INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

UNIVERSITY OF CALIFORNIA RIVERSIDE

Econometric Analysis of Household Surveys

A Dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy
in
Economics
by
Robert V. Breunig
June, 1998

Dissertation Committee:

Dr. Aman Ullah, Chairperson

Dr. Keith Griffin

Dr. Gloria Gonzalez-Rivera

Dr. R. Robert Russell

The Dissertation of Robert V. Breunig is approved:

Chairperson

University of California, Riverside

Acknowledgements

I would like to thank George Durnford for getting me interested in Economics, and Professor Dennis Moran and Professor Gloria Rudolf for teaching me to think critically. I could not have possibly completed my thesis without the help of Tala Martinez, Lilia Liderbach-Vega, Martha Ponce, Benicia Chatman, and especially Karen Smith. All the faculty in the economics department have been supportive of me. I would particularly thank Professor Jang-Ting Guo, who knew when to be critical and when to be supportive and is the kind of professor that every student would dream of having. I credit Walter Oakes, Jeffrey Schoonover, and Norris Turner for keeping me laughing and generally preserving my sanity.

All of the members of my committee, from whom I have benefited intellectually and personally, deserve my heartfelt thanks. Professor Keith Griffin provided helpful and timely comments and was a wonderful sounding board. Professor Gloria Rudolf was always available for me and very helpful with her comments. Professor R. Robert Russell was generous with his time, was a pleasure to work and write with, and a great mentor. I had a fantastic chair in Professor Aman Ullah, who was tremendously helpful, always available, and has provided me with a wealth of suggestions for research ideas.

Chapter 2 contains material that appears in *Handbook of Applied Economic Statistics*. Aman Ullah, the co-author listed in that publication, directed and supervised the research that forms the basis for this dissertation.

Véronique Danjou provided me daily with unflagging support and encouragement. She made the bad days livable and the good days great.

ABSTRACT OF THE DISSERTATION

Econometric Analysis of Household Surveys

by
Robert V. Breunig

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 1998
Professor Aman Ullah, Chairperson

Survey data is widely used in economics to draw conclusions about policy and consumer behavior. A large number of techniques, such as sampling without replacement, stratified sampling, cluster sampling, and systematic sampling, have been developed, and are employed (individually or in combination) in the surveys used to gather economic data. Despite the prevalence of such survey techniques (particularly in the data used by labor and development economists) relatively little attention has been paid to its analysis.

We show that standard econometric techniques which fail to account for the survey structure of the data lead to biased and inconsistent estimates. Standard errors estimated in the usual way will be incorrect, thus inferences drawn from inappropriately treated survey data may well be inaccurate.

Three distinct areas of interest to applied economists and econometricians are considered in this paper. After a brief introduction to the history of survey sampling, we

show through simulation the large bias which arises in estimates of simple parameters such as means and variances when the sample structure is ignored. Techniques, some of which are available in the statistics literature, are provided for conducting unbiased estimation and proper inference.

In the next section, these results are extended to inequality measures. Stratification and unequal probability sampling are shown to lead to biases of 20% or more in commonly used inequality measures when the sampling structure is ignored. Methods are developed for unbiased estimation of inequality measures and these are shown to perform well in simulation. We provide a method for calculating a "design effect" which can be used to inflate the usual standard errors for inequality measures to account for clustering in the data. This method is applied to the Coefficient of Variation, a frequently used inequality measure. The techniques in this section are applied to survey data on household income and expenditure from China, Mexico, and Kenya.

In the last section, we consider non-parametric density estimation under stratified and clustered samples. We develop a weighted, non-parametric density estimator to account for unequal probability sampling and provide data-based methods for choosing a new optimal bandwidth parameter based upon the survey structure of the data. These methods are shown to perform well in simulation for stratified and for clustered data.

Econometric Analysis of Household Surveys

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

List of Tables

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Figures

List of Figures (continued)

List of Figures (continued)

1 INTRODUCTION

The past fifty years have seen a tremendous growth in research in theoretical and applied econometrics. Particularly in the last twenty years, the explosive growth of research in econometrics has resulted in the appearance of two distinct sub-fields within econometrics. New time-series methodologies have been developed in response to empirical and data issues in macroeconomics and finance. This has led to the creation of and widespread growth of research in methods to analyze non-stationary data, conditional heteroscedasticity., high frequency data, and regime switches/structural breaks to name just a few areas.

Parallel developments have occurred in response to changes in labor economics, development economics and applied microeconomics. Cross-sectional econometric methods have evolved to deal with simultaneous equation models and endogeneity problems, truncated data, discrete data, and self-selection. These theoretical developments are now widely applied in labor and development economics and have considerably added to the quality of quantitative analysis.

Despite these developments, econometric inference methods (and this is particularly the case for cross-sectional econometrics) have been confined to the assumption of data being generated as a simple random sample with replacement, or that the data are coming from an infinite population, see Johnston (1991), Greene (1993), and Davidson and MacKinnon (1993). This is certainly an invalid assumption in the case

of most applied cross-sectional econometrics. The vast majority of applied econometric work in development and labor (as well as in many other areas) uses survey data which violates the assumption of random sampling from an infinite population. Surveys are generally conducted on finite populations, which are first subdivided into many groups, and samples may be drawn using different methods for different sub-groups. In most cases, population elements enter the sample with unequal probabilities. In addition, commonly used sampling methods introduce large correlation into cross-sectional data–a problem which most econometricians seem to believe exists only in time series data.

The use of sampling schemes which differ substantially from random sampling with replacement (or from an infinite population) has long been a topic of statistical research. The past four decades have seen an extensive literature on systematic sampling, stratified random sampling, and cluster sampling, either alone or in combination. See the now-classic books of Kish (1965), Cochran (1953), and Sukhatme and Sukhatme (1984) for the early developments in this literature. For a more modern approach, see Levy and Lemeshow (1991) or Thompson (1992)–both of which are widely used in statistics courses. An excellent, concise text, though highly mathematical, is Gourieroux (1981). Two accessible texts for the non-statistician are Kalton (1983) and Dalenius (1988).

This statistics literature has developed over the past 150 years, prodded on by questions of data and empirical estimation. We give a brief overview of this history

below. In section 1.3, we will examine some of the most commonly used survey designs to see how they are implemented in practice. Despite the existence of this literature in statistics on estimation and inference under various sampling designs, very little application of these results in econometric analysis has occurred. For economists, another problem is that the statistics literature has been overwhelmingly concerned with estimating means and variances under different sampling schemes. The effects of the survey design on regression, inequality measurement, non-parametric density and regression estimation, and other parameters of interest to economists have not been dealt with in the statistics literature. In section 1.4, we will preview the contents of this paper, where we extend the results from the mean model to other models of greater interest to economic analysts.

1.1 History

The history of survey sampling can be traced back to the early eighteenth century, and even earlier–see Hansen (1987), Bellhouse (1988a), and Deaton (1997) for detailed references. The idea of gathering a sample of data to estimate a population total or a general model is one that goes back quite far, though the preferred method was always considered to be total enumeration. (The early French statistician Quetelet is the first known person to express the view of a complete census being the optimal sample–see Stigler (1986).)

Stigler (1954) surveys the early use of samples to study consumer behavior. He

describes how Reverend David Davies, in 1795, and Sir Frederick Morton Eden, in 1797, gathered non-representative samples to study the living conditions of the working classes and poor in industrial revolution-era England. Another example of such sample gathering is Engel (1857), who used a non-representative sample of 200 Belgian households gathered by Ducpetiaux to establish that the share of the budget allocated to food is higher for the poor. One of the first known expositions of the idea of gathering sample data to estimate population totals was LaPlace, who was interested in estimating the population of France. (His methodology and ideas are described in Cochran (1978).)

The Norwegian statistician A.N. Kiaer is generally credited with putting forward the idea of random sampling by a "representative method" as an effective and cost-efficient method of gathering a sample to estimate a population total. Between 1895 and 1903, Kiaer relentlessly pursued his ideas (despite the fierce opposition of many eminent statisticians of the day) before the International Statistical Institute (ISI) at their annual meetings. He met with only limited success, however, as representative sampling was not adopted as a practical technique.

Kiaer's (1897) (see Kiaer (1976) for translation) paper was perhaps the first in which a large-scale, representative sample was used in practical application. He also discussed the principles, uses, and limitations of various sampling designs, though not in the language we use to describe such designs today. Kiaer did not introduce the idea of randomization in surveys, but merely stressed the selection of representative

samples. Lucien March, another French statistician, introduced the concepts of cluster sampling and simple random sampling without replacement in his discussion of Kiaer's paper at the International Statistical Institute meeting of 1903. This was the same meeting where the ISI passed a resolution supporting and promoting the use of the "representative method" of sampling. Despite his role in the development of probability models for sampling, March remained skeptical of their usage.

From 1903 to 1925, randomization and representative sampling went undiscussed at the ISI meetings. The turning point for the use of survey sampling came in the 1920s, though much of the groundwork was laid by Arthur Bowley, who promoted and implemented Kiaer's ideas throughout the first three decades of this century.

Bowley (1913) conducted a study of poverty in Reading in 1912 based upon a survey sample gathered through randomization techniques and compared his studies with those of other statisticians who did not use survey techniques. Bowley (1907) also came perilously close to discovering a central limit theorem for simple random sampling, providing an empirical demonstration through random sampling from a list of numbers in an almanac.

Bowley was quite active in the commissioning of an ISI study on representative sampling conducted in 1924 and reported at the ISI meetings in 1925. (The results of this commission are discussed extensively in Yates (1946).) The ISI at that point passed a resolution supporting both random and purposive sampling techniques. Bowley (1926) provided a summarization of survey techniques available at the time and

laid out the framework for stratified sampling with proportional allocation.[1]

The breakthrough in the 1920s continued with the pathbreaking work in statistical estimation theory and practice of R. A. Fisher at the Rothamsted experimental station (Fisher (1925)). Fisher emphasized randomization, replication, and stratification in sampling design. His work lead to the calculation of statistical estimates and their precision by Yates and Zacapony (1935) and Cochran (1939). Indeed, Fisher's work paved the way for Neyman's (1934) classic paper which, for the first time, gave a systematic discussion of inference from random samples drawn from a finite population, contained a comparison of purposive sampling and random sampling, introduced the concept of the confidence interval, established the asymptotic normality of the sample average, and provided the optimal sample sizes within strata independently of Tschuprow's (1923) work. Later Neyman (1938) developed the theory of what is known as "two-stage" sampling.

In the U.S., around this period, important research work on sampling design was conducted by the researchers at the Department of Agriculture and the statistical lab at Iowa State University. Morris Hansen and William Hurwitz made their impact on the development of survey methods at the U.S. Bureau of the Census. (Some of these developments are discussed in Hansen, et. al. (1983).) They were quite instrumental

[1] It is interesting to note that the influential work of Gini (1928) on the Gini coefficient was undertaken at this time, though Gini did not use representative methods. He was affected by the discussions however, as he purposively selected certain data out of the Italian census of 1921 with the intention of matching sample averages of seven economic variables with their known population values. His method did not work very well, however, as his sample averages differed radically from the census averages for many other important variables.

in gaining acceptance of survey methods in many government agencies. Duncan and Shelton (1978) provide a discussion of this very interesting period in statistics history. Hansen's introduction of sampling methods were widely fought on political grounds and his diplomacy is credited with gaining acceptance for these methods. The politics surrounding the introduction of sampling in the 1940 U.S. Census (and the discussion of them by Duncan and Shelton) make for interesting reading, particularly in the light of current attempts by the Republican Party in the U.S. to prevent the use of sampling in the 2000 Census.

In the 1930s and 1940s, labor force surveys dealing with issues such as measuring employment and labor force participation were under particular pressure to provide accurate estimates at low cost. These surveys, using a low-cost, systematic sampling design, were necessitated by the Great Depression–see Stephan (1948) for detailed references. Sampling with probability proportional to measures of size at the successive stages of sampling was also introduced during the work on labor force studies of that period.

In the 1930s, parallel developments on the application of survey sampling took place under the leadership of P. C. Mahalanobis at the Indian Statistical Institute and P. V. Sukhatme at the Indian Council of Agriculture Research. Mahalanobis introduced the concepts of developing sampling designs based on cost and variance estimates, and methods of evaluating survey errors. His famous survey of the Bengali jute crop (Mahalanobis (1940)) and his ability to predict accurately the total output

of jute from samples of less than 5% are well known.

The work during the 1930s and the 1940s revolutionized the collection of household survey data after the war. Major developments include the first national sample survey data developed annually (1950-1970) and then every five years at the Indian Statistical Institute, the household survey data now collected in U.S., U.K., and Taiwan, the Living Standard Measurement Study (LSMS) survey of Peru and the Ivory Coast by the World Bank, and Malaysian family survey data by the Rand Corporation. These household data are now extensively used in development economics to study poverty, income distribution and economic welfare.

Another important development since the 1960s has been the challenge presented to the dominant framework by V. P. Godambe–first laid out in his work Godambe (1952). He has proposed a model-based framework instead of the finite population based framework upon which the bulk of statistical theory for survey sampling has been based. He proposes using estimating functions to get at model parameters instead of the classic approach which has focused on estimating finite population parameters by correctly inflating sample values. For his view of the literature on estimation in finite populations, see Godambe (1976). His challenge to the dominant paradigm is laid out in a series of articles, Godambe (1995) and Godambe (1997), and in a collection of works, edited by Godambe (1991).

1.2 Survey Designs

Consider the problem of designing a survey to estimate some variable from a population. Statistical theory has shown that it is possible to get quite accurate estimates of many variables without appealing to a population-wide census. By selecting a sample of elements from the population and estimating the variable(s) of interest from the sample, one can get a good idea of the true underlying parameters.

The obvious starting point is to make a listing of all the elements in the population. This is typically called a frame, and in a household survey it would be a list of all the households in the population. It may be built upon phone numbers, driver's licenses, tax returns, etc. Obviously making a complete listing is not necessarily straightforward. A list of all phone numbers, for example, would exclude people without phones. We will not deal with problems of frame and frame selection here, but will refer interested readers to the general texts on sampling listed at the end of this section. Many techniques exist for solving problems of missing elements–from supplemental door-to-door interviews to building frames from a combination of lists. For our purposes, we will assume the existence of a frame which includes all the elements of the population, while realizing that this will rarely hold in practice.

How should one go about conducting a survey? One possibility is to randomly select elements, with equal probability of selection assigned to each element, out of the entire pooled population. Once an element is chosen, it may either be included in the

population for future random selections or it may be excluded from the population. If an element is included in the pool for future selections after having been selected once, the probability of any element being sampled on any draw will be the same for all draws. This is called random sampling with replacement, and is the typical econometrics textbook assumption (often implicit) of how the sample has been gathered. If an element is excluded, then the probability of selection for elements will change at each draw. An element already drawn, for example, will have zero probability of selection in future draws. An element not drawn, will have a slightly higher probability of selection for future draws, since there is now one less element in the entire pool. This is called random sampling without replacement. Both random sampling with replacement (RSWR) and random sampling without replacement (RSWOR) will be discussed in detail below.

Another possibility is that the population may be subdivided into different units prior to sampling. Then a separate sample may be taken in each unit. For example, in the United States, the entire population may be divided up by state or by county and then samples drawn in each different geographic unit. For a survey of the labor force, we may first divide the working population into different industries or job categories. For a survey of banking practices, the population may be divided into Federal Deposit Insurance Corporation (FDIC) member and non-member banks.

Several reasons exist why the population may be so sub-divided. Perhaps different populations will be surveyed by different administrative units and therefore having a

clear demarcation of jurisdiction is important. Different states may have very different systems of listing population and keeping them different may save on administrative problems. There also may be reasons connected to the nature of the survey. Suppose that we want to compare the behavior of FDIC-member and non-member banks. By first sub-dividing the population and then drawing a sample within each sub-population, we can guarantee a large enough sample from each population of interest to conduct our analysis.

