2009). A comparison of sample set restriction procedures. *Survey Methodology*. To appear.

Lehtonen, R. and Pahkinen, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. Wiley, New York.

Lessler, J.T. (1984). Measurement errors in surveys. In: C.F. Turner and E. Martin (eds.), *Surveying Subjective Phenomena*, Vol. 2. Russell Sage Foundation, New York, pp. 405–440.

Levy, P.S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association* **72**, 758–763.

Lin, D.Y. (2000). On fitting Cox's proportional hazards model to survey data. *Biometrika* **87**, 37–47.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association* **77**, 237–250.

Little, R.J.A. (1983a). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association* **78**, 596–604.

Little, R.J.A. (1983b). Superpopulation models for nonresponse, the ignorable case. In: W.G. Madow, I. Olkin, and D.B. Rubin (eds.),

Incomplete Data in Sample Surveys, Vol. 2. Academic Press, New York, pp. 341–382.

Little, R.J.A. (1986). Survey nonresponse adjustments. *International Statistical Review* **54**, 139–157.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* **6**, 287–301.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics* **7**, 405–424.

Little, R.J.A. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association* **88**, 1001–1012.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* **99**, 546–556.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

Liu, T.P. and Thompson, M.E. (1983). Properties of estimators of quadratic finite population functions: the batch approach. *Annals of Statistics* **11**, 275–285.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Lohr, S.L. and Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association* **95**, 271–280.

Lu, W.W., Brick, J.M. and Sitter, R.R. (2006). Algorithms for constructing combined strata variance estimators. *Journal of the American Statistical Association* **101**, 1680–1692.

Lyberg, L. (1983). The development of procedures for industry and occupation coding at Statistics Sweden. (Essays in honor of Tore E. Dalenius.) *Statistical Review* **21**, 139–156.

Madow, W.G. (1948). On the limiting distribution of estimates based on samples from finite universes. *Annals of Mathematical Statistics* **19**, 535–545.

Madow, W.G. and Madow, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics* **15**, 1–24.

Madow, W.G., Nisselson, H. and Olkin, I. (eds.) (1983). *Incomplete Data in Sample Surveys*, Vol. 1. Academic Press, New York.

Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society B* **60**, 115–126.

Mahalanobis, P.C. (1939). A sample survey of the acreage under jute in Bengal. *Sankhya* **4**, 511–531.

Mahalanobis, P.C. (1944). On large-scale sample surveys. *Philosophical Transactions of the Royal Society of London B* **231**, 329–451.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society* **109**, 325–370.

Mahalanobis, P.C. (1952). Some aspects of the design of sample surveys. *Sankhya* **12**, 1–7.

Mann, H.B. and Wald, A. (1943). On stochastic limit and order relationships. *Annals of Mathematical Statistics* **14**, 217–226.

Maritz, J.S. and Jarrett, R.G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association* **73**, 194–196.

McCallion, T. (1992). Optimum allocation in stratified random sampling with ratio estimation as applied to the Northern Ireland December agricultural sample. *Applied Statistics* **41**, 39–45.

McCarthy, P.J. (1965). Stratified sampling and distribution-free confidence intervals for a median. *Journal of the American Statistical Association* **60**, 772–783.

McCarthy, P.J. (1969). Pseudo-replication: half-samples. *Review of the International Statistical Institute* **37**, 239–264.

McCarthy, P.J. and Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, 2–95. Public Health Service Publications 85–1369. U.S. Government Printing Office, Washington, DC.

McLeod, A. I. and Bellhouse, D.R. (1983). A convenient algorithm for drawing a simple random sample. *Applied Statistics* **32**, 182–184.

Meeks, S.L. and D'Agostino, R.B. (1983). A note on the use of confidence limits following rejection of a null hypothesis. *The American Statistician* **37**, 134–136.

Meng, X.L. (1994). Multiple-imputation inferences with uncogenial sources of input (with discussion), *Statistical Science* **9**, 538–573.

Mickey, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association* **54**, 594–612.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of sizes. *Annals of the Institute of Statistical Mathematics* **3**, 99–107.

Montanari, G.E. (1987). Post-sampling efficient Q-R prediction in large-sample surveys. *International Statistical Review* **55**, 191–202.

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology* **24**, 69–77.

Montanari, G.E. (1999). A study on the conditional properties of finite population mean estimators. *Metron* **57**, 21–35.

Morel, J.G. (1989). Logistic regression under complex survey designs. *Survey Methodology* **15**, 203–223.

Moser, C.A. and Kalton, G. (1971). *Survey Methods in Social Investigation*. 2nd ed. Heinemann, London.

Mukhopadhyay, P. (1993). Estimation of a finite population total under regression models: a review. *Sankhya* **55**, 141–155.

Mukhopadhyay, P.K. (2006). Extensions of small area models with applications to the national resources inventory. Ph.D. dissertation. Iowa State University, Ames, IA.

Mulry, M.H. and Wolter, K.M. (1981). The effect of Fisher's Z-transformation on confidence intervals for the correlation coefficient. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 113–121.

Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta, India.

Murthy, M.N. (1983). A framework for studying incomplete data with a reference to the experience in some countries of Asia and the Pacific. In: W.G. Madow and I. Olkin (eds.), *Incomplete Data in Sample Surveys*, Vol. 3. Academic Press, New York, pp. 7–24.

Mussa, A.S. (1999). A new bias-reducing modification of the finite population ratio estimator and a comparison among proposed alternatives. *Journal of Official Statistics* **15**, 25–38.

Namboodiri, N.K. (ed). (1978). *Survey Sampling and Measurement*. Academic Press, New York.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **3**, 169–175.

Nascimento Silva, P.L.D. and Skinner, C.J. (1995). Estimating distribution functions with auxiliary information using poststratification. *Journal of Official Statistics* **11**, 277–294.

Nathan, G. (1969). Tests of independence in contingency tables from stratified samples. In: N.L. Johnson and H. Smith (eds.), *New Developments in Survey Sampling*. Wiley, New York, pp. 578–600.

Nathan, G. (1975). Tests of independence in contingency tables from stratified proportional samples. *Sankhya C* **37**, 77–87; corrigendum: (1978), *Sankhya C* **40**, 190.

Nathan, G. (1981). Notes on inference based on data from complex sample designs. *Survey Methodology* **7**, 110–129.

Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society B* **42**, 377–386.

Nedyalkova, D. and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika* **95**, 521–537.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society A* **135**, 370–384.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558–625.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* **33**, 101–116.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics* **5**, 223–239.

Nusser, S.M., Breidt, F.J. and Fuller, W.A. (1998). Design and estimation for investigating the dynamics of natural resources. *Ecological Applications* **8**, 234–245.

Nusser, S.M., Carriquiry, A.L., Dodd, K.W. and Fuller, W.A. (1996). A semiparametic transformation approach to estimating usual daily intake distribution. *Journal of the American Statistical Association* **91**, 1440–1449.

Nusser, S.M. and Goebel, J.J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics* **4**, 181–204.

Nygård, F. and Sandström, A. (1985). Estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics* **1**, 399–412.

Oh, H.L. and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In: W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. Academic Press, New York, pp. 143–184.

Oñate, B.T. and Bader, J. M. O. (1990). *Sampling Surveys and Applications*. College, Laguna, Philippines.

Papageorgiou, I. and Karakostas, K.X. (1998). On optimal sampling designs for autocorrelated finite populations. *Biometrika* **85**, 482–486.

Park, M. (2002). Regression estimation of the mean in survey sampling. Unpublished Ph.D. dissertation. Iowa State University, Ames, IA.

Park, M. and Fuller, W.A. (2002). A two-per-stratum design. Unpublished manuscript. Iowa State University, Ames, IA.

Park, M. and Fuller, W.A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology* **31**, 85–93.

Park, M. and Fuller, W.A. (2009). The mixed model for survey regression estimation. *Journal of Statistical Planning and Inference* **139**, 1320–1331.

Pasqual, J.N. (1961). On simple random sampling with replacement. *Sankhya A* **24**, 287–302.

Patak, Z., Hidiroglou, M. and Lavallée, P. (1998). The methodology of the workplace employee survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 83–91.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society B* **12**, 241–255.

Pfeffermann, D. (1984). Note on large sample properties of balanced samples. *Journal of the Royal Statistical Society B* **46**, 38–41.

Pfeffermann, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *Journal of the American Statistical Association* **83**, 824–833.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review* **61**, 317–337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research* **5**, 239–262.

Pfeffermann, D. and Barnard, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics* **9**, 73–84.