This type of sampling is called stratified sampling. The overall population is first divided into sub-populations, called strata, and then a separate sample is taken in each stratum. The stratum-specific samples may or may not be of equal size and different techniques may be used in different strata. Often strata with small populations are relatively more heavily sampled to provide enough observations to make accurate inference. As we will see below, if the proportion of elements selected in each stratum is proportional to that stratum's population, stratified sampling poses no particular challenge. However, when sampling is not proportional, parameter estimates from stratified samples may be biased and inconsistent.

It would be easy for policy analysts and social scientists if every survey could be designed so as to maximize the amount of information available in the survey. Realistically, surveys are usually conducted under fixed budgets. Survey practitioners are often quickly-trained students, without specialized knowledge of survey design. For both ease of implementation and to save money, surveys are often conducted in

several stages. By this, we mean that the random selection of elements is done in several stages and that the elements over which the random selection is being done are not necessarily the ultimate units of interest.

An example will serve to clarify. Consider a survey of dairy farms in Minnesota and Wisconsin to assess the impact of bovine growth hormone usage in the Midwest. Assume that a desired sample size of 100 has been chosen. One possible way to conduct the survey would be to lump all dairy farms in Minnesota together and randomly select 100 farms to be surveyed. A cheaper and easier alternative would be to randomly select 20 townships in Minnesota and then within each of the 20 selected townships, to further select five farms for inclusion in the survey. It is easy to see that the first method may well involve travelling to 80 to 90 different far-flung townships while the second would involve only 20 such trips.

This type of survey is called cluster sampling. It is also referred to as two-stage sampling, since there are two successive levels of sample selection. We generally think of clusters as being geographical units or units of the population which are in proximity to one another, though this may not be the case. (We may be selecting clusters of records out of a computer which, though sequentially listed in the computer, are really thousands of miles apart in real space.)

Cluster sampling may also be implemented by choosing an element randomly and then including nearby elements in the sample. For a household survey, for example, a single household is randomly selected from the population and then a number of

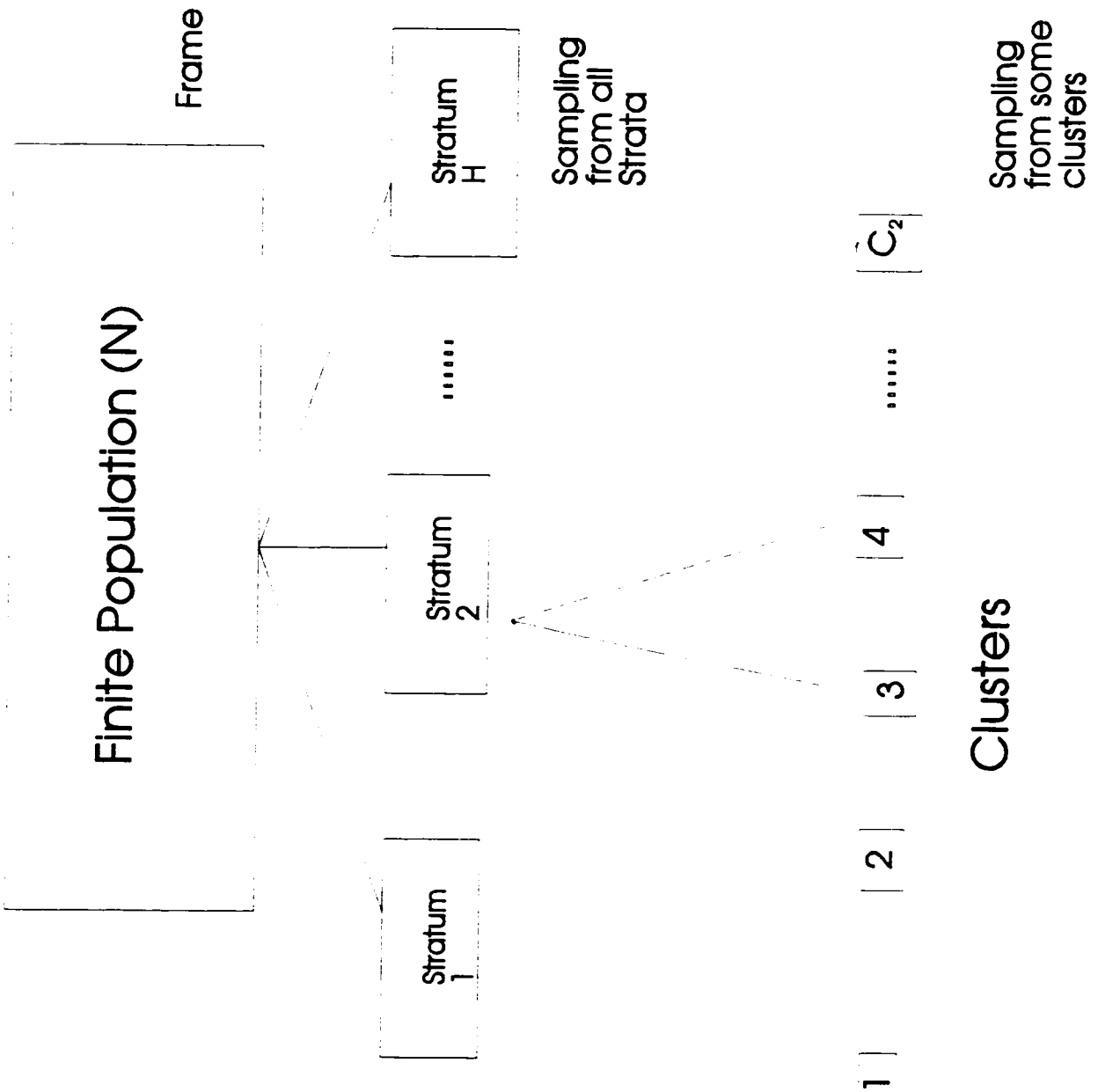
households in the immediate proximity are also included in the sample. Using this method, for example, one could draw 100 households randomly from the population and turn this into a sample of size 1000 by including 9 nearby houses in the survey for each household chosen in the original sample.

Cluster sampling introduces dependence into the data, and elements in the same cluster will tend to exhibit positive correlation. Households within the same village (or cluster) can be assumed to face similar conditions–for example we expect heating fuel costs to be correlated for households in the same area.

Yet another variant of cluster sampling is called systematic sampling. The great advantage of systematic sampling is that it involves making only one random selection from the population. Once that selection is made, every k-th element in the population will be sampled until the desired sample size is reached. This method is quite easy to teach to survey gatherers in the field: "pick one of the first five houses randomly and then survey every fifth house in order." This directive does not take a degree in statistics to implement!

Figure 1.1 provides a visual picture of the framework which has been presented here. We have some finite population of elements, listed in a frame. Selection may occur at this level, either with or without replacement, or it may occur in the various sub-population levels. We have a first layer of sub-populations called strata which exhaustively cover the entire finite population.

FIGURE 1.1
Complex Sampling in Finite Population Framework



If we employ stratified sampling, some elements from each stratum are surveyed. Within each stratum, we have clusters of elements that may be geographical clusters or job title clusters or diagnosis clusters, etc. In cluster sampling, in contrast with stratified sampling, some but not all clusters are selected.

There are several excellent statistics books which discuss various sampling schemes such as systematic sampling, stratified random sampling, and cluster sampling in detail: see Kish (1965), Sukhatme and Sukhatme (1984), and Thompson (1992) among others. Kalton (1983) and Dalenius (1988) are nice introductions to the subject. In this paper, I assume that the data has already been gathered and that the analyst has information about the structure of the data.

Economists use survey data all the time. Many commonly used data sets considered by economists such as the Living Standards Measurement Surveys (LSMS) of the World Bank, the Survey on Income and Program Participation (SIPP) of the US Census Bureau, and the Labor Force Survey (LFS) of Statistics Canada are gathered through surveys which combine techniques of stratification, clustering, and/or systematic sampling.

Table 1.1 provides a list of data sets which have been widely analyzed by economists. As one can see, all of them deviate significantly from the RSWR assumption. However, little or no attention is paid by most applied economists to the problems which arise from the survey structure of the data.

Table 1.1
Sample Design of some Household Surveys

Sample	Stratified?	Clustered?	Stages
SRS	No	No	One
Pakistan (1991)	Yes	Yes	Two
Ghana (1987)	No	Yes	Two
Russia (1993)	No	Yes (twice)	Three
India (1976)	Yes	Yes	Two
Kenya (1986)	Yes	Yes	Two
China (1989)	Yes	No	One
Canada (LFS)	Yes	Yes	Two
Mexico (1989)	Yes	Yes	Two
U.S. SIPP (1989-92)	Yes	Yes	Two

Despite three decades worth of developments in statistical inference based on survey data, econometric analysis is carried on under the false assumption of RSWR, although for notable exceptions see the excellent works of Pudney (1989), Deaton (1997), and Howes and Lanjouw (1994). This is especially a matter of concern in development economics where measures of income inequality, poverty and elasticity are used in policy-making by governments and international agencies. When survey data are analyzed using standard econometric techniques, large problems of bias and incorrect standard errors may arise. Policy-making may thus be significantly affected.

The failure to take into account information about the survey design when conducting econometric analysis may be due to the statistical complexity of the various sampling designs, which makes analysis difficult for the average applied economist. It may also be due to the almost total lack of exposure to the statistical literature on survey design in econometrics texts. Given this deficiency, a systematic development of the parametric and non-parametric econometric inference (estimation and testing) of various econometric models, under various practical sampling designs, is needed. This paper is a beginning step towards that goal.

1.3 Preview

The plan of this paper is as follows. In Section 2 we present the estimation of the finite population mean and inequality measures under random sampling without replacement (RSWOR), stratification and clustering. Ullah and Breunig (1998) have provided a unified econometric framework of the five decades of diverse statistical literature on estimating the population mean. These results are summarized here and the implications of mis-specifying the sampling design for mean estimation are considered. These results are then extended to several other models. Section 3 deals with the estimation of inequality and poverty in complex samples. We first consider the bias problem which arises in estimating inequality from a small sample and how to correct for that bias. We then consider inequality estimation and inference for RSWR, RSWOR, stratified, and clustered samples. Section 4 examines non-parametric kernel density estimation for RSWOR, stratified and clustered sampling. In Section 5, we provide a brief conclusion and point the way to future areas of research.

2 EFFECT OF SAMPLING DESIGN ON ESTIMATION AND INFERENCE IN MEAN MODEL

2.1 Mean Model

Let us consider a finite population of size N for an economic variable Y, and write the population mean model as

$$Y_i = \mu + U_i \ , \ i = 1, ..., N \tag{1}$$

where Y_i is the *i-th* population observation, U_i is the *i-th* error, and μ and σ^2 are the population mean and variance, respectively, given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i \ , \ \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2 = \frac{N-1}{N}S^2, \tag{2}$$

$S^2 = \sum_{i=1}^{N}(Y_i - \mu)^2/(N-1)$. The non-sampling errors, U_i, sum to zero by the definition of μ in (2). U_i and Y_i are therefore non-stochastic variables. However, if we treat the finite population model (1) as having been generated from an infinite population or super-population model, then U_i and Y_i are stochastic.

We will consider the estimation of the mean when the usual textbook assumption of sampling with replacement (or sampling from an infinite population) is violated. We will first consider random sampling without replacement for a finite population, and then the general case of unequal probability sampling. We then consider the most

commonly used techniques for gathering economic data–stratification, clustering, and systematic sampling–and how our estimation and inference about the mean must change from the random sampling with replacement case.

2.2 Random Sampling Without Replacement (RSWOR) and Random Sampling With Replacement (RSWR)

A random sample without replacement of size n, often referred to in the statistics literature as a simple random sample (SRS), is taken from the above finite population. We denote the sample observations as y_i and write for these observations

$$y_i = \mu + u_i \ , \ i = 1, ..., n \tag{3}$$

where u_i is now the *i-th* sampling error. We may also write this in more compact form,

$$y = \iota\mu + u \tag{4}$$

where ι is an $n \times 1$ vector of ones.

Since the sampling is without replacement,

$$\pi_r(i) = P[y_i = Y_r] = \frac{1}{N} \tag{5}$$

is the probability that the *r-th* population unit is selected in the *i-th* draw and

$$\pi_{rs}(i,j) = P[y_i = Y_r \text{ and } y_j = Y_s] = \frac{1}{N(N-1)} \tag{6}$$

is the probability that the (r,s)*-th* unit is selected in the (i,j)*-th* draw where

$i, j = 1, ..., n$ and $r, s = 1, ..., N\,(i \neq j, r \neq s)$. In a sample of size n.

$$\pi_r = \sum_{i=1}^{N} \pi_r(i) = \frac{n}{N} \tag{7}$$

will be the probability of selection of the *r-th* population unit in the sample and

$$\pi_{rs} = \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \pi_{rs}(i,j) = \frac{n}{N}\left(\frac{n-1}{N-1}\right) \tag{8}$$

the probability of selection of the (r, s)*-th* population unit in the sample.

In view of (3) to (8), we get

$$E\,u_i = 0, E\,u_i^2 = \sigma^2 = \frac{N-1}{N} S^2 \tag{9}$$

and, for $i \neq j$,

$$E\,u_i\,u_j = \sigma_{12} = -\frac{\sigma^2}{N-1} = -\frac{S^2}{N} = \rho\sigma^2 \tag{10}$$

where $\rho = -1/(N-1)$ is the intra-class population correlation between y_i and y_j.

It is easy to verify that

$$\begin{aligned}
E\,u_i^3 &= \gamma_1\sigma^3 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^3, \qquad (11)\\
E\,u_i^4 &= (\gamma_2 + 3)\sigma^4 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^4\\
E\,u_i^2\,u_j &= \sigma_{112} = \frac{1}{N(N-1)}\sum_{\substack{i=1 \\ i\neq j}}^{N}\sum_{j=1}^{N}(Y_i - \mu)^2(Y_j - \mu) = -\frac{\gamma_1\sigma^3}{N-1}\\
E\,u_i^2\,u_j^2 &= \sigma_{1122} = \frac{1}{N(N-1)}\sum_{\substack{i=1 \\ i\neq j}}^{N}\sum_{j=1}^{N}(Y_i - \mu)^2(Y_j - \mu)^2\\
&= \frac{N - (\gamma_2 + 3)}{N-1}\sigma^4\\
E\,u_i^3\,u_j &= \sigma_{1112} = \frac{1}{N(N-1)}\sum_{\substack{i=1 \\ i\neq j}}^{N}\sum_{j=1}^{N}(Y_i - \mu)^3(Y_j - \mu) = -\frac{(\gamma_2 + 3)\sigma^4}{N-1}
\end{aligned}$$

and for $i \neq j \neq k \neq l$

$$
\begin{aligned}
E\,u_i\,u_j\,u_k &= \sigma_{123} = \frac{1}{N(N-1)(N-2)} \sum_i \sum_{\substack{j \\ i \neq j \neq k}} \sum_k (Y_i - \mu)(Y_j - \mu)(Y_k - \mu) \\
&= \frac{2\gamma_1}{(N-1)} \frac{\sigma^3}{(N-2)} \\
E\,u_i\,u_j\,u_k\,u_l &= \sigma_{1234} = \frac{1}{N(N-1)(N-2)(N-3)} \sum_i \sum_{\substack{j \\ i \neq j \neq k \neq l}} \sum_k \sum_l \qquad (12) \\
&\quad (Y_i - \mu)(Y_j - \mu)(Y_k - \mu)(Y_l - \mu) \\
&= \frac{3[N - 2(\gamma_2 + 3)]}{(N-1)(N-2)(N-3)} \sigma^4 \\
E\,u_i^2\,u_j\,u_k &= \sigma_{1123} = \frac{1}{N(N-1)(N-2)} \sum_i \sum_{\substack{j \\ i \neq j \neq k}} \sum_k (Y_i - \mu)^2 (Y_j - \mu)(Y_k - \mu) \\
&= \frac{2(\gamma_2 + 3) - N}{(N-1)(N-2)} \sigma^4
\end{aligned}
$$

where γ_1 and γ_2 are Pearson's measures of skewness and excess kurtosis (see Kendall and Stuart, 1977). For normal distribution, $\gamma_1 = \gamma_2 = 0$. From (10) through (12) it is clear that random sampling without replacement represents a set of n identically distributed but correlated, random variables y_i.