Pfeffermann, D. and Holmes, D.J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society A* **148**, 268–278.

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica* **8**, 1087–1114.

Pfeffermann, D. and Nathan, G. (1977). Regression analysis of data from complex samples. *Bulletin of the International Statistical Institute* **49**, 699–718.

Pfeffermann, D. and Nathan, G. (1981). Regression analysis of data from a cluster sample. *Journal of the American Statistical Association* **76**, 681–689.

Pfeffermann, D., Skinner, C., Goldstein, H., Holmes, D. and Rasbash, J. (1998). Weighting for unequal probabilities in multilevel models. *Journal of the Royal Statistical Society B* **60**, 23–40.

Pfeffermann, D. and Smith, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review* **53**, 37–59.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhya B* **61**, 166–186.

Pla, L. (1991). Determining stratum boundaries with multivariate real data. *Biometrics* **47**, 1409–1422.

Plackett, R.L. and Burman, J.P. (1946). The design of optimum multifactorial experiments. *Biometrika* **33**, 305–325.

Platek, R. and Gray, G.B. (1983). Imputation methodology. In: W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. Academic Press, New York, pp. 255–293.

Platek, R., Singh, M.P., Rao, J.N.K. and Särndal, C.E. (eds.), (1987). *Small Area Statistics: An International Symposium*. Wiley, New York.

Politz, A. and Simmons, W. (1949). An attempt to get not-at-homes into the sample without callbacks. *Journal of the American Statistical Association* **44**, 9–31.

Porter, R.D. (1973). On the use of survey sample weights in the linear model. *Annals of Economic and Social Measurement* **2**, 141–158.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of the small area estimators. *Journal of the American Statistical Association* **85**, 163–171.

Prasad, N.G.N. and Rao, J.N.K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology* **25**, 67–72.

Pratt, J.W. (1959). On a general concept of "In Probability." *Annals of Mathematical Statistics* **30**, 549–558.

Pritzker, L., Ogus, J. and Hansen, M.H. (1965). Computer editing methods: some applications and results. *Bulletin of the International Statistical Institute* **41**, 442–466.

Pulum, T.W., Harpham, T. and Ozsever, N. (1986). The machine editing of large-sample surveys: the experience of the world fertility survey. *International Statistical Review* **54**, 311–326.

Purcell, N.J. and Kish, L. (1979). Estimation for small domains. *Biometrics* **35**, 365–384.

Quenouille, M.H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics* **20**, 355–375.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353–360.

Raj, D. (1964). On forming strata of equal aggregate size. *Journal of the American Statistical Association* **59**, 481–486.

Raj, D. (1968). *Sampling Theory*. McGraw-Hill, New York.

Raj, D. (1972). *The Design of Sample Surveys*. McGraw-Hill, New York.

Raj, D. and Chandhok, P. (1998). *Sample Survey Theory*. Narosa, London.

Raj, D. and Khamis, S.H. (1958). Some remarks on sampling with replacement. *Annals of Mathematical Statistics* **29**, 550–557.

Rancourt, E., Särndal, C.E., and Lee, H. (1994). Estimation of the variance in the presence of nearest neighbor imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 888–893.

Randles, R.H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* **10**, 462–474.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association* **3**, 173–180.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika* **60**, 125–133.

Rao, J.N.K. (1975). Unbiased variance estimation for multistage designs. *Sankhya C* **60**, 125–133.

Rao, J.N.K. (1978). Sampling designs involving unequal probabilities of selection and robust estimation of a finite population total. In: H.A. David (ed.), *Contributions to Survey Sampling and Applied Statistics*. Academic Press, New York, pp. 69–87.

Rao, J.N.K. (1979a). On deriving mean square errors and their non-negative unbiased estimators. *Journal of the Indian Statistical Association* **17**, 125–136.

Rao, J.N.K. (1979b). Optimization in the design of sample surveys. In: J.S. Rustagi (ed.), *Optimizing Methods in Statistics: Proceedings of an International Conference*. Academic Press, New York, pp. 419–434.

Rao, J.N.K. (1982). Some aspects of variance estimation in sample surveys. *Utilitas Mathematica B* **21**, 205–225.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology* **11**, 15–31.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics* **10**, 153–165.

Rao, J.N.K. (1996). On variance estimation with imputed data. *Journal of the American Statistical Association* **91**, 499–506.