In the case of random sampling with replacement (RSWR) the draws are independent and thus

$$
E\,u_i = 0,\ E\,u_i^2 = \sigma^2,\ E\,u_i^3 = \gamma_1 \sigma^3,\ E\,u_i^4 = (\gamma_2 + 3)\sigma^4 \qquad (13)
$$

because $\sigma_{12} = \sigma_{1112} = \sigma_{123} = \sigma_{1234} = \sigma_{1123} = 0$ and $\sigma_{1122} = \sigma^4$. This also holds if we are sampling from an infinite population ($N \to \infty$)— the case usually considered in econometrics. In what follows we analyse the effect of assuming RSWR when the

true design is RSWOR.

2.2.1 Estimation of Parameters: RSWOR

From (4) using (9) and (10) we have

$$E\{uu'\} = V(u) = \sigma^2\Omega = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \\ \rho & \rho & & 1 \end{bmatrix}. \tag{14}$$

The n x n matrix Ω can be rewritten as

$$\Omega = (1-\rho)[I + \frac{\rho}{1-\rho}\iota\iota'] \tag{15}$$

which has inverse

$$\Omega^{-1} = \frac{1}{(1-\rho)(1+\rho(n-1))}[(1+\rho(n-1))I - \rho\iota\iota']. \tag{16}$$

The least squares (LS) estimator of μ in (2) is obtained by minimizing $u'u$ with respect to μ. This gives

$$\bar{y} = (\iota'\iota)^{-1}\iota' y = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{17}$$

The estimator $\bar{y}$ is unbiased, $E\bar{y} = \mu$, and its variance is

$$\begin{aligned} V(\bar{y}) &= \sigma^2(\iota'\iota)^{-1}\iota'\Omega\iota(\iota'\iota)^{-1} \\ &= \frac{\sigma^2}{n}[1+(n-1)\rho] \\ &= \frac{N-n}{nN}S^2 = \frac{1}{n}(1-\frac{n}{N})S^2 \end{aligned} \tag{18}$$

where the last equality gives the familiar expression of the variance of the sample mean under RSWOR. The term $(1 - \frac{n}{N})$ is known as the finite population correction (fpc).[2] We can see that as the sample size converges to the population, that the variance of $\overline{y}$ will become zero.

The efficient generalized least squares (GLS) estimator of μ is

$$\overline{y}_{GLS} = (\iota'\Omega^{-1}\iota)^{-1}\iota'\Omega^{-1}y = (\iota'\iota)^{-1}\iota' y \tag{19}$$

where the second equality follows by using the expression given above for the inverse of Ω. Thus $\overline{y}_{GLS}$ and $\overline{y}$ are the same.

When the sampling is RSWR or if the population is infinite, $\Omega = I$ because $\rho = 0$. In this case, $E\overline{y} = \mu$ and

$$V(\overline{y}) = \frac{\sigma^2}{n}, \tag{20}$$

which also follows from the last equality of (18) where $\frac{n}{N} \to 0$ as $N \to \infty$.

From (18) and (20)

$$\frac{V(\overline{y}_{RSWOR})}{V(\overline{y}_{RSWR})} = [1 + (n-1)\rho] = \frac{N-n}{N-1} \leq 1 \tag{21}$$

The above results indicate that the LS estimator $\overline{y}$ is unbiased for both RSWOR and RSWR. However, if the sampling is actually without replacement, the variance formula in (20) is wrong and gives an over-estimate of the correct variance (14). To obtain the correct variance, we calculate an unbiased estimator of S^2,

[2] The finite population correction was first noted by Isserlis (1918), though one rarely sees him credited with the result.

$s^2 = (n-1)^{-1}\sum(y_i - \bar{y})^2$ and then deflate (20) by $(N-n)/(N-1)$. For example if $n = 20$ and $N = 100$ the correct variance will be approximately 20% smaller than the variance of a sample of size 20 drawn with replacement. The smallness of the variance of $\bar{y}_{RSWOR}$ is due to the negative correlation ρ, which results from the way in which the sampling is conducted (without replacement).

Of course, the finite population correction will not have a significant effect on our analysis in the case where the sample size is quite small relative to the overall population size. More significant deviations from the RSWR case include potential bias which arises from unequal probability sampling and correlation induced in the data by clustered sampling. It is to these problems that we turn now.

2.3 Sampling with Unequal Probabilities

It is often the case with data used by economists that not all elements of the population enter the sample with the same probability. This may be due to intentional over- or under-sampling of certain segments of the population, or it may be due to differing response rates from different parts of the population. In this section, we consider the general problem of how to conduct unbiased estimation of the mean when sampling is done with unequal probabilities. Below, we will consider the specific case of unequal probability sampling arising from stratification.

When the sampling is with unequal probabilities, $\pi_r(i)$ in (5), π_r in (7), $\pi_{r,s}(i,j)$ in (6), and $\pi_{r,s}$ in (8) are not constants. In this case, we first transform (4) by

$$W^{1/2}y = W^{1/2}\iota\mu + W^{1/2}u \tag{22}$$

and then obtain the LS estimator of μ by minimizing the weighted squared error $u'Wu = (y-\iota\mu)'W(y-\iota\mu)$, where W = Diag. $(w_1, ..., w_n)$ is an $n \times n$ stochastic diagonal weight matrix whose elements w_i, also known as the normalized expansion factors, satisfy $\iota' w \iota = \sum_1^n w_i = 1$. This gives the weighted LS estimator of μ as

$$\bar{y}_w = \iota' W y = \sum_{i=1}^{n} w_i y_i \tag{23}$$

The stochastic weights w_i are chosen such that (using (5) and (7)) the sample is representative of the population in the sense that the sample mean, on average, is identical to the population mean. That is,

$$\begin{aligned} E\bar{y}_w &= E(\sum_{i=1}^{n} w_i y_i) = \sum_{i=1}^{n}\sum_{j=1}^{N} w_j Y_j \pi_j(i) \\ &= \sum_{j=1}^{N} w_j Y_j \pi_j = \mu. \end{aligned} \tag{24}$$

This gives

$$w_i = \frac{1}{N\pi_i} \tag{25}$$

and $\bar{y}_w = N^{-1}\sum_1^n (y_i/\pi_i)$ where π_i is the probability of selection of the *i-th* population unit in the sample. When $\pi_i = n/N$ we get $\bar{y}_w = \bar{y}$ as given in (17).

An alternative way to obtain the weighted estimator of μ is to write

$$\begin{aligned} E\,\overline{y}_w &= E(\textstyle\sum_1^n w_i\, y_i) = E(\sum_1^N w_i\, d_i\, Y_i) \\ &= \textstyle\sum_1^N (w_i . E\, d_i)\, Y_i = \sum_1^n w_i\, \pi_i\, Y_i \end{aligned} \tag{26}$$

where d_i is a dummy random variable which takes value 1 when Y_i is in the sample, and zero otherwise. Since the probability of selection of the i-*th* population unit in the sample is π_i we can verify the following

$$\begin{aligned} E\,d_i &= \pi_i \\ V(d_i) &= \pi_i(1-\pi_i) \\ cov(d_i, d_j) &= \pi_{ij} - \pi_i\,\pi_j \end{aligned} \tag{27}$$

where π_{ij} is defined in (8). This in fact gives us an easier way to calculate the variance of the weighted estimator. Using (27) we see that

$$V(\overline{y}_w) = \sum_{i=1}^{N} w_i^2\, Y_i^2\, \pi_i(1-\pi_i) + \sum_{i\neq j}^{N} w_i\, w_j\, Y_i\, Y_j(\pi_{ij} - \pi_i\,\pi_j). \tag{28}$$

When $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(n-1)}$ we get (18). We point out here again that both the mean and the variance of $\overline{y}_w$ contain Y_i since the error in this finite population model results from the sampling, not from the errors in the superpopulation. (In other words, we are not making any assumption about a data generating process.)

Let us now turn to the specific case of stratified sampling which is frequently found in data used by economists.

2.4 Stratified Sampling

In addition to using sampling without replacement, most surveys use stratified random sampling. The finite population is first divided into different groups (sub-populations), typically by geographical regions, such as rural/urban or states, or by certain characteristics such as blue-collar and white-collar workers. RSWOR is then used within each stratum.

There are several advantages to this method as opposed to RSWOR or RSWR from the entire population. First, the stratification provides a more representative sample overall, and so reduces the variance–especially when the variation within strata is small but the variation between strata is large. Since sampling is done independently in each stratum, the variance of our estimators will only be a function of within-strata variation. Stratification allows for different types of sampling schemes in different strata, desirable perhaps because of cost considerations. For example, one can perform SRS in urban areas where households are closely concentrated, and cluster sampling in rural areas, where households are widely dispersed (see below). Finally, stratification helps to obtain enough sample observations from small sub-populations of special interest.

The population mean model (re-writing (1) above) for stratified sampling can be written as

$$Y_{hi} = \mu_h + U_{hi}\,,\ h = 1, ..., H,\ i = 1, ..., N_h \tag{29}$$

where Y_{hi} is the *i-th* unit in the *h-th* stratum, μ_h is the mean of the *h-th* stratum

$$\mu_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi} \tag{30}$$

and U_{hi} is the error. The variance of the *h-th* stratum is

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 = \frac{N_h}{N_h - 1} \sigma_h^2 \tag{31}$$

In a more compact form we can write the population model for the *h-th* stratum

$$Y_h = \iota_h \mu_h + U_h \tag{32}$$

where Y_h is an $N_h \times 1$ vector of observations, U_h is an $N_h \times 1$ vector of errors, and ι_h is an $N_h \times 1$ vector of ones. The population size is $N = \sum_1^H N_h$.

The stratified sample observations, generated by RSWOR in each stratum, follow

$$y_{hi} = \mu_h + u_{hi}, \qquad h = 1, ..., H, \qquad i = 1, ..., n_h \tag{33}$$

or more compactly

$$y_h = \iota_h \mu_h + u_h, \tag{34}$$

$$V(u_h) = \sigma_h^2 \Omega_h \tag{35}$$

where ι_h is $n_h \times 1$ vector of unit elements, y_h is an $n_h \times 1$ vector of sample observations (y_{hi}) in the *h-th* stratum, and Ω_h is Ω in (15) with $n = n_h$. The total size of the sample is $n = \sum_{h=1}^H n_h$.

2.4.1 Estimation of Parameters: Stratification

The model (34) for the *h-th* subpopulation (stratum) is the same as that of the population model in (4). Thus the results from Section 2.2.1 go through for the estimation of the *h-th* stratum parameters. For example the LS estimator of μ_h is

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} = (\iota_h'\iota_h)^{-1}\iota_h' y_h \tag{36}$$

and

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2. \tag{37}$$

The parameter of interest will usually be the overall mean of the population. That is

$$\mu = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} Y_{hi} = \sum_{h=1}^{H}\theta_h\,\mu_h \tag{38}$$

where $\theta_h = \frac{N_h}{N}$ is the proportion of the total population in the *h-th* stratum. An estimator of μ is then

$$\bar{y}_{st} = \sum_{h=1}^{H}\theta_h\,\bar{y}_h \tag{39}$$

which is unbiased. Further

$$V(\bar{y}_{st}) = \sum_{h=1}^{H}\theta_h^2\,V(\bar{y}_h) = \sum_{h=1}^{H}\theta_h^2\,(1 - \frac{n_h}{N_h})\frac{S_h^2}{n_h}, \tag{40}$$

provided that the sample is chosen independently for each stratum and that sampling is without replacement within each stratum.

When there are many strata, individual strata means, μ_h, may not be of interest. Also, when strata are strictly administrative divisions with no economic interest, we

do not want to estimate each stratum-specific mean individually and then sum as in (39). This may also be impossible or impractical if there are many strata with very small sample sizes from each stratum. In most cases, we are interested in estimating the population parameter μ directly from the pooled sample of data.

To see the least squares (LS) solution of μ we rewrite (33) as

$$y_{hi} = \mu + \mu_h^* + u_{hi} \tag{41}$$

where $\mu_h^* = \mu_h - \mu$, $E\,u_{hi} = 0$, $E\,u_{hi}^2 = \sigma_h^2$, $E\,u_{hi}\,u_{h'i'} = \rho_h\sigma_h^2$ for $h = h'$, $i \neq i'$ and 0 when $h \neq h'$. [3] In the model (41), however, μ_h^* and μ are not identifiable. We can identify the population parameter by imposing the restriction $\sum_{h=1}^{H} \frac{n_h}{n}\mu_h^* = 0$. Using this we get the LS estimators of μ_h and μ by minimizing $\sum_{h=1}^{H}\sum_{i=1}^{n_h}(y_{hi} - \mu - \mu_h^*)^2$. This gives $\bar{y}_h*$ and hence $\bar{y}_h$ as in (36), and the LS estimator of μ

$$\bar{y} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} y_{hi} = \sum_{h=1}^{H} p_h\,\bar{y}_h \tag{42}$$

where $p_h = \frac{n_h}{n}$ are the sampling proportions for each stratum. Note that this is just the estimate of μ we get by assuming that the sample is a simple random sample and pooling all the data, thus ignoring the stratification. We will thus refer to $\bar{y}$ as the pooled estimator, $\bar{y}_p$. We also note that

$$E\,\bar{y} = \sum_{h=1}^{H} p_h\,\mu_h \tag{43}$$

[3] Note here that we are considering the case where a random sample without replacement is drawn from each strata. Since we have strata population sizes N_h we will have finite population effects within each stratum, giving $\rho_h = \frac{-1}{N_h - 1}$.

and

$$V(\bar{y}) = \sum_{h=1}^{H} p_h^2 \, V(\bar{y}_h) \tag{44}$$

Under stratification, it will usually be the case that the inclusion probabilities for each element (7) will no longer be identical. The probability that the *i-th* element in the *h-th* stratum enters the sample will be the number of elements from the *h-th* stratum divided by the population total in the *h-th* stratum, $\pi_{hi} = \pi_h = n_h/N_h$. Thus the LS estimator $\bar{y}$ will be biased unless one of the following two conditions hold

$$\begin{aligned} (i) \quad & \frac{n_h}{N_h} = \frac{n}{N} \\ (ii) \quad & \mu_h = \mu. \end{aligned} \tag{45}$$

The first case is when the combined sample is SRS; this is known as proportional stratified sampling, a special case of stratification in which the data will be self-weighting. Even when stratification is designed to be self-weighting, however, differing response rates in different strata will often result in unequal sampling probabilities. The second case, in which the population is homogenous with respect to means, will result in unbiased estimation of μ using $\bar{y}$, however the variance will still have to be adjusted since the sampling is independently conducted in the different strata. Neither (45),(i) nor (45),(ii) will hold in most surveys.

The problem of bias and inconsistency is potentially serious, as demonstrated in the simulation discussed below in section 2.7. For example, in the case of two strata with slightly dissimilar means (specifically μ_1 =1 and μ_2 =1.5) the bias of $\bar{y}$

will be 25% when the sampling probabilities differ by a factor of 10. The sampling disproportion is much larger than this in many data sets; consider for example, the Survey on Income and Program Participation (SIPP) of the U.S. Census Bureau, where the sampling disproportion is nearly 1000 between some subgroups (SIPP Users Guide, 1991).

The solution to this bias problem is to use the weighted estimator discussed in (23) above. To see that this is the weighted least squares (WLS) estimator, we can minimize $\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}\,(y_{hi}-\mu-\mu_h^*)^2$ with the restrictions that $\sum\limits_{h=1}^{H}\sum\limits_{i=1}^{n_h} w_{hi}\,\mu_h^* = 0$ and $\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi} = 1$. This gives

$$\bar{y}_w = \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}\,y_{hi} \tag{46}$$

and

$$\bar{y}_{h,w} = \sum_{i=1}^{n_h} w_{hi}\,y_{hi} \tag{47}$$

The inflation, or expansion, factors w_{hi} are chosen such that $E\,\bar{y}_w = \mu$ and $E\,\bar{y}_{h,w} = \mu_h$. This gives

$$\begin{aligned} \bar{y}_w &= \sum_{h=1}^{H}\sum_{i=1}^{n_h} \frac{1}{N\pi_{hi}}\,y_{hi}\,, \\ \bar{y}_{h,w} &= \sum_{i=1}^{n_h} \frac{1}{N\pi_{hi}}\,y_{hi} \end{aligned} \tag{48}$$

Thus for stratification we have

$$w_i = \frac{1}{N\pi_i} \tag{49}$$

where π_i is the probability of selection of the *i-th* population unit in the sample. When $\pi_i = n/N$ (i.e. when condition (i) above is met) we get $\bar{y}_w = \bar{y}$ as given in

(42). If we substitute $\pi_{ij} = n_i/N_i$, where this is no longer a constant, we get $\overline{y}_w = \overline{y}_{st}$ and $\overline{y}_{h,w} = \overline{y}_h$.

We can thus see that the weighted least squares is equivalent to transforming the model (4) by

$$W^{1/2}y = W^{1/2}\iota\mu + W^{1/2}u \tag{50}$$

and using the weights from (49) and the method discussed in section 2.3 above. Weights are identical for units in the same stratum.

We can also think of the stochastic weights w_i as being chosen such that the sample is representative of the population in the sense that the sample mean, on average, is identical to the population mean. That is,

$$\begin{aligned} E\overline{y}_w &= E(\sum_{i=1}^{n} w_i\, y_i) = \sum_{i=1}^{n}\sum_{j=1}^{N} w_j\, Y_j\, \pi_j\,(i) \\ &= \sum_{j=1}^{N} w_j\, Y_j \pi_j = \mu. \end{aligned} \tag{51}$$

where $\pi_j\,(i)$ is defined as in (5), the probability that the j-th population unit is selected in the i-th draw. Larger weights are thus given to those elements which are under-represented in the sample. This matches our intuition that over-sampled strata should have less impact on our analysis.

The variance of $\overline{y}_w$ will be the same as in (40). However, if we ignore the fpc, we can use the consistent, weighted estimator of the variance

$$\widehat{Var}(\overline{y}_w) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}\,(y_{hi} - \overline{y})^2\,. \tag{52}$$

This will slightly overestimate the variance, however, since we are not using the information that the between strata variances are zero. Sometimes we only have information on sampling probabilities (weights) and do not know to which stratum each data point belongs. (52) will usually give an adequate approximation of the variance in this case.

If we (mistakenly) assume $\mu_i = \mu$ and pool all the observations (ignoring strata and treating the data as a random sample of size n from a population of size N) then the variance of the pooled estimator is

$$V(\overline{y}) = V(y_P) = (\frac{1}{n} - \frac{1}{N})S^2. \tag{53}$$

This will no longer be appropriate for most surveys.

This is only the same as $V(\overline{y}_{st})$ or $V(y_P)$ if $n_h/n = N_h/N$ and $\mu_h = \mu$. In this case the population is homogenous and the combined sample is a simple random sample. In general, $V(y_P) > V(\overline{y}_{st})$, meaning that ignoring the stratification in the data will cause us to overestimate the variance of the sample mean. This problem will be especially important if within-strata heterogeneity is low and between-strata heterogeneity is high. To see this, consider $V(\overline{y}_{st})$ for case of proportional stratified sampling, where $n_h = nN_h/N$. In this case

$$\frac{V(\overline{y}_{st})}{V(y_P)} = \frac{\sum_{h=1}^{H} \theta_h S_h^2}{\sum_{h=1}^{H} \theta_h(\mu_h - \mu)^2 + \sum_{h=1}^{H} \theta_h S_h^2} \tag{54}$$

where we use $S^2 \simeq N^{-1} \sum_{h=1}^{H} \sum_{i=1}^{N_h} (Y_{hi} - \mu)^2 \simeq \sum_{h=1}^{H} \theta_h S_h^2 + \sum_{h=1}^{H} \theta_h(\mu_h - \mu)^2$. Thus,

if between-strata variation is zero, $V(y_P) = V(\bar{y}_{st})$ But if the between-strata variance is non-zero, $V(\bar{y}_{st})$ is smaller than $V(y_P)$.

One of the advantages to conducting a stratified survey is that this information may be used to design surveys which minimize the variance of the estimator. If something is known about the within-stratum variances (say from previous samples or from similar surveys of related populations), the stratum-specific sample sizes can be chosen to minimize the overall variance of $\bar{y}_w$.

In order to minimize the variance of the weighted estimator, $\bar{y}_{st}$, one must choose that n_h for which $V(\bar{y}_{st})$ is minimum subject to $\sum_{h=1}^{H} n_h = n$. It can be shown using simple calculus that the optimal n_h is

$$n_h^* = \frac{nN_h s_h}{\sum_{h=1}^{H} n_h}. \tag{55}$$

In practice, the unbiased estimator of $V(\bar{y}_{st})$ can be calculated by substituting S_i^2 with s_i^2 in (40). Further, if $\mu_h = \mu$, the unbiased estimate of the variance of $\bar{y}_P$ can be calculated as

$$\hat{V}(\bar{y}_P) = (\frac{1}{n} - \frac{1}{N})[\frac{\sum_{h=1}^{H}(N_h - 1)s_h^2}{N-1}] \tag{56}$$

Alternatively,

$$\hat{V}(\bar{y}_P) = (\frac{1}{n} - \frac{1}{N})s^2 \tag{57}$$

where $s^2 = (n-H)^{-1}\sum_{h=1}^{H}(n_h - 1)s_h^2$ is the pooled estimator if $S_h^2 = S^2$. If $S_h^2 \neq S^2$ but $n_h/n = N_h/N$, one may consider $s_*^2 = (n-1)^{-1}\sum_{h=1}^{H}\sum_{i=1}^{n_h}(y_{hi} - \bar{y})^2 = (n-1)^{-1}[\sum_{h=1}^{H}(n_h - 1)s_h^2 + \sum_{h=1}^{H} n_h \bar{y}_h^2 - (\sum_{h=1}^{H} n_h \bar{y}_h)^2/n]$. However, if $n_h/n \neq N_h/N$ and

$\mu_h \neq \mu$, then

$$\hat{V}(\bar{y}_P) = (\frac{1}{n} - \frac{1}{N})[\frac{\sum_{h=1}^{H}(N_h - 1)s_h^2}{N-1} + q] \tag{58}$$

where q is an unbiased estimator of $(N-1)^{-1}\sum_{h=1}^{H} N_h(\mu_h - \mu)^2$, see Cochran (1953).

From the above analysis it is clear that if the sample observations y_{hi} are generated by stratified random sampling then they should be reweighted to resemble the population by replicating (inflating) sampling units using the inflation or expansion factor, and treating the enlarged sample as if it were the population. The inflation factor, θ_{hi}, for each sampling unit i in the h-*th* stratum is the reciprocal of its sampling probability, that is $\theta_{hi} = \frac{1}{\pi_{hi}} = \frac{N_h}{n_h}$. If we multiply each sample observation by its inflation factor θ_{hi} we obtain an unbiased estimate of the population total. Alternatively if we multiply the sample observations by their weights $w_{hi} = \theta_{hi}/\sum\sum\theta_{hi} = \frac{1}{N\pi_{hi}}$, the normalized inflation factor, we get an unbiased estimate of population mean, as shown in (51) and in section 2.3. Exactly the same procedures can be used to obtain estimates of medians, variances, and other parameters.

Now we turn to another type of sampling frequently found in economic data, cluster sampling. In stratified sampling considered above, elements are selected from all strata. An important difference in cluster sampling is that population elements are divided up into various clusters, sharing common traits, and only some clusters are sampled. This will cause our usual variance estimates to under-estimate, quite substantially, the true variance of the estimated parameters. In other words it affects the variance in the opposite way that stratification does.

2.5 Cluster Sampling

Let us consider Y_{hcj} to be the population observations of the *c-th* group or cluster, $c = 1, ..., C_h$ in the *h-th stratum*, $h = 1, ..., H$. In each cluster, there are M_{hc} elements such that $N_h = \sum_{c=1}^{C} M_{hc}$ is the number of elements in the *h-th* stratum. Cluster sampling or single-stage sampling involves drawing a sample of k_h clusters in each stratum h and then sampling all M_c elements in the chosen clusters. This is also referred to as first-stage clustering. However, if we further take a random sample of m_{hc} elements out of M_{hc} at the second stage then the overall sampling is called sub-sampling or two-stage sampling. In two-stage sampling, the sample inclusion probabilities π_{hci} will depend upon the probability of selection at both stages of the sample.[4]

The first-stage units, (sometimes called primary sampling units), could be villages or street blocks, and the second-stage units could be households. The primary advantage of cluster sampling is that it drastically reduces survey cost per second-stage unit. The disadvantage is that it leads to a higher variance than a SRS of the same size due to the correlation among elements within the same cluster. The standard assumption of uncorrelated observations in cross-sectional data fails dramatically for clustered samples.

We begin by considering the problem of estimating population means μ (or μ_h)

[4] The probability of selection of every element in the chosen cluster is $m_{hc}\, n_h / M_{hc} N_h$.

under single-stage (cluster) sampling. For this we consider the population model as

$$Y_{hci} = \mu_{hc} + U_{hci} \tag{59}$$

where $\sum_c \sum_i U_{hci} = 0$ by definition of μ_{hc}. We generally assume a common mean in each stratum, so that

$$Y_{hci} = \mu_h + U_{hci} \tag{60}$$

We can also think of the model as being an error-components model with cluster specific effects which have expected value zero over the entire stratum. We thus re-write (59) as

$$Y_{hci} = \mu_h + \alpha_c + \epsilon_{hci}. \tag{61}$$

For simplicity, we will consider the case of estimating a stratum-specific mean[5] . For this case, we suppress the h subscript, and consider the population model

$$\begin{aligned} Y_{ci} &= \mu + U_{ci} \\ &= \mu + \alpha_c + \epsilon_{ci}. \end{aligned} \tag{62}$$

The sample model is

$$y_{ci} = \mu + u_{ci}, \qquad c = 1, ..., C, \; i = 1, ..., M_c \tag{63}$$

where we assume RSWR such that

$$\begin{aligned} E\, u_{ci} &= 0, \; E\, u_{ci}^2 = \sigma^2 , \; E\, u_{ci} u_{cj} = \rho\sigma^2 , \; i \neq j \\ E\, u_{ci}\, u_{c'j} &= 0, \; c \neq c'; \end{aligned} \tag{64}$$

[5] This is also appropriate for estimating a population mean when clustering is present without stratification.

$\rho > 0$ is called the intra-cluster correlation coefficient. (64) implies that the elements within clusters are correlated, but are uncorrelated across clusters. The total sample size is $n = \sum_c M_c$. Thus

$$E\,u = 0 \text{ and } E\,u\,u' = \text{Diag. } (\sum_1, \ldots, \sum_C) = \Sigma \tag{65}$$

where Σ is an $n \times n$ block diagonal matrix with $\sum_c = \sigma^2[(1-\rho)I + \rho\,\iota_c\,\iota_c']$. ($\iota_c$ is $M_c \times 1$ vector of ones.) The LS estimator of μ is

$$\bar{y} = \frac{1}{n}\sum_{c=1}^{C}\sum_{i=1}^{M_c} y_{ci} \ . \tag{66}$$

Provided that clusters share a common mean, $\mu_c = \mu$, the estimator $\bar{y}$ is unbiased. Further, its variance is given by

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n^2}\sum_{c=1}^{C}(\iota_c'\sum_c \iota_c) \\ &= \frac{\sigma_u^2}{n}[1 + (\overline{M} - 1)\rho] \end{aligned} \tag{67}$$

where $\overline{M} = n^{-1}\sum_c M_c^2$ is the weighted mean of cluster sizes.

As shown above in section 2.2.1, the LS estimator $\bar{y}$ is the same as the GLS estimator for the structure of $\sum_c$ in (65).

If we ignore the clustering and assume that the sample is drawn as RSWR of size n, we will underestimate the variance of $\bar{y}$ by the factor

$$d = (1 + (\bar{n} - 1)\rho). \tag{68}$$

This is known as the design effect (see Kish (1965)). Failure to account for the intra-cluster correlation will lead to serious underestimation of the true variance in (65) since ρ will generally be positive.

In practice, we use an unbiased estimator of the variance of u

$$\hat{\sigma}_u^2 = \frac{1}{n-d}\,\hat{u}'\,\hat{u} = \frac{1}{n-d}\sum_{c=1}^{C}\sum_{i=1}^{M_c}\hat{u}_{ci}^2 \tag{69}$$

or the consistent estimator $\hat{u}'\,\hat{u}/n$. Further, we estimate ρ by

$$\hat{\rho} = \frac{\sum_{c=1}^{C}\sum_{i=1}^{M_c}\sum_{j\neq j}^{M_c}\hat{u}_{cj}\,\hat{u}_{cj'}}{\hat{\sigma}_u^2 n(\overline{M}-1)} \tag{70}$$

and use these in (67) to get an estimate of the variance of $\overline{y}$. Failure to inflate the usual RSWR estimate of the variance of $\overline{y}$, $\frac{s^2}{n}$, will result in underestimation of the true variance.

Deaton (1997) provides numerical examples of the effect of ignoring ρ in the calculation of standard errors of $\overline{y}$. He shows that for estimated food price elasticities in Pakistani Villages, ρ is between .3 and .6, leading to underestimation of $V(\overline{y})$ by a factor greater than 2 for mean cluster size is 12. Design effects between 2 and 3.5 are very common in economic data, meaning standard errors for $\overline{y}$ calculated using $\widehat{V(\overline{y})} = \frac{s^2}{n}$ will be up to three-and-a-half times too large!

Kish (1965) provides extensive discussion of cluster sampling, including the case of two- and three-stage cluster sampling and the appropriate variance formulas.

2.6 Systematic Sampling

Systematic sampling is one of the most common techniques used in development economics. In systematic sampling, the sampling units are (usually) arranged in random order with respect to the variable of interest.[6] Of the first K units, one is selected at random. Then every Kth unit is sampled in order. This sampling design is the easiest to implement, because it involves drawing only one sample. Systematic sampling can be thought of as a kind of one-stage cluster sampling. The population is arranged into K clusters, each with n elements. One of these clusters is chosen and every element within that cluster is sampled. For simplicity, we assume below that there are N elements in the population and that N/K is an integer. (In other words, the N elements are exhaustively and uniquely assigned to the K clusters, each of which has the same number of elements, n—an integer.)

Consider the finite population model

$$Y_i = \mu + U_i \tag{71}$$

as in (1) where the data are randomly ordered. The population mean and variance are defined as in section 2.1. If the population is divided into K clusters, we can write the model as.

$$Y_{kj} = \mu_k + U_{kj}\,,\ k = 1, ..., K \text{ and } j = 1, ..., N. \tag{72}$$

[6] Imagine, for example, hospital records by alphabetical order. If one wanted to survey hospital patients, the sample could be drawn by choosing a random number between 1 and 10 and then taking every 10th person after that. Presumably any medical conditions of interest covered in the survey would be uncorrelated with a person's last name. This method would be quite easy to teach someone totally unskilled in survey methods.

Our sample would consist of one randomly chosen cluster k, written as:

$$y_j = \mu + u_j, \quad j = 1, ..., n. \tag{73}$$

The LS estimator of the mean, when cluster k is chosen, is

$$\bar{y}_k = \bar{y}_{\text{SYS}} = \frac{1}{n}\sum_{j=1}^{n} y_j, \tag{74}$$

which will be unbiased when N/k is an integer. Further

$$V(\bar{y}_{\text{SYS}}) = \frac{1}{K}\sum_{k=1}^{K}(\bar{y}_k - \mu)^2. \tag{75}$$

In general, we will not be able to estimate this variance. In the case where our data consists of one systematic sample, the population mean, μ, is unknown as are the remaining $(K-1)$ unsampled clusters. In some surveys, re-sampling is possible. In this case, information can be gathered about the within and across-cluster heterogeneity and an approximation for $V(\bar{y}_{\text{SYS}})$ as a function of the intracluster correlation coefficient, ρ:

$$V(\bar{y}_{\text{SYS}}) = \frac{\sigma^2}{n}\left[1 + \rho\,(n-1)\right]. \tag{76}$$

In general, if the data are randomly arranged with respect to the variable of interest, systematic sampling should give broad coverage of the population, the estimate of the mean should be unbiased, and the approximation of $V(\bar{y}_{\text{SYS}})$ by assuming simple random sampling will not be too unreasonable. If the data are clustered, and the clusters ordered, then a systematic sample will perform better than either SRS or clustered sampling. This follows because the systematic sample, picking every *kth*

elements, will cover most, if not all, clusters. This very broad coverage will give a precise estimate of the mean.