Rao, J.N.K. (2003). *Small Area Estimation.* Wiley, Hoboken, NJ.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: an appraisal. *Survey Methodology* **31**, 117–138.

Rao, J.N.K. and Bellhouse, D.R. (1990). History and development of survey based estimation and analysis. *Survey Methodology* **16**, 3–29.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society B* **24**, 482–491.

Rao, J.N.K. and Hidiroglou, M.A. (1981). Chi square tests for the analysis of categorical data from the Canada health survey. *Bulletin of the International Statistical Institute* **49**, 699–718.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**, 365–375.

Rao, J.N.K., Kumar, S. and Roberts, G. (1989). Analysis of sample survey data involving categorical response variables: method and software. *Survey Methodology* **15**, 161–185.

Rao, J.N.K. and Nigam, A.K. (1992). "Optimal" controlled sampling: a unified approach. *International Statistical Review* **60**, 89–98.

Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* **76**, 221–230.

Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics* **12**, 46–60.

Rao, J.N.K. and Scott, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics* **15**, 385–397.

Rao, J.N.K. and Scott, A.J. (1992). A simple method for the analysis of clustered data. *Biometrics* **48**, 577–585.

Rao, J.N.K. and Scott, A.J. (1999). A simple method for analyzing overdispersion in clustered Poisson data. *Statistics in Medicine* **18**, 1373–1385.

Rao, J.N.K., Scott, A.J. and Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica* **8**, 1059–1070.

Rao, J.N.K. and Shao, J. (1991). Jackknife variance estimation with survey data under hot deck imputation. *Technical Report 172*. Carleton University, Ottawa, Ontario, Canada.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811–822.

Rao, J.N.K. and Singh, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 57–64.

Rao, J.N.K. and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika* **82**, 453–460.

Rao, J.N.K. and Thomas, D.R. (1991). Chi-squared tests with complex survey data subject to misclassification error. In: P.P. Biemer et al. (eds.), *Measurement Errors in Surveys.* Wiley, New York, pp. 637–663.

Rao, J.N.K. and Vijayan K. (1997). On estimating the variance with probability proportional to aggregate size. *Journal of the American Statistical Association* **72**, 579–584.

Rao, J.N.K. and Wu, C.F.J. (1984). Bootstrap inference for sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 106–112.

Rao, J.N.K. and Wu, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association* **80**, 620–630.

Rao, J.N.K. and Wu, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data: some recent work. *Bulletin of the International Statistical Institute* **52**, 5–21.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association*, **83**, 231–241.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys, *Survey Methodology* **18**, 209–217.

Rao, P.S.R.S. (1983). Callbacks, follow-ups, and repeated telephone calls. In: W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. Academic Press, New York, pp. 33–44.

Rao, P.S.R.S. (1988). Ratio and regression estimators. In: P.R. Krishnaiah and C.R. Rao (eds.), *Handbook of Statistics*, Vol. 6. North-Holland, Amsterdam, pp. 449–468.

Rao, R.R. (1962). Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics* **33**, 659–680.

Rivest, L.-P.(1994). Statistical properties of winsorized means for skewed distributions. *Biometrika* **81**, 373–383.

Roberts, G., Rao, J.N.K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**, 1-12.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15–51.

Robinson, P.M. and Särndal, C.E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya B* **45**, 240–248.

Rosén, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement, I and II. *Annals of Mathematical Statistics* **43**, 373–397, 748–776.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for casual effects. *Biometrika* **70**, 41–55.

Rossi, P.H., Wright, J.D. and Anderson, A.B., (eds.) (1983). *Handbook of Survey Research.* Academic Press, New York.

Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485–497.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377–387.

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association* **71**, 657–664.

Royall, R.M. (1981). The finite population linear regression estimator and estimators of its variance: an empirical study. *Journal of the American Statistical Association* **76**, 924–930.

Royall, R.M. (1986). The prediction approach to robust variance estimation in two stage cluster sampling. *Journal of the American Statistical Association* **81**, 119–123.

Royall, R.M. (1992). Robust and optimal design under prediction models for finite populations. *Survey Methodology* **18**, 179–185.

Royall, R.M. and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association* **73**, 351–358.

Royall, R.M. and Cumberland, W.G. (1981). The finite population linear regression estimator and estimators of its variance, an empirical study. *Journal of the American Statistical Association* **76**, 924–930.

Royall, R.M. and Pfeffermann, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika* **69**, 401–409.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581–590.

Rubin, D.B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 20–34.

Rubin, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics* **9**, 130–134.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. In: J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.), *Bayesian Statistics*, Vol. 2. North-Holland, Amsterdam, pp. 463–472.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

Rubin, D.B and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366–374.

Rubin-Bleuer, S., and Kratina, I.S. (2005). On the two phase framework for joint model and design-based inference. *The Annals of Statistics* **33**, 2789–2810.

Saleh, A.K. Md.E. (2006). *Theory of Preliminary Test and Stein-Type Estimation with Applications*. Wiley, New York.

Sampford, M.R. (1962). *An Introduction to Sampling Theory*. Oliver & Boyd, London.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499–513.

Sande, I.G. (1982). Imputation in surveys: coping with reality. *The American Statistician* **36**, 145–152.

Sande, I.G. (1983). Hot-deck imputation procedures. In: W.G. Madow and I. Olkin (eds.), *Incomplete Data in Sample Surveys*, Vol. 3. Academic Press, New York, pp. 339–349.

Sandström, A., Wretman, J.G. and Waldén, B. (1988). Variance estimators of the Gini coefficient–probability sampling. *Journal of Business and Economic Statistics* **6**, 113–119.

Särndal, C.E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics* **5**, 27–52.

Särndal, C.E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* **67**, 639–650.

Särndal, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. *Bulletin of the International Statistical Institute* **49**, 494–513.

Särndal, C.E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference* **7**, 155–170.

Särndal, C.E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association* **79**, 624–631.

Särndal, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology* **18**, 241–252.

Särndal, C.E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistics Association* **91**, 1289–1300.

Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* **33**, 99–119.

Särndal, C.E. and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association* **84**, 266–275.

Särndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* **55**, 279–294.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika* **76**, 527–537.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics* **2**, 110–114.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Schaible, W.L. (1979). A composite estimator for small area statistics. In: C. Steinberg (ed.), *Synthetic Estimates for Small Areas*. National Institute on Drug Abuse, Research Monograph 24. U.S. Government Printing Office, Washington, DC, pp. 36–83.

Scheaffer, R.L., Mendenhall, W. and Ott, L. (1990). *Elementary Survey Sampling*, 4th ed. PWS-Kent, Boston.

Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Annals of Statistics* **16**, 1550–1566.

Schnell, D., Kennedy, W.J., Sullivan, G., Park, H.J. and Fuller, W.A. (1988). Personal computer variance software for complex surveys. *Survey Methodology* **14**, 59–69.

Scott, A. and Smith, T.M.F. (1974). Linear superpopulation models in survey and sampling. *Sankhya C* **36**, 143–146.

Scott, A. and Wu, C.F. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association* **76**, 98–102.

Scott, A.J. (1977). Large samples posterior distributions for finite populations. *Annals of Mathematical Statistics* **42**, 1113–1117.

Scott, A.J. and Holt, D. (1982). The effect of two stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77**, 848–854.

Scott, A.J. and Rao, J.N.K. (1981). Chi-squared tests for contingency tables with proportions estimated from survey data. In: D. Krewski, R. Platek, and J.N.K. Rao (eds.), *Current Topics in Survey Sampling*. Academic Press, New York, pp. 247–265.

Scott, D.W. (1992). *Multivariate Density Estimation*. Wiley, New York.

Searle, S.R. (1971). *Linear Models*. Wiley, New York.

Sedransk, J. and Meyer, J. (1978). Confidence intervals for the quantiles of a finite population: simple random and stratified simple random sampling. *Journal of the Royal Statistical Society B* **40**, 239–252.

Sedransk, N. and Sedransk, J. (1979). Distinguishing among distributions using data from complex sample designs. *Journal of the American Statistical Association* **74**, 754–760.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.

Sen, P.K. (1988). Asymptotics in finite populations. In: P.R. Krishnaiah and C.R. Rao (eds.), *Handbook of Statistics*, Vol. 6. North-Holland, Amsterdam, pp. 291–331.

Shah, B.V., Holt, M.M. and Folsom, R.E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute* **47**, 43–57.

Shao, J. (1989a). The efficiency and consistency of approximations to the jackknife variance estimators. *Journal of the American Statistical Association* **84**, 114–119.