There remains much which has been written and much which can be said about systematic sampling. It is perhaps the most widely used method in statistics, because it is so easy to implement and so easy to teach to someone who knows nothing of statistics. For more information on systematic sampling, see the chapter in Thompson (1992) or the surveys by Iachan (1982) and Bellhouse (1988b).

2.7 Simulation: mean model

In this section, we present a summary of the results from a detailed simulation of the mean model under complex sampling. The two primary objectives of the simulation are: 1) to illustrate the effect of ignoring sample design in data analysis; and 2) to ascertain the properties of our estimators under various sample designs where there do not exist analytical results.

The first step in our simulation was the creation of several finite "populations." The populations created ranged in size from 50 to 20,000, with means ranging from 1 to 2000. They were drawn from an "infinite" population of randomly distributed, normal numbers. From these finite "populations" we then drew n observations using the sampling design in question. For the questions under consideration in this section, the shape of the distribution is irrelevant, so only normal random numbers were

considered. For analyzing other variables, such as the distribution of $V(s^2)$ considered in Ullah and Breunig (1996), the shape of the distribution does matter and conclusions based on simulations using only normally distributed populations should be made with caution.

To demonstrate the first point, a sample of size n was drawn using the sample design of interest (stratified, clustered, etc.), then $\overline{y}$ and $V(\overline{y})$ were estimated using the information on how the sample was drawn. Then, taking this same sample, and ignoring the sample design, we have calculated $\overline{y}_{\text{RSWR}}$ and $V(\overline{y}_{\text{RSWR}})$–i.e., treating the sample as if it were a random sample drawn with replacement. These values are averaged over 1000 repetitions. We then compare the average Bias $(\overline{y})$ and Bias $(\overline{y}_{\text{RSWR}})$ and the ratio of the averages of the two variances, which can be interpreted as the degree of over- or under-estimation arising from ignoring (or mis-specifying) the sample design.

The second type of simulation we have undertaken, to answer the second question raised above, involves drawing a separate sample for each sampling design of interest. From these distinct samples, we calculate $\overline{y}$ and $V(\overline{y})$ for each of the sample designs. After r repetitions, we compute the average bias and the simulation variance of the estimator for each design.

By way of example, let us consider a simulation of both kinds comparing sampling

under finite population and infinite population from section 2.2. Recall that

$$V(\overline{y}_{\text{SRS}}) = \frac{s^2}{n}d = V(\overline{y}_{\text{RSWR}})\left(1 - \frac{n}{N}\right). \tag{77}$$

Following the first method outlined above, we take 1000 samples of size n from our population and calculate $V(\overline{y}_{\text{SRS}})$ using the fact that the sample has been drawn without replacement from a finite population. Then we will calculate $V(\overline{y}_{\text{RSWR}})$ and compare the ratio of the average of these two over the 1000 repetitions.

The ratio should be the inverse of the finite population correction. Indeed this is confirmed in the results in Table 2.1. Results from the second type of simulation are presented in Table 2.2. Here, two separate samples are drawn from the same population–one under SRS, the other under sampling with replacement. Results for 10,000 repetitions are reported. As expected, the results closely approximate those in Table 1. One way to interpret these results is that for the same sample size, sampling without replacement is more precise than sampling with replacement. (Since the variance of the estimator $\overline{y}$ under SRS, is on average, smaller.) We can also think of the ratio as representing the "cost" of assuming that sampling is from an infinite population when in fact sampling is from a finite population.

Tables 2.3 through 2.7b present comparisons between RSWR and stratified sampling without replacement, conducted under the first method described above. In Table 2.3, we consider two strata, each with a population of 1000, where a sample of size n_i is drawn from each stratum. Since $n_1 \neq n_2$ and both strata have a population

of 1000, the sampling probabilities are unequal. As we saw in section 2.3 above, the unweighted estimator of μ will be biased. As we can see from Table 2.3, the more unequal the sampling probabilities, the greater the bias in the unweighted estimator, $\bar{y}_{RSWR}$, and the greater the ratio of mean squared errors. Most data in labor and development economics is stratified and the most common case is unequal sampling probabilities, either by design or because of different rates of non-response across strata. Thus, as the simulation shows, a potentially serious bias problem exists even in calculating a simple mean. The general intuition behind these results extends to the regression case.

In Table 2.4 we consider the case of "spurious" stratification. A simgle population of 200 is arbitrarily divided up into two strata. Then equal probability samples are drawn from the two strata. Since we have equal probability sampling and the two strata are identical, both condition (i) and condition (ii) of section 2.4 are met, so there is no bias problem. And as we can see from columns 5 and 6, there is a small gain in variance and mean squared error. This is due to the fact that we are using sampling without replacement (which has lower variance) in both strata, and that by drawing a stratified sample we have zero variance for elements in different strata. From Table 2.4, we thus see that even when strata are identical, there can be some gain in variance by selecting a stratified sample provided the samples are chosen independently in the different strata.

In Tables 2.5 to 2.7, we present results from the stratified case, but with equal

probabilities of selection in both strata. In Table 2.5, we see that even though $\bar{y}_{\text{RSWR}}$ (unweighted) is unbiased under equal sampling probability, it is not efficient compared to $\bar{y}_{w.}$ In Tables 2.6 and 2.7, we consider across-strata heterogeneity in the mean and the element variances separately, since both affect the ratio between $\bar{y}_{\text{RSWR}}$ and $\bar{y}_{w\cdot}$. In Table 2.6 we first consider the case where $\sigma_1^2 = \sigma_2^2$, but $\mu_1 \neq \mu_2$. Table 2.7 presents the case where $\sigma_1^2 \neq \sigma_2^2$, but $\mu_1 = \mu_2$. We note that the increase in precision as measured by the ratio of variances is increasing as the distance between the two strata means, μ_1 and μ_2, increases. It is not uncommon in development economics to encounter stratified samples where the urban mean income is three times that of rural mean income. In the case where $\mu_1 = 200$ and $\mu_2 = 600$, we see that this can lead to an over-estimate of the variance of the population mean by a factor of 20. It increases for large sample sizes, because the finite population correction has a proportionally larger effect. From Table 2.7, we see that stratification does not improve efficiency when the strata have the same mean regardless of the difference in within-stratum variance. The small increases in efficiency that we see are the result of the increasing effect of the finite population correction as the sample size increases. In other simulations, not reported here, we show that assuming a stratified structure when the data does not have one, leads to no gains in efficiency. This result follows intuitively from the results presented in Table 2.7.

Table 2.8a and Table 2.8b present the cost of ignoring the one-stage clustered sample design and assuming that the sample is actually a RSWR. Here we have drawn

a clustered sample from one stratum with a population of 1000, which is divided into 50 clusters, each of size 10. We present the ratio of variances: $\frac{V(\bar{y}_c)}{V(\bar{y}_{\text{RSWR}})}$, where we have calculated $V(\bar{y}_c)$ using the estimated sample value of $\hat{\rho}$. We compare this to the expected Kish Design effect $(1 + (m-1)\rho)$ given our knowledge of the true value of ρ. $\hat{\rho}$ gives a slight under-estimation, which disappears as $n \rightarrow N$.

Tables 2.9 and 2.10 present results comparing RSWOR (SRS), stratified sampling, cluster sampling, and systematic sampling. We use the second method described above for simulation.

Table 2.9 presents results from 5000 replications comparing SRS, clustered, and systematic samples drawn from the same population. The last three columns compare the variances of the different estimators. Systematic sampling performs best in the simulation reported in Table 2.9. Since the data are ordered by cluster, the systematic sample gives the broadest coverage of the population–taking at least one observation from each cluster. (See section 2.5) As expected, clustered sampling gives the highest variance for given sample size. Table 2.10 compares stratified and systematic sampling in a stratified population. Results are for 10,000 repetitions. The systematic sample does not perform as well as the stratified random sample since stratification will more evenly cover the population over repeated sampling.

Table 2.1
Effect of Ignoring Sample Design (Sampling with Replacement)

Population size	Sample size: 5	10	20	50	100
50	1.11	1.25	1.67	x	x
100	1.05	1.11	1.25	2.00	x
500	1.01	1.02	1.04	1.11	1.25
1000	1.00	1.01	1.02	1.05	1.11

Table entries show the degree of over-estimation of the variance of the sample mean: $\frac{\text{var}(\bar{y}_{RSWR})}{\text{var}(\bar{y}_{SRS})}$.

Table 2.2
Efficiency Gains from Sampling Without Replacement vs. Sampling with Replacement

Population size	Sample size 5	10	20	50	100
50	1.08	1.22	1.63	x	x
100	1.063	1.082	1.24	1.98	x
500	.993	1.009	1.054	1.12	1.253
1000	1.00	1.00	1.005	1.055	1.108

Entries in table are $\frac{\text{var}(\bar{y}_{RSWR})}{\text{var}(\bar{y}_{SRS})}$ averaged over 10,000 repetitions of each design.

Table 2.3
Stratified Sampling with Unequal Probabilities

2 Strata: $\mu_1 = 200$; $\sigma_1 = 60$;
$\mu_2 = 300$; $\sigma_2 = 75$;

	Penalty of not considering sampling structure			
Sample size (Stratum 1, 2)	Bias $\bar{y}_w$	Bias $\bar{y}_{RSWR}$	Ratio of Variances $Var(\bar{y}_{RSWR}) / Var(\bar{y}_w)$	Ratio of MSEs $MSE(\bar{y}_{RSWR}) / MSE(\bar{y}_w)$
(5,10)	0.94	17.97	1.46	1.81
(5,20)	-0.13	30.11	1.15	2.99
(5,50)	0.30	41.62	0.60	4.75
(10,5)	0.24	-16.62	1.23	1.37
(20,5)	1.16	-29.68	0.72	1.97
(50,5)	-0.99	-41.37	0.27	3.07

Population mean: $\mu = 250$

N=1000 for both strata

$$\frac{n_1}{N_1} = \frac{n_2}{N_2}$$

Table 2.4
Stratified Sampling with Equal Probabilities:
"Spurious" Stratification

Sample sizes (Strata 1, Strata 2)	Population Mean	Population Variance	Bias	Ratio of $Var(\bar{y})$ / $Var(\bar{y}_{st})$	Ratio of MSE
(5, 5)	1	1.00	N	1.10	1.05
	2	1.34	E	1.09	1.05
	10	1.79	G	1.10	1.05
	140	216.62	L	1.10	1.05
	1215	3524.67	I	1.10	1.05
			G		
(10, 10)	1	1.00	I	1.24	1.12
	2	1.34	B	1.27	1.14
	10	1.79	L	1.24	1.12
	140	216.62	E	1.24	1.12
	1215	3524.67		1.24	1.12
(20, 20)	1	1.00		1.72	1.36
	2	1.34		1.66	1.32
	10	1.79		1.67	1.34
	140	216.62		1.65	1.32
	1215	3524.67		1.68	1.34
(30, 30)	1	1.00		2.48	1.74
	2	1.34		2.48	1.77
	10	1.79		2.57	1.78
	140	216.62		2.48	1.74
	1215	3524.67		2.52	1.77

Population: Randomly divided into two strata with sub-population means $\mu_1 = \mu_2 = \mu$

Population of both strata=100
Total Population: 200

Table 2.5a
Stratified Sampling with Equal Probabilities:
Improved Efficiency for Unequal Strata Means

$$\sigma_1 = \sigma_2 = 50$$

		Penalty of not considering sampling structure		
Sample sizes (Stratas 1, 2)	**Population means (Strata 1, Strata 2)**	Bias $\bar{y}$ = Bias $\bar{y}_{st}$	Ratio of Variances $Var(\bar{y}) / Var(\bar{y}_{st})$	Ratio of MSEs
(5, 5)	**(200, 300)**	8.6	2.20	1.56
	(200, 400)	13.8	5.39	3.21
	(200, 500)	5.6	11.64	6.12
	(200, 600)	5.0	20.22	10.44
	(200, 800)	4.5	43.15	22.86
	(200, 1000)	2.73	74.95	37.82
	(200, 1500)	-11.4	187.33	92.98
(10, 10)	**(200, 300)**	7.2	2.17	1.60
	(200, 400)	-10.0	5.31	3.18
	(200, 500)	-3.0	10.80	6.06
	(200, 600)	24.3	18.94	9.68
	(200, 800)	2.5	41.51	21.79
	(200, 1000)	-10.5	69.11	35.16
	(200, 1500)	-2.1	183.68	93.7
(20, 20)	**(200, 300)**	-16.0	2.16	1.54
	(200, 400)	-14.9	5.22	3.12
	(200, 500)	4.9	10.86	5.89
	(200, 600)	6.8	18.44	9.67
	(200, 800)	-1.7	40.89	22.03
	(200, 1000)	0.2	68.06	34.86
	(200, 1500)	1.03	174.16	86.20

Stratified sampling without replacement in each stratum.
2 Strata: with sub-population means μ_1, μ_2 given above.

Total Population: N=2000
Population equally divided between two strata
Population mean: $\frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 = \mu$

Table 2.5b
Stratified Sampling with Equal Probabilities:
Improved Efficiency for Unequal Strata Means

$$\sigma_1 = \sigma_2 = 50$$

		Penalty of not considering sampling structure		
Sample sizes (Stratas 1, 2)	**Population means (Strata 1, Strata 2)**	Bias $\bar{y}$ = Bias $\bar{y}_{st}$	Ratio of Variances $Var(\bar{y}) / Var(\bar{y}_{st})$	Ratio of MSEs
(50, 50)	**(200, 300)**	-2.2	2.20	1.61
	(200, 400)	-8.5	5.29	3.09
	(200, 500)	-5.2	10.93	6.11
	(200, 600)	-10.2	18.97	9.76
	(200, 800)	1.2	41.88	21.54
	(200, 1000)	0.3	69.58	34.51
	(200, 1500)	-3.58	179.50	90.10
(100, 100)	**(200, 300)**	4.1	2.32	1.66
	(200, 400)	-1.3	5.54	3.28
	(200, 500)	3.7	11.50	6.20
	(200, 600)	-3.8	19.81	10.05
	(200, 800)	-3.7	43.77	22.27
	(200, 1000)	0.6	72.75	35.55
	(200, 1500)	3.3	189.39	94.76
		-0.21	5.08	

Stratified sampling without replacement in each stratum.
2 Strata: with sub-population means μ_1, μ_2 given above.

Total Population: N=2000
Population equally divided between two strata
Population mean: $\frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 = \mu$

Table 2.6
Stratified Sampling with Equal Probabilities:
Improved Efficiency for Unequal Strata Variances

$$\mu_1 = \mu_2 = 1000$$

		Penalty of not considering sampling structure		
Sample sizes (Stratas 1, 2)	**Population σ^2 (Strata 1, Strata 2)**	Bias $\bar{y}$ = Bias $\bar{y}_{st}$	Ratio of Variances $Var(\bar{y}_{srs}) / Var(\bar{y}_{st})$	Ratio of MSEs
(5, 5)	**(2500, 3520)**	-12.3	1.01	1.00
	(2500, 4909)	29.6	1.01	1.00
	(2500, 6527)	-13.7	0.99	0.99
	(2500, 8089)	32.1	1.00	1.00
	(2500, 9651)	13.4	1.01	1.00
	(2500, 41696)	-12.9	1.01	1.01
(10, 10)	**(2500, 3520)**	2.5	1.01	1.01
	(2500, 4909)	9.2	1.01	1.00
	(2500, 6527)	2.1	1.01	1.01
	(2500, 8089)	22.6	1.01	1.01
	(2500, 9651)	14.3	1.01	1.01
	(2500, 41696)	13.2	1.01	1.01
(20, 20)	**(2500, 3520)**	25.3	1.02	1.01
	(2500, 4909)	18.7	1.02	1.01
	(2500, 6527)	20.9	1.02	1.01
	(2500, 8089)	21.9	1.02	1.01
	(2500, 9651)	-5.0	1.02	1.01
	(2500, 41696)	-6.5	1.02	1.01
(50, 50)	**(2500, 3520)**	-5.5	1.05	1.03
	(2500, 4909)	3.5	1.05	1.03
	(2500, 6527)	2.1	1.05	1.03
	(2500, 8089)	18.2	1.05	1.03
	(2500, 9651)	-6.9	1.05	1.03
	(2500, 41696)	-4.9	1.05	1.03

Stratified sampling without replacement in each stratum.
2 Strata: with sub-population variances σ_1, σ_2 given above.
Total Population: N=2000
Population equally divided between two strata

Table 2.7a
Stratified Sampling with Equal Probabilities:
Improved Effeciency for Unequal Strata Means and Variances

		Penalty of not considering sampling structure		
Sample sizes (Stratum 1, 2)	**Population means (Stratum 1, 2)**	Bias $\bar{y}$ = Bias $\bar{y}_{st}$	Ratio of Variances $Var(\bar{y}) / Var(\bar{y}_{st})$	Ratio of MSEs
(5, 5)	**(200, 300)**	0.33	1.59	1.30
	(200, 400)	0.29	2.29	1.64
	(200, 600)	0.12	3.53	2.24
	(200, 800)	2.22	5.46	3.19
	(200, 1000)	0.17	5.10	2.96
	(200, 1500)	-.81	8.88	5.05
(10, 10)	**(200, 300)**	-0.61	1.58	1.29
	(200, 400)	0.09	2.18	1.60
	(200, 600)	0.29	3.51	2.24
	(200, 800)	-0.19	5.41	3.20
	(200, 1000)	-3.41	4.79	2.85
	(200, 1500)	0.56	8.33	4.80
(20, 20)	**(200, 300)**	0.07	1.59	1.30
	(200, 400)	-0.38	2.17	1.58
	(200, 600)	0.22	3.42	2.22
	(200, 800)	-1.67	5.28	3.13
	(200, 1000)	0.41	4.76	2.92
	(200, 1500)	-0.40	8.42	4.81

Ratios in last two columns show improved efficiency from using stratified mean estimator when both strata means and variances exhibit heterogeneity between strata.

Stratified sampling without replacement in each stratum.
2 Strata: with sub-population means μ_1, μ_2 given above.

Total Population: N=2000
Population equally divided between two strata

Population mean: $\frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 = \mu$

Table 2.7b
Stratified Sampling with Equal Probabilities:
Improved Effeciency for Unequal Strata Means and Variances

		Penalty of not considering sampling structure		
Sample sizes (Stratum 1, 2)	**Population means (Stratum 1, 2)**	Bias $\bar{y}$ = Bias $\bar{y}_{st}$	Ratio of Variances $Var(\bar{y}) / Var(\bar{y}_{st})$	Ratio of MSEs
(50, 50)	**(200, 300)**	0.19	1.65	1.33
	(200, 400)	-0.75	2.22	1.63
	(200, 600)	0.06	3.49	2.25
	(200, 800)	0.07	5.42	3.26
	(200, 1000)	0.71	4.88	2.82
	(200, 1500)	-0.66	8.45	4.67
(100, 100)	**(200, 300)**	-0.20	1.74	1.36
	(200, 400)	0.15	2.37	1.67
	(200, 600)	0.05	3.68	2.35
	(200, 800)	0.26	5.97	3.32
	(200, 1000)	-0.21	5.08	3.07
	(200, 1500)	-0.40	8.91	4.81

Ratios in last two columns show improved efficiency from using stratified mean estimator when both strata means and variances exhibit heterogeneity between strata.

Stratified sampling without replacement in each stratum.
2 Strata: with sub-population means μ_1, μ_2 given above.

Total Population: N=2000
Population equally divided between two strata

Population mean: $\frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 = \mu$

Table 2.8a
One-stage Cluster Sampling Without Replacement:
Effect of Changing Values of ρ

No. of clusters sampled (c)	Total sample size (n=c*m)	True Pop. Mean	Pop ρ	$\hat{\rho}$	Ratio of Variances $Var(\bar{y}_{clu}) / Var(\bar{y})$	Expected Kish Design effect (d)
5	100	1000	.12	.088	3.13	3.28
			.17	.13	5.00	4.23
			.26	.20	6.48	5.94
			.44	.344	7.24	9.36
			.48	.38	10.02	10.12
			.64	.519	14.44	13.16
10	200	1000	.12	.11	2.39	3.28
			.17	.15	3.89	4.23
			.26	.24	5.78	5.94
			.44	.41	10.36	9.36
			.48	.45	11.78	10.12
			.64	.59	12.85	13.16

Column 6 is "design effect" in simulation.
Column 7 gives the expected design effect given the estimated value of r.

Sampling with equal probabilities

1 Stratum: Total population of stratum=1000
50 Clusters: 20 elements in each cluster (balanced clusters)

$$Var(\bar{y}_{clu}) = \left(\frac{C-c}{C}\right)\frac{s_u^2}{c} \quad \text{where} \quad s_u^2 = \frac{1}{c-1}\sum_{j=1}^{c}\left(\bar{y}_c - \bar{y}\right)^2$$

$$\text{or} \quad = \frac{s^2}{n}d \quad \text{where d is the Kish (1965) design effect } 1+(m-1)\rho$$

Table 2.8b
One-stage Cluster Sampling Without Replacement:
Effect of Changing Values of ρ

No. of clusters sampled (c)	Total sample size (n=c*m)	True Pop. Mean	Pop ρ	$\hat{\rho}$	Ratio of Variances $Var(\bar{y}_{clu}) / Var(\bar{y})$	Expected Kish Design effect (d)
20	400	1000	.12	.117	3.25	3.28
			.17	.169	4.22	4.23
			.26	.257	5.93	5.94
			.44	.434	9.47	9.36
			.48	.47	10.28	10.12
			.64	.62	13.49	13.16
25	500	1000	.12	.119	3.69	3.28
			.17	.169	4.29	4.23
			.26	.256	5.42	5.94
			.44	.439	9.69	9.36
			.48	.48	10.09	10.12
			.64	.63	13.64	13.16

Column 6 is "design effect" in simulation.
Column 7 gives the expected design effect given the estimated value of r.

Sampling with equal probabilities

1 Stratum: Total population of stratum=1000
50 Clusters: 20 elements in each cluster (balanced clusters)

$$Var(\bar{y}_{clu}) = \left(\frac{C-c}{C}\right)\frac{s_u^2}{c} \quad \text{where} \quad s_u^2 = \frac{1}{c-1}\sum_{j=1}^{c}\left(\bar{y}_c - \bar{y}\right)^2$$

$$\text{or} \quad = \frac{s^2}{n}d \quad \text{where d is the Kish (1965) design effect } 1+(m-1)\rho$$

Table 2.9
Comparison of SRS, Clustered, and Systematic Sampling Designs

# of clusters sampled	Total sample size	True Population Mean	Correla-tion coefficient	Ratio of Variances: Clustered/ SRS	Ratio of Variances: SRS/ Systematic	Ratio of Variances: Clustered/ Systematic
5	100	1007.4	.12	.28	1.31	4.60
		986.8	.17	.21	1.65	8.04
		983.9	.26	.15	1.88	12.4
		1004.9	.44	.10	1.62	15.74
		997.7	.48	.09	2.30	25.74
		986.9	.64	.06	2.92	47.37
10	200	1007.4	.12	.24	1.09	4.52
		986.8	.17	.19	1.69	8.98
		983.9	.26	.11	1.69	14.7
		1004.9	.44	.08	3.35	39.46
		997.7	.48	.08	6.24	81.15
		986.9	.64	.07	1.04	14.78
20	400	1007.4	.12	.18		
		986.8	.17	.15		
		983.9	.26	.11	infeasible	
		1004.9	.44	.07		
		997.7	.48	.06		
		986.9	.64	.05		

50 Clusters: Population = 20 in each cluster (balanced clusters)
Variable correlation coefficient

Table 2.10
Systematic Sampling Compared with
Stratified Systematic Sampling

Sample size	Strata means	Strata Pop.	Bias $\bar{y}_{st}$	Bias $\bar{y}_{sys}$	Ratio of Variances $Var(\bar{y}_{sys}) / Var(\bar{y}_{st})$	Ratio of MSEs
5	(10, 100) (2, 900) (1000, 1000)	50 100 500	0.01 3.21 0.09	-0.00 -0.09 -0.54	0.45 1.06 1.05	0.73 1.03 1.02
10	(10, 100) (2, 900) (1000, 1000)	50 100 500	-0.02 -4.67 1.28	0.00 0.28 1.94	0.62 0.62 1.36	0.81 0.81 1.18
25	(10, 100) (2, 900) (1000, 1000)	50 100 500	0.01 4.11 -0.17	-0.00 -3.11 -0.04	0.24 2.55 1.46	0.62 1.78 1.23
50	(2, 900) (1000, 1000)	100 500	1.89 -0.16	0.30 0.77	1.89 1.74	1.44 1.37
100	(1000, 1000)	500	-0.13	0.14	1.03	1.02

$\bar{y}_{st} = \sum_{h=1}^{2} \frac{N_h}{N} \bar{y}_h$ sample of size n drawn without replacement

$\binom{N}{n}$ Possible samples

$\bar{y}_{sys} = \frac{1}{(n)} \sum_{h=1}^{2} \sum_{i=1}^{n_i} y_{hi}$ sample of size 1 drawn, then every N/n-th unit sampled

$\frac{N}{n}$ Possible samples

2.8 Conclusion

As seen in this chapter, problems of inference and estimation arise when data are gathered under a complex sampling design. The simulation demonstrates that these are of more than trivial interest. Unequal sampling probabilities are the rule, not the exception, and treating such data as having been drawn under RSWR will lead to biased estimators. Even where the disproportion is 2 to 1, this leads to large bias as shown in Table 2.3.

As different strata will usually have different parameter means, ignoring stratification will lead to large overestimates of the true variance of our estimate of μ. A recent survey of income in Kenya showed that average rural income was one-third that of average urban income. Ignoring stratification when calculating a population mean in this case will lead to confidence intervals which are 20 times too wide. The exact opposite problem occurs in clustering. Intra-class correlation coefficients of .5 are common in developing country studies. The simulation shows that ignoring the sample design leads to an underestimate of the variance by a factor of 10–or more if the average cluster size is greater than 20.

Bias problems and mis-estimates of standard errors are exacerbated in more complex sample designs which combine different aspects of stratification, clustering, and systematic sampling. And clearly the same problems will arise in the regression case.

The simulation demonstrates that assuming away sample design effects as trivial is unjustified. Instead, more careful attention should be paid to using available methods of analysis and information on sampling to construct unbiased, and more precise, estimates.

I would conclude this section by providing further proof of both the importance of the issues discussed above and the frequent mis-understandings created by survey sampling design. Consider the following quote from a Federal Trade Commission report on proposed United States Department of Agriculture regulations concerning the contents labeling of milk:

> "Furthermore, sampling from the inspection lot, rather than some larger production lot, would not cause the statistical finding to be biased against the manufacturer. Of course, it is possible that an inspection lot is significantly more underfilled, on average, than some larger set of packages of which it is part. But it should be just as likely that an inspection lot is significantly more overfilled, on average, than the large production lot. Data from the milk study support the conclusion that inspection of packages in retail stores will not result in any bias against the dairy industry."
>
> (Federal Trade Commission, 1997)

The authors of the report correctly claim that bias will not be a problem here. However, they completely ignore the problems of standard error estimation from the

clustered sampling which is described here. Even worse, it turns out that the actual regulations (and proposed fines) will be based upon the number of standard errors away from the labelled quantity the actual quantity is found to be. But if that standard error is being estimated from a clustered sample, it may be underestimated by 300% or more as we saw in section 2.5!

This failure to deal with the clustering in the data and its effect on the estimated standard errors will have an important policy impact. Furthermore, this type of ignorance about the relevance of the survey sampling literature is rampant in economics. Economists and econometricians need to pay more attention to the nature of the survey design when analyzing data.

3 INEQUALITY MEASUREMENT FROM SURVEY SAMPLES

Measuring inequality and poverty and making comparisons between regions, between countries, and across different time periods are important activities for governments, aid organizations, and economists. Program and policy goals, both in developing and in over-developed nations, often include poverty or inequality reduction. Testing claims of policy success or failure and achieving a quantitative understanding of poverty and inequality are thus important objectives. Problems that arise include the difficulty of measurement and data collection and basic philosophical questions about what constitutes inequality. Index number measures of inequality are frequently employed because they give a simple aggregate of a complex income distribution, allowing easy comparison across different states of time and different regions. Many such indices have been proposed. Most inequality indices are ratios of random variables, as is shown below. One problem that arises is that these measures will be biased in small samples. In section 3.3, we show that this bias is potentially quite large.

We can think of most inequality indices as being straightforward extensions of the mean. But rather than summing values of y_i we are calculating functions of y_i. As with the mean model, if we calculate inequality indices ignoring the sample structure of the data, our estimators may be biased. This will be a problem when

we face a sample gathered through unequal probability sampling, whether stratified or not. Another problem in the inequality and applied literature is the frequent lack of inference when inequality measurement is conducted. Inequality indices are often reported, as are their evolution over time. Confidence intervals, however, are rarely given. Without inference, the statistical significance of changes in these indices is unknown. The lack of reporting of confidence intervals for inequality measures partly stems from the difficulty of implementing the results on the distribution of inequality indices and partly from an ignorance that such indices are random variables which have their own distributions. We will discuss some of the literature where people have attempted to create confidence intervals for inequality measures. We will then show how these results change when data are stratified and clustered.

3.1 Inequality measurement using indices: some issues

We define an inequality index as a function which maps from the space of income distributions, Θ, to the real line,

$$I : \Theta \rightarrow \Re$$

such that as inequality (however defined) increases, the value of I will increase. Clearly, different definitions of inequality will lead to different definitions of the function I. Two approaches have generally been taken in the literature: 1) identifying a social welfare function that incorporates our judgments on inequality and transforming this welfare function into an index number; or 2) identifying a list of properties

to which we expect our inequality index to adhere and then choosing an index based upon the particular set of axioms that we find most desirable[7] .

Here, the social welfare function (SWF) is not to be seen as a function that is being maximized by any agency or polity. Instead, it is an aggregator which turns a distribution of income into a number representing the judgment of that distribution from a particular perspective.

Dalton (1920) first expressed the connection between social welfare and the measurement of inequality. Atkinson (1970) developed formally the concept of looking at inequality measurement from a social welfare perspective. Atkinson's index (see (83) below) corresponds to a SWF which exhibits constant relative inequality aversion. (The parameter ϵ captures the degree of inequality aversion.) Sen (1992), among others, discusses this approach in some depth.

Another approach to inequality measurement is to specify axioms that an inequality index should satisfy. After specifying a number of desirable axioms, one then goes about finding an index which will correspond to the desired axioms. Frequently proposed axioms include:

- Symmetry (Anonymity): If any two individuals in the economy trade places in the income distribution, $I(\cdot)$ should remain unchanged. (In other words, only the list of incomes matters, not whose name is alongside which income.)

[7] There is a third approach which seeks to develop direct measures of well being, such as the Human Development Index, and which lead to inequality comparisons across different functionings/capabilities. See Sen (1992).

- Scale Independence: If $y_i = \alpha x_i$ for every $i = 1, \ldots, N$, then $I(f(y)) = I(g(x))$ (i.e., our index of inequality will be the same if income is measured in rupees or in dollars.)

- Proportionality: If a proportionate number of persons are added at all income levels, then $I(\cdot)$ should remain unchanged.