Shao, J. (1989b). Asymptotic distribution of weighted least squares estimator. *Annals of the Institute of Mathematical Statistics* **41**, 365–382.

Shao, J. (1994). L-statistics in complex survey problems. *Annals of Statistics* **22**, 946–967.

Shao, J. and Rao, J.N.K. (1994). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankya B* **55**, 393–414.

Shao, J. and Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer-Verlag, New York.

Sheather, S.J. (2004). Density estimation. *Statistical Science* **19**, 588–597.

Shen, X., Huang, H-C., and Ye, J. (2004). Inference after model selection. *Journal of the American Statistical Association* **99**, 751–762.

Silva, P.L.D.N. and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology* **23**, 23–32.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Singh, A.C. and Folsom, R.E. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 610–615.

Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian labour force survey with a rotating design. *Survey Methodology* **27**, 33–44.

Singh, A.C. and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology* **22**, 107–115.

Singh, D. and Chaudhary, F.S. (1986). *Theory and Analysis of Sample Survey Design*. Wiley, New York.

Singh, S. and Singh, R. (1979). On random non-response in unequal probability sampling. *Sankhya C* **41**, 127–137.

Sirken, M.G. (1972). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association* **67**, 224–227.

Sirken, M.G. (2001). The Hansen–Hurwitz estimator revisited: PPS sampling without replacement. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association* **92**, 780–787.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In: C.J. Skinner, D. Holt, and T.M.F. Smith (eds.), *Analysis of Complex Surveys*. Wiley, New York, pp. 59–87.

Skinner, C.J. (1994). Sample models and weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 133–142.

Skinner, C.J. (1998). Logistic modeling of longitudinal survey data with measurement errors. *Statistica Sinica* **8**, 1045–1058.

Skinner, C.J., Holmes, D.J. and Smith, T.M.F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association* **81**, 789–798.

Skinner, C.J., Holt, D. and Smith, T.M.F. (eds.) (1989). *Analysis of Complex Surveys*. Wiley, New York.

Smith, P. and Sedransk, J. (1983). Lower bounds for confidence coefficients for confidence intervals for finite population quantiles. *Communications in Statistics Theory and Methods* **12**, 1329–1344.

Smith, T.M.F. (1976). The foundations of survey sampling: a review (with discussion). *Journal of the Royal Statistical Society A* **139**, 183–204.

Smith, T.M.F. (1988). To weight or not to weight, that is the question. In: J.M. Bernado, M.H. DeGroot, and D.V. Lindley (eds.), *Bayesian Statistics*, Vol. 3. Oxford University Press, Oxford, UK.

Solomon, H. and Stephens, M.A. (1977). Distribution of a sum of weighted chi-square variables. *Journal of the American Statistical Association* **72**, 881–885.

Srinath, K.P. and Hidiroglou, M.A. (1980). Estimation of variance in multi-stage sampling. *Metrika* **27**, 121–125.

Statistics Canada (1987). Statistics Canada's policy on informing users of data quality and methodology. *Journal of Official Statistics* **3**, 83–92.

Stefanski, L.A. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics* **21**, 249–259.

Stephan, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal tables are known. *Annals of Mathematical Statistics* **13**, 166–178.

Stuart, A. (1984). *The Ideas of Sampling*, rev. ed. Griffin, London.

Stuart, A. and Ord, J.K. (1991). *Kendall's Advanced Theory of Statistics*, Vol. 2. Oxford University Press, New York.

Sudman, S. (1976). *Applied Sampling*. Academic Press, New York.

Sugden, R.A. and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495–506.

Sukhatme, P.V. (1947). The problem of plot size in large-scale surveys. *Journal of the American Statistical Association* **42**, 297–310.

Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Applications*. Asia Publishing House, London.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*, 3rd ed. Iowa State University Press, Ames, IA.

Sunter, A.B. (1977a). Response burden, sample rotation, and classification renewal in economic surveys. *International Statistical Review* **45**, 209–222.

Sunter, A.B. (1977b). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics* **26**, 261–268.

Sunter A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review* **54**, 33–50.

Sverchkov, M. and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology* **30**, 79–103.

Swensson, B. (1982). A survey of nonresponse terms. *Statistical Journal of the United Nations ECE* **1**, 241–251.

Tallis, G.M. (1978). Note on robust estimation infinite populations. *Sankya C* **40**, 136–138.