- Principle of transfers: Define a progressive mean-preserving transfer (MPT) as one in which income is transferred from a rich person to a poor person, but not so much income that the previously rich person becomes more poor than the poor person originally was. Then the principle of transfers states that if $g(y)$ can be created from $f(y)$ by a progressive MPT, $I(g(y)) < I(f(y))$.

This last axiom is sometimes called the Pigou-Dalton principle of transfers.

Other possible axioms include limiting the range of the index to an interval (such as $[0, 1]$), or requiring the index to be decomposable by income groups or types. We may also be interested in other types of transfer properties. We may not care much about slight increases in inequality at higher parts of the distribution if they are compensated by relatively large decreases in inequality in the lower part of the distribution. Some authors have attempted to require various restrictions on transfer sensitivity at different levels of income. For examples, see Kakwani (1980) and Davies, Green and Paarsch (1998).

Blackorby and Donaldson (1978) bridge the gap between these two approaches by

showing how any inequality index can be transformed to reveal its underlying, implicit social welfare function. They also demonstrate how different inequality indexes weight different income transfers unequally. Thus the choice of inequality measure implicitly reflects a judgment on the overall picture of inequality as well as a judgement about the welfare implications of transfers at different levels of income.

The advantage to both of these approaches is that once a particular index has been chosen (or derived from the preferred SWF) the index will exhaustively rank all possible income distributions. Another advantage to the above two approaches is that they force an explicit (or implicit, but recoverable) definition of inequality. The obvious disadvantage of the above approaches is that disagreement about the axioms or about the structure of the SWF will lead to different indices which may contradict each other for certain income distributions.

Having touched on a few issues, I will not attempt to survey the vast literature on measuring inequality, constructing inequality indices, the theory of inequality measurement and inequality indices. For the economist, Kakwani's (1980) book is probably the most useful. The most commonly used indices are treated in depth, and interesting examples are provided which are quite illustrative. Cowell (1995) is also useful. For a non-technical approach, I suggest Coulter (1989). This book was written from a non-mathematical veiwpoint and its intended audience is any social scientist who is interested in looking at distribution. With few equations, but many examples, it manages to cover the theory and measurement of inequality quite nicely.

3.2 Inequality indices

There is a considerable amount of controversy in the simple question: what is income? We will abstract from that problem and issues of whether inequality is best considered over income or some potentially smoother (and potentially more welfare-consistent) measure such as consumption. Below, I will use the word income as a general term which may be taken to mean consumption or income including imputed income from household production, etc. The important point is that the measure of "income" should be a reasonably good measure of welfare.

We will use the model from (1) above. Consider a finite population of size N and the variable income, Y, allowing us to write the population model as

$$Y_i = \mu + U_i \ , \ i = 1, ..., N \tag{78}$$

where Y_i is income of the *i-th* population unit (which may be a household or an individual), U_i is the *i-th* error, and μ and σ^2 are the population mean and variance, respectively, given by

$$\mu = \frac{1}{N} \sum_{i=1}^{N} Y_i \ , \ \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2. \tag{79}$$

U_i, sum to zero by the definition of μ and are thus not products of some data generating process, but rather deviations from the common population mean as it is defined above.

We can define the following inequality measures for the population:

the coefficient of variation (CV)

$$CV = \frac{\sigma}{\mu}; \tag{80}$$

Theil's two measures of inequality (I_1 and I_0)

$$I_1 = \mu^{-1}\frac{1}{N}\sum_{i=1}^{N} Y_i \ln(Y_i) - \ln(\mu), \tag{81}$$

and

$$I_0 = \ln(\mu) - \frac{1}{N}\sum_{i=1}^{N} \ln(Y_i); \tag{82}$$

and Atkinson's measure

$$A(\epsilon) = 1 - \mu^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}}. \tag{83}$$

When $\epsilon = 1$ in Atkinson's measure, it takes the form

$$A(1) = 1 - \mu^{-1}e^{\frac{1}{N}\sum_{i=1}^{N}\ln(Y_i)}. \tag{84}$$

All of these measures meet the four axioms suggested for inequality indices above. We can think of these as being minimum criteria for a reasonable inequality index. I_0 and I_1 were first introduced by Theil (1967). They are based upon entropy and information theory from engineering. The CV has been around for a long time as a measure of inequality and spread, whose property of scale independence has long been recognized[8] . Atkinson (1970) introduced $A(\epsilon)$, deriving it directly from a social welfare function that depends positively on total income and on the equally distributed

[8] CV is discussed further in section 3.3 below.

equivalent level of income captured by the choice of ϵ and the function

$$\left(\frac{1}{N}\sum_{i=1}^{N} Y_i^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}}. \tag{85}$$

Larger values of ϵ imply higher aversion to inequality. If $\epsilon = 0$, inequality will always be zero. This is a measure which is completely insensitive to how the income is distributed in society. As $\epsilon \to \infty$, Atkinson's measure becomes Rawlsian–it is only concerned with the level of income of the poorest person in society. (See Rawls (1972).) Choice of ϵ is not without controversy. Atkinson, in his own work, used Kuznets (1963) data and demonstrated how different choices of ϵ resulted in considerably different rankings between countries in that sample. Whether there exists some criteria upon which to base an appropriate choice of ϵ remains to be seen.

Below we will discuss estimation and inference for inequality indices using the above measures as examples.

3.2.1 Inequality indices: Estimation

Given some sample of data

$$y_i = \mu + u_i \ , \ i = 1, ..., n \tag{86}$$

we can estimate the above inequality measures by replacing population values with their sample estimators:

the coefficient of variation (CV)

$$\widehat{CV} = \frac{s}{\bar{y}}; \tag{87}$$

Theil's two measures of inequality (I_1 and I_0)

$$\widehat{I_1} = \overline{y}^{-1}\frac{1}{n}\sum_{i=1}^{n} y_i \ln(y_i) - \ln(\overline{y}), \tag{88}$$

and

$$\widehat{I_0} = \ln(\overline{y}) - \frac{1}{n}\sum_{i=1}^{n} \ln(y_i); \tag{89}$$

and Atkinson's measure

$$\widehat{A(\epsilon)} = 1 - \overline{y}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} y_i^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}}. \tag{90}$$

When $\epsilon = 1$ in Atkinson's measure, it takes the form

$$\widehat{A(1)} = 1 - \overline{y}^{-1} e^{\frac{1}{n}\sum_{i=1}^{n}\ln(y_i)} \tag{91}$$

$$= 1 - e^{-\widehat{I_0}}. \tag{92}$$

The estimators in (87) to (92) will be consistent, but as they are ratios of random variables, they will not be unbiased. To see this, we note that all of the above indices (as well as many other commonly used income inequality measures) can be written as a function of the mean and some other function, $g(y)$, of income

$$I(y) = F(g(y), \mu). \tag{93}$$

Our sample estimate of $I(y)$ will then be

$$\widehat{I}(y) = F(\widehat{g}(y), \overline{y}) \tag{94}$$

where $\widehat{g}(y)$ will take the general form $\widehat{g}(y) = n^{-1}\sum_{i=1}^{n} g(y_i)$.

From Cramer (1946), the expected value of $\hat{I}$ will be of the form

$$E\hat{I} = F_0 + O\left(\frac{1}{n}\right) \tag{95}$$

where F_0 is the function F evaluated at $Eg(y)$ and μ. Consistency thus follows directly by noting that $\lim_{n\to\infty} \hat{I} = I$. As we will see below in section 3.3, these terms of $O\left(\frac{1}{n}\right)$ may be significant in non-normal populations, even for values of n which we usually think of as being "large."

We will use (93) and (94) below to derive estimates of standard errors for inequality measures. They will also be useful for writing inequality measures in a manner which allows us to adjust for the bias in (95) when samples are small.

3.2.2 Inequality indices: Inference

One of the most important reasons to calculate inequality indices is to use them as a policy analysis tool to compare across regions, across countries, and within the same country/region over time. Indeed, an inequality number in and of itself is almost useless. The important questions always tend to be of the nature, "has inequality decreased over time?" or "Are the policies in country A more conducive to equality than those of country B?". It is almost scandalous, therefore, how little attention is paid to inference and the construction of confidence intervals for inequality measures. How can such comparative questions be answered without some sense of how much change in a measure is significant change?

There has in fact been a substantial amount of theoretical work on inference for inequality measures. But a quick perusal of much of the applied literature shows that it has not been heeded. Nygård and Sandström (1981) develop asymptotic standard errors for several dozen inequality measures. Kakwani (1990) develops standard errors using a similar method for the Gini coefficient, the entire class of decomposable measures, and the Theil's index. Confidence intervals for CV were published by Cramer (1946). Gastwirth (1974) developed asymptotic standard errors for the relative mean deviation. Sandström (1983) devoted an entire monograph to inference for the Gini coefficient and other measures related to the Lorenz curve. Nygård and Sandström (1985) considered the Gini coefficient and Theil's measures. Other work on the Gini coefficient includes that of Gastwirth and Gail (1985) and Gastwirth, et. al. (1986). Cowell (1995) gives what he calls "rule of thumb" estimates of standard errors, but these are based on underlying distributions (such as normal) which are unlikely to hold for income distributions. Cowell, however, does not provide any theoretical nor empirical justification for why these might be good "rules of thumb." Nygård and Sandström (1981) cite a paper by Nygård (1981)[9] , in which violations of non-normality in the Finnish income distribution are shown to be quite great and result in serious mis-estimation of standard errors derived under the assumption of a normal distribution of income.

[9] This manuscript is unpublished and apparently not translated from the Finnish.

Here we follow the approach of Cramer (1946). This has also been called the δ method by Rao (1973). Given that we can write the inequality measures above in the form (93), we can make use of the following result from Cramer (1946, p. 353)

Lemma 1 *Given a function $G(h_1(x), h_2(x))$ which is continuous and has continuous first and second derivatives and meets certain other regularity conditions.*

$$Var(G) = Var(h_1(x)) \left(\frac{\partial G}{\partial h_1(x)} \right)^2 + 2Cov(h_1, h_2) \frac{\partial G}{\partial h_1(x)} \frac{\partial G}{\partial h_2(x)} + Var(h_2(x)) \left(\frac{\partial G}{\partial h_2(x)} \right)^2 + O\left(\frac{1}{n^{3/2}} \right).$$

The partial derivatives, $\frac{\partial G}{\partial h_1(x)}$ and $\frac{\partial G}{\partial h_2(x)}$, must be evaluated at the points $E(h_1(x))$ and $E(h_2(x))$. In order to implement this result for income inequality measures, the regularity conditions imply that all incomes must be positive. If we have any negative incomes in our sample, we must therefore normalize so that all incomes are greater than zero. I have found in simulation that the standard errors for $A(\epsilon = 1)$ and Theil's two measures, all of which use logs of income, may behave strangely when there are incomes less than 1 in the sample. Some normalization to avoid this problem is therefore suggested.

Cramer also demonstrates that under fairly general conditions the function $G(\cdot)$ above will be asymptotically normally distributed. We can thus make use of these two results to calculate asymptotic variances for our inequality measures, conduct inference and build confidence intervals.

Using (93) and Lemma 1, we have

$$Var(\widehat{I}(y)) = Var(\widehat{g}(y))\left(\frac{\partial I(y)}{\partial \widehat{g}(y)}\right)^2 + 2Cov(\widehat{g}(y),\overline{y})\frac{\partial I(y)}{\partial \widehat{g}(y)}\frac{\partial I(y)}{\partial \overline{y}} + Var(\overline{y})\left(\frac{\partial I(y)}{\partial \overline{y}}\right)^2 + O\left(\frac{1}{n^{3/2}}\right). \quad (96)$$

In order to implement this result, we need to re-write the inequality measure of interest in the form $\widehat{I}(y) = F(\widehat{g}(y),\overline{y})$. As an example, consider the coefficient of variation. The population coefficient of variation is

$$CV = \frac{\sigma}{\mu}.$$

The sample estimate is

$$\widehat{CV} = \frac{s}{\overline{y}} = \left(\frac{s^2}{\overline{y}^2}\right)^{\frac{1}{2}} = \left(\frac{\widehat{\beta}_2 - \overline{y}^2}{\overline{y}^2}\right)^{\frac{1}{2}} \quad (97)$$

where $\widehat{\beta}_2 = \frac{1}{n}\sum y_i^2$ is the sample, non-central second moment. (This is our $\widehat{g}(y)$ in (96).)

For the CV,

$$\frac{\partial I(y)}{\partial g(y)} = \frac{1}{2\sigma\mu} \quad (98)$$

$$\frac{\partial I(y)}{\partial \overline{y}} = -\frac{\beta_2}{\sigma\mu^2} = -\frac{1}{\sigma} - \frac{\sigma}{\mu^2} \quad (99)$$

As we saw above, in RSWR

$$Var\,(\overline{y}) = \frac{\sigma^2}{n}. \quad (100)$$

The $Var(\widehat{g}(y))$ can be found by writing

$$\begin{aligned} Var(\widehat{\beta}_2) &= E\left[\widehat{\beta}_2 - E\widehat{\beta}_2\right]^2 \\ &= E\left[\frac{1}{n}\sum y_i^2 - \mu^2 - \sigma^2\right]^2 \end{aligned} \tag{101}$$

Under RSWR, this provides

$$\begin{aligned} Var(\widehat{\beta}_2) &= \frac{1}{n}\left\{(\gamma_2+3)\,\sigma^4 + 4\mu^2\sigma^2 + 4\mu\sigma^3\gamma_1 - \sigma^4\right\} \\ &= \frac{1}{n}\left\{(\gamma_2+2)\,\sigma^4 + 4\mu^2\sigma^2 + 4\mu\sigma^3\gamma_1\right\}. \end{aligned} \tag{102}$$

Likewise, the $Cov(\widehat{\beta}_2, \overline{y})$ can be shown to equal

$$Cov(\widehat{\beta}_2, \overline{y}) = \frac{1}{n}\left[\sigma^3\gamma_1 + 2\mu\sigma^2\right]. \tag{103}$$

The derivations of $Var(\widehat{\beta}_2)$ and $Cov(\widehat{\beta}_2, \overline{y})$ are given in Appendix B.

Combining (97) through (103) yields, upto $O(\frac{1}{n})$

$$Var(\widehat{CV}) = \frac{1}{n}\left\{(\gamma_2+2)\,\frac{\sigma^2}{4\mu^2} - \frac{\sigma^3\gamma_1}{\mu^3} + \frac{\sigma^4}{\mu^4}\right\}. \tag{104}$$

We can also write this in terms of the coefficient of variation itself

$$Var(\widehat{CV}) = \frac{(CV)^2}{n}\left\{\frac{\gamma_2+2}{4} - (CV)\,\gamma_1 + (CV)^2\right\}. \tag{105}$$

If we impose the assumption of normality of the distribution of income, then $\gamma_1 = \gamma_2 = 0$, and we have

$$Var(\widehat{CV}) = \frac{(CV)^2}{4n}\left[2 + 4\,(CV)^2\right] \tag{106}$$

which matches Cowell (1995, p. 118). However, since γ_1 and γ_2 are easily computed with any standard software, and since income distributions are non-normal (and frequently highly so), it doesn't appear sensible to use this approximation.

Cramer provides another equivalent version of the variance of the coefficient of variation

$$Var(\widehat{CV}) = \frac{\mu^2(\beta_4 - \beta_2^2) - 4\mu\beta_2\beta_3 + 4\beta_2^3}{4n\mu^4\beta_2}. \tag{107}$$

β_c is the cth non-central moment of Y. In practice we can consistently estimate the variance by replacing μ with $\overline{y}$ and β_c with $\widehat{\beta_c} = \frac{1}{n}\sum y_i^c$ in (107) or by replacing CV with $\widehat{CV}$ in (105).