Tam, S.M. (1986). Characterization of best model-based predictors in survey sampling. *Biometrika* **73**, 232–235.

Tam, S.M. (1988). Some results on robust estimation in finite population sampling. *Journal of the American Statistical Association* **83**, 242–248.

Ten Cate, A. (1986). Regression analysis using survey data with endogenous design. *Survey Methodology* **12**, 121–138.

Tepping, B.J. (1968). Variance estimation in complex surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 11–18.

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association* **94**, 635–644.

Théberge, A. (2000). Calibration and restricted weights. *Survey Methodology* **26**, 99–107.

Thomas, D.R. (1989). Simultaneous confidence intervals for proportions under cluster sampling. *Survey Methodology* **15**, 187–201.

Thomas, D.R. and Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association* **82**, 630–636.

Thomas, D.R., Singh, A.C. and Roberts, G.R. (1996). Tests of independence on two way tables under cluster sampling: an evaluation. *International Statistical Review* **64**, 295–311.

Thompson, M.E. (1997). *Theory of Sample Surveys.* Chapman & Hall, London.

Thompson, S.K. (2002). *Sampling*, 2nd ed. Wiley, New York.

Thompson, S.K. and Seber, G.A.F. (1996). *Adaptive Sampling.* Wiley, New York.

Thompson, W.R. (1936). On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. *Annals of Mathematical Statistics* **7**, 122–128.

Thomsen, I. (1978). Design and estimation problems when estimating a regression coefficient from survey data. *Metrika* **25**, 27–35.

Tillé, Y. (1998). Estimation in surveys using conditional probabilities: simple random sampling. *International Statistical Review* **66**, 303–322.

Tillé, Y. (1999). Estimation in surveys using conditional probabilities: complex design. *Survey Methodology* **25**, 57–66.

Tillé, Y. (2002). Unbiased estimation by calibration on distribution in simple sampling designs without replacement. *Survey Methodology* **28**, 77–85.

Tillé, Y. (2006). *Sampling Algorithms*. Springer-Verlag, New York.

Tin, M. (1965). Comparison of some ratio estimators. *Journal of the American Statistical Association* **60**, 294–307.

Tollefson, M. and Fuller, W.A. (1992). Variance estimation for samples with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 140–145.

Tremblay, V. (1986). Practical criteria for definition of weighting classes. *Survey Methodology* **12**, 85–97.

Tschuprow, A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* **2**, 646–680.

Tukey, J.W. (1958). Bias and confidence in not-quite large samples (abstract). *Annals of Mathematical Statistics* **29**, 614.

U.S. Bureau of the Census (1986). Content reinterview study: accuracy of data for selected population and housing characteristics as measured by reinterview. *1980 Census of Population and Housing*, PHC 80-E2. U.S. Department of Commerce, U.S. Government Printing Office, Washington, DC.

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey Methodology* **28**, 103–114.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Vijayan, K. (1968). An exact π ps sampling scheme: generalization of a method of Hanurav. *Journal of the Royal Statistical Society B* **30**, 556–566.

Wang, J. and Fuller, W.A. (2002). Small area estimation under a restriction. *Proceedings of the Joint Statistical Meetings*, CD-ROM.

Wang, J. and Fuller, W.A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association* **98**, 716–723.

Wang, J., Fuller, W.A. and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology* **34**, 29–36.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **57**, 622–627.

Watson, D.J. (1937). The estimation of leaf area in field crops. *Journal of Agricultural Science* **27**, 474–483.

Wilks, S.S. (1962). *Mathematical Statistics*. Wiley, New York.

Williams, W.H. (1961). Generating unbiased ratio and regression estimators. *Biometrics* **17**, 267–274.

Williams, W.H. (1963). The precision of some unbiased regression estimators. *Biometrics* **19**, 352–361.

Williams, W.H. (1978). *A Sampler on Sampling*. Wiley, New York.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2nd ed. Springer-Verlag, New York.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association* **47**, 635–646.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* **66**, 411–414.

Woodruff, R.W. and Causey, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* **71**, 315–321.

Wooldridge, J.M. (2006). *Introductory Econometrics: A Modern Approach*, 3rd ed. Thompson/South-Western, Mason, OH.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association* **78**, 879–884.

Wright, T. and Tsao, H.J. (1983). A frame on frames: an annotated bibliography. In: T. Wright (ed.), *Statistical Methods and the Improvement of Data Quality*. Academic Press, New York, pp. 25–72.

Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* **96**, 185–193.

Wu, C.F.J. and Deng, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In: G.E.P. Box et al. (eds.), *Scientific Inference, Data Analysis and Robustness*. Academic Press, New York, pp. 245–277.

Wu, Y. (2006). Estimation of regression coefficients with unequal probability samples. Unpublished Ph.D. dissertation. Iowa State University, Ames, IA.

Wu, Y. and Fuller, W.A. (2006). Estimation of regression coefficients with unequal probability samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 3892–3899.

Xiang, X. (1994). Bahadur representation of kernel quantile estimators. *Scandinavian Journnal of Statistics* **21**, 169–178.

Yamane, T. (1967). *Elementary Sampling Theory*. Prentice-Hall, Englewood Cliffs, NJ.

Yansaneh, I.S. and Fuller, W.A. (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology* **24**, 31–40.

Yates, F. (1948). Systematic sampling. *Philosophical Transactions of the Royal Society A* **241**, 345–377.

Yates, F. (1949). *Sampling Methods for Census and Surveys*. Griffin, London.

Yates, F. (1981). *Sampling Methods for Censuses and Surveys*, 4th ed. Griffin, London.

Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B* **15**, 253–261.

You, Y. and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics* **38**, 431–439.

You, Y. and Rao, J. N. K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference* **111**, 197–208.

Yu, C.L. and Legg, J.C. (2009). Protocol calibration in the National Resources Inventory. *Journal of Agricultural, Biological and Enviornmental Statistics*. To appear.

Yung, W. and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology* **22**, 23–31.

Zaslavsky, A.M., Zheng, H. and Adams, J. (2008). Optimal sample allocation for design-consistent regression in a cancer services survey when design variables are known for aggregates. *Survey Methodology* **34**, 65–78.

Zhang, L.C. (2005). On the bias in gross labour flow estimates due to nonresponse and misclassification. *Journal of Official Statistics* **21**, 591–604.

Zyskind, G. (1976). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics* **38**, 1092–1109.

INDEX

Sampling Statistics. By Wayne A. Fuller

WILEY SERIES IN SURVEY METHODOLOGY

Established in Part by WALTER A. SHEWHART AND SAMUEL S. WILKS

The ***Wiley Series in Survey Methodology*** covers topics of current research and practical interests in survey methodology and sampling. While the emphasis is on application, theoretical discussion is encouraged when it supports a broader understanding of the subject matter.

The authors are leading academics and researchers in survey methodology and sampling. The readership includes professionals in, and students of, the fields of applied statistics, biostatistics, public policy, and government and corporate enterprises.

*Now available in a lower priced paperback edition in the Wiley Classics Library.

LESSLER and KALSBEEK · Nonsampling Error in Surveys
LEVY and LEMESHOW · Sampling of Populations: Methods and Applications, *Fourth Edition*
LYBERG, BIEMER, COLLINS, de LEEUW, DIPPO, SCHWARZ, TREWIN (editors) · Survey Measurement and Process Quality
MAYNARD, HOUTKOOP-STEENSTRA, SCHAEFFER, VAN DER ZOUWEN · Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview
PORTER (editor) · Overcoming Survey Research Problems: New Directions for Institutional Research, No. 121
PRESSER, ROTHGEB, COUPER, LESSLER, MARTIN, MARTIN, and SINGER (editors) · Methods for Testing and Evaluating Survey Questionnaires
RAO · Small Area Estimation
REA and PARKER · Designing and Conducting Survey Research: A Comprehensive Guide, *Third Edition*
SARIS and GALLHOFER · Design, Evaluation, and Analysis of Questionnaires for Survey Research
SÄRNDAL and LUNDSTRÖM · Estimation in Surveys with Nonresponse
SCHWARZ and SUDMAN (editors) · Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research
SIRKEN, HERRMANN, SCHECHTER, SCHWARZ, TANUR, and TOURANGEAU (editors) · Cognition and Survey Research
SUDMAN, BRADBURN, and SCHWARZ · Thinking about Answers: The Application of Cognitive Processes to Survey Methodology
UMBACH (editor) · Survey Research Emerging Issues: New Directions for Institutional Research No. 127
VALLIANT, DORFMAN, and ROYALL · Finite Population Sampling and Inference: A Prediction Approach