The same procedure that we followed for the coefficient of variation can be employed to derive the variances of I_0, I_1, and $A(\epsilon)$. Below we provide Kakwani's (1990) results for the variance of the inequality measures we are considering.

$$Var(\widehat{I}_0) = \frac{\sigma^2}{\mu^2} + E(t_2) - (t_0)^2 - \frac{2}{\mu}(t_1 - \mu t_0) \tag{108}$$

$$Var(\widehat{I}_1) = \frac{1}{\mu^2}\left[E(t_3) - (t_1)^2\right] + \frac{(t_1+\mu)^2\sigma^2}{\mu^4} - \frac{2(t_1+\mu)}{\mu^3}(t_4 - \mu t_1) \tag{109}$$

$$\begin{aligned} Var(\widehat{A}(\epsilon)) \;=\; & \frac{1}{\mu^2(1-\epsilon)^2}(\alpha_{1-\epsilon})^{\frac{2\epsilon}{1-\epsilon}}\left(\alpha_{2-2\epsilon} - \alpha_{1-\epsilon}^2\right) \\ & + \frac{\sigma^2}{\mu^4}(\alpha_{1-\epsilon})^{\frac{2}{1-\epsilon}} - \frac{2(\alpha_{1-\epsilon})^{\frac{1+\epsilon}{1-\epsilon}}}{\mu^3(1-\epsilon)}(\alpha_{2-\epsilon} - \mu\alpha_{1-\epsilon}) \end{aligned} \tag{110}$$

where

$$t_0 = \frac{1}{N}\sum_{i=1}^{N} \ln(Y_i) \tag{111}$$

$$t_1 = \frac{1}{N}\sum_{i=1}^{N} Y_i \ln(Y_i) \tag{112}$$

$$t_2 = \frac{1}{N}\sum_{i=1}^{N} (\ln(Y_i))^2 \tag{113}$$

$$t_3 = \frac{1}{N}\sum_{i=1}^{N} (Y_i \ln(Y_i))^2 \tag{114}$$

$$t_4 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 \ln(Y_i) \tag{115}$$

and

$$\alpha_{1-\epsilon} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{1-\epsilon} \tag{116}$$

$$\alpha_{2-\epsilon} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{2-\epsilon} \tag{117}$$

etc. In practice, we can replace these with their sample analogs and get consistent estimators for the variance[10] .

For Atkinson's measure, when $\epsilon = 1$, we can use (92) to write the variance of $\hat{A}(1)$ in terms of $\hat{I}_0$. In general to find the variance of a function of the form $f(g(x))$, we

[10]When $\epsilon \neq 1$ the expression in (110) is not valid for the variance of Atkinson's measure–see below.

write the Taylor's series expansion

$$f(g(x) = f(g_0(x)) + f'(g_0(x))\,(g(x) - g_0(x)) + ... \tag{118}$$

where $(g_0(x)) = E\left[(g(x))\right]$. Then,

$$f(g(x) - f(g_0(x)) = f'(g_0(x))\,(g(x) - g_0(x)) + ... \tag{119}$$

and

$$E\left[f(g(x) - f(g_0(x))\right]^2 = \left(f'(g_0(x))\right)^2 E\left[g(x) - g_0(x)\right]^2 + ... \tag{120}$$

$$Var(f(g(x)) = \left(f'(g_0(x))\right)^2 Var(g(x)) \tag{121}$$

In this case, $Var(g(x)) = Var(\hat{I}_0)$ and $\left(f'(g_0(x))\right)^2 = e^{-2\hat{I}_0}$. So when $\epsilon = 1$

$$Var(\hat{A}(1)) = Var(\hat{I}_0)e^{-2\hat{I}_0}. \tag{122}$$

Once we have calculated the estimated variance of the inequality measure, θ, we can exploit Cramer's result (Cramer, 1946, p. 367) that $\sqrt{n}\left(\hat{\theta} - \theta\right)$ is asymptotically normal to build confidence intervals around our estimate, $\hat{\theta}$. An asymptotic 95% confidence interval may be built as

$$\hat{\theta} \pm 1.96\sqrt{Var(\theta)}. \tag{123}$$

For comparison between two inequality estimates from different countries or in different years, the asymptotically normal test statistic

$$\frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\frac{Var(\hat{\theta}_1)}{n_1} + \frac{Var(\hat{\theta}_2)}{n_2}}} \tag{124}$$

will have zero mean and variance one and can thus be used for hypothesis testing.

Without providing the details, we note here that the same approach can be used for poverty measurement. For example, for the Foster-Greer-Thorbecke family of measures,

$$P(y_i) = \left(\frac{q_i}{z}\right)^{\alpha} \tag{125}$$

where z is the poverty line and q_i is the poverty gap for a household or individual below the poverty line. Assuming an exogenously given poverty line, we can write this as a function of $\bar{y}$ and $f(y_i) = q_i = y_i - z$ and following the method above. See Kakwani (1980) for an excellent review of inequality/poverty measurement. Ravallion (1994) and Sen (1992) also provide excellent reviews. These measures were first introduced by Foster, et. al. (1984).

3.2.3 Simulation

The asymptotically normal standard errors presented above may give poor approximations for highly non-normal data or for small samples. We conducted a simulation study of the effects of sample size on these approximations for both normal and non-normal populations. For the normal data, the simulation was conducted by drawing random numbers from a normal distribution with mean 100 and variance 20. This leads to a CV of .2 for the population, which as seen in Table 3.1, is estimated with only small bias which diminishes as the sample size grows. The data was truncated at y=2. Any values of y less than 2 were made equal to 2. Since this is almost 5

standard deviations from the mean, such occurrences were quite rare. This truncation was taken to avoid the problem mentioned above when incomes are less than one.

For the log-normal data, numbers were generated based upon a normal distribution following the method of Evans, Hastings and Peacock (1993). Given a standard normal N(0,1) variable,

$$L = me^{\sigma N(0,1)} \tag{126}$$

will be lognormally distributed with median m and variance of $logL$, σ^2. We chose to keep the population value of CV identical as in the normal case. The standard deviation of the variable L is $m\sqrt{e^{\sigma^2}(e^{\sigma^2}-1)}$ and its mean will be $me^{\frac{1}{2}\sigma^2}$. Setting the mean to 100 and the standard deviation to 20 gives two equations with two unknowns and solving for σ and m gives $\sigma = .198$ and $m = 98.058$. Plugging these values into (126) allows generation of log-normal random variables.

Tables 3.1 through 3.5 summarize the results from this simulation. Column six on each table provides a ratio of the estimated variance to the simulated variance over 1000 repetitions. The first thing to notice is that the variance (both true and estimated) diminishes as sample sizes increases. The problem of mis-estimation for small samples varies greatly with the inequality measure considered. Also, the approximations for the variance used above may give under or over-estimation. For a sample of size 10, the estimated variance for CV underestimates the true variance by about 40%. For Theil's first measure, $I(1)$, the estimated variance is an overestimate of the true variance by a factor of 140! For sample sizes of 5000, the asymptotic

approximation of variance performs well for all of the measures considered. Notice that for Theil's first measure, however, the variance is over-estimated by 250% even for sample of size 1000. Kakwani (1990), for example, uses these approximations to conduct inference for samples of around 300, so his results must be considered with caution.

The results for the lognormal populations are quite similar to the normal case. However, as we expect, the approximations are worse for small samples than in the normal case. As sample size grows to 5000 however, the normal approximation works quite well for all of the measures considered. The degree of skewness considered here $\gamma_1 = (e^{\sigma^2}+2)(e^{\sigma^2}-1)^{\frac{1}{2}} = .608$ is quite small.[11] The degree of under/over-estimation of the variance will increase as the coefficient of skewness increases.

In the next section, we discuss the problem of unbiased estimation for inequality measures in small samples. In section 3.4, below, we will consider first the case of random sampling without replacement and compare it to random sampling with replacement considered above. In section 3.5, we discuss estimation and inference in the presence of stratification. In section 3.6, we discuss the effects on inference when clustering is present in the data. In section 3.7, we give an example where both stratification and clustering are present and demonstrate how our analysis changes when we take into account the sample structure of the data. In section 3.8, we conclude this portion on inequality measurement.

[11]See the example from Kenya in section 3.3 where the sample estimate of the coefficient of skewness is 20.74.

Table 3.1
Simulation Results Comparing Asymptotically Normal Variance to True (Simulated) Variance

Coefficient of Variation

Normal distribution

Sample size	CV	Var(CV)	$Var_{sim}(CV)$	Var(CV)/ $Var_{sim}(CV)$
10	0.19654	0.001555	0.002375	0.654607
20	0.197626	0.000881	0.001143	0.770922
50	0.198907	0.000395	0.000451	0.876503
100	0.19989	0.000208	0.000223	0.932879
500	0.200331	4.35E-05	4.36E-05	0.998416
1000	0.199873	2.15E-05	2.09E-05	1.025054
2000	0.199882	1.08E-05	1.08E-05	1.001189
5000	0.199952	4.32E-06	4.03E-06	1.072036

Log-normal distribution

Sample size	CV	Var(CV)	$Var_{sim}(CV)$	Var(CV)/ $Var_{sim}(CV)$
10	0.192288	0.001244	0.00265	0.469363
20	0.196894	0.000862	0.001159	0.74414
50	0.199529	0.000408	0.000443	0.922061
100	0.199634	0.000219	0.000241	0.910056
500	0.200253	4.65E-05	5.07E-05	0.916056
1000	0.199792	2.31E-05	2.28E-05	1.014029
2000	0.200173	1.17E-05	1.30E-05	0.900932
5000	0.200002	4.66E-06	4.89E-06	0.953541

Table 3.2
Simulation Results Comparing Asymptotically Normal Variance to True (Simulated) Variance

Theil's Measure: I(0)

Normal distribution

Sample size	I(0)	Var(I(0))	$Var_{sim}(I(0))$	Var(I(0))/ $Var_{sim}(I(0))$
10	0.019833	0.000522	0.000131	4.000432
20	0.020328	0.000152	5.90E-05	2.585689
50	0.020904	3.85E-05	2.43E-05	1.586897
100	0.021257	1.60E-05	1.25E-05	1.279772
500	0.021491	2.73E-06	2.61E-06	1.045818
1000	0.021355	1.28E-06	1.17E-06	1.094934
2000	0.021388	6.36E-07	6.15E-07	1.033893
5000	0.021401	2.52E-07	2.29E-07	1.103084

Log-normal distribution

Sample size	I(0)	Var(I(0))	$Var_{sim}(I(0))$	Var(I(0))/ $Var_{sim}(I(0))$
10	0.017687	0.000452	8.29E-05	5.45218
20	0.018684	0.000134	3.83E-05	3.498581
50	0.019384	3.09E-05	1.45E-05	2.136173
100	0.019494	1.17E-05	8.09E-06	1.445888
500	0.019632	1.72E-06	1.66E-06	1.037388
1000	0.019569	8.12E-07	7.65E-07	1.061461
2000	0.019632	4.02E-07	4.31E-07	0.933526
5000	0.019608	1.57E-07	1.65E-07	0.953766

Table 3.3
Simulation Results Comparing Asymptotically Normal Variance to True (Simulated) Variance

Theil's Measure: I(1)

Normal distribution

Sample size	I(1)	Var(I(1))	$Var_{sim}(I(1))$	$Var(I(1))/Var_{sim}(I(1))$
10	0.018889	0.01303	9.38E-05	138.9854
20	0.019473	0.003217	4.64E-05	69.30649
50	0.020006	0.000524	1.92E-05	27.35546
100	0.020312	0.000137	9.78E-06	13.95553
500	0.020498	7.04E-06	1.96E-06	3.582963
1000	0.020395	2.22E-06	9.11E-07	2.434338
2000	0.020413	7.96E-07	4.75E-07	1.676699
5000	0.020428	2.43E-07	1.78E-07	1.367596

Log-normal distribution

Sample size	I(1)	Var(I(1))	$Var_{sim}(I(1))$	$Var(I(1))/Var_{sim}(I(1))$
10	0.0176	0.012629	8.45E-05	149.4856
20	0.018645	0.003195	3.90E-05	81.91792
50	0.019365	0.000525	1.51E-05	34.79963
100	0.019477	0.000135	8.37E-06	16.10118
500	0.019639	6.71E-06	1.74E-06	3.846431
1000	0.019566	2.07E-06	7.91E-07	2.610638
2000	0.019636	7.25E-07	4.51E-07	1.606414
5000	0.019608	2.13E-07	1.71E-07	1.245437

Table 3.4
Simulation Results Comparing Asymptotically Normal Variance to True (Simulated) Variance

Atkinson's Measure: A(1)

Normal distribution

Sample size	A(1)	Var(A(1))	$Var_{sim}(A(1))$	Var(A(1))/ $Var_{sim}(A(1))$
10	0.019574	0.000493	0.000123	4.018762
20	0.020094	0.000145	5.62E-05	2.582206
50	0.020675	3.68E-05	2.32E-05	1.584168
100	0.021026	1.53E-05	1.19E-05	1.278816
500	0.021261	2.61E-06	2.50E-06	1.045599
1000	0.021128	1.23E-06	1.12E-06	1.09471
2000	0.021161	6.09E-07	5.89E-07	1.033639
5000	0.021174	2.41E-07	2.19E-07	1.10308

Log-normal distribution

Sample size	A(1)	Var(A(1))	$Var_{sim}(A(1))$	Var(A(1))/ $Var_{sim}(A(1))$
10	0.017491	0.000431	7.92E-05	5.448522
20	0.018492	0.000128	3.67E-05	3.4974
50	0.019191	2.97E-05	1.39E-05	2.134069
100	0.019301	1.12E-05	7.77E-06	1.445415
500	0.01944	1.65E-06	1.59E-06	1.037366
1000	0.019378	7.81E-07	7.36E-07	1.06146
2000	0.019441	3.86E-07	4.14E-07	0.933558
5000	0.019417	1.51E-07	1.59E-07	0.953736

Table 3.5
Simulation Results Comparing Asymptotically Normal Variance to True (Simulated) Variance

Atkinson's Measure: A(2)

Normal distribution

Sample size	A(2)	Var(A(2))	$Var_{sim}(A(2))$	Var(A(2))/ $Var_{sim}(A(2))$
10	0.040496	0.000812	0.000679	1.194787
20	0.041493	0.000323	0.000289	1.119406
50	0.04287	0.000124	0.000123	1.013632
100	0.043726	6.43E-05	6.41E-05	1.003015
500	0.044442	1.99E-05	2.15E-05	0.926103
1000	0.044083	8.86E-06	8.63E-06	1.02662
2000	0.044162	3.92E-06	3.84E-06	1.021862
5000	0.044191	1.65E-06	1.48E-06	1.115514

Log-normal distribution

Sample size	A(2)	Var(A(2))	$Var_{sim}(A(2))$	Var(A(2))/ $Var_{sim}(A(2))$
10	0.034619	0.000559	0.000297	1.884942
20	0.03662	0.000216	0.000141	1.52681
50	0.038004	6.83E-05	5.26E-05	1.300129
100	0.038247	3.19E-05	2.96E-05	1.076452
500	0.038485	5.82E-06	5.99E-06	0.97228
1000	0.038383	2.86E-06	2.81E-06	1.0161
2000	0.038497	1.44E-06	1.55E-06	0.925087
5000	0.038454	5.70E-07	6.05E-07	0.942828

3.3 Small-sample bias in inequality measures: Coefficient of Variation

3.3.1 Introduction

The sample coefficient of variation (cv) is extensively used in applied sciences for various reasons. In economics, cv has been considered to compare income inequality across regions or groups, see Sen (1992) and Ravallion (1994). Recently Beach, Davidson, and Slotsve (1994) have used cv to develop a test for third-order stochastic dominance. (Although their test normalizes income in a way that can lead to contradictory dominance results.)

Despite the widespread use of the cv, not much is known about its sampling properties, especially under non-normality. Sampling of $\widehat{cv}$ under normality has been well-analyed going back to McKay (1932). Warren (1982) provides an approximation of the sampling distribution of cv under normality and surveys the literature on the coefficient of variation under normality. More recently, Gupta and Ma (1996) analyze $\widehat{cv}$ in k-variate normal populations. Notable exceptions where the properties of $\widehat{cv}$ have been studied under non-normality include Singh (1973) who used simulation methods to study the sample $\widehat{cv}$ under a variety of distributions, Bowman and Shenton (1981), who have considered the mean and variance of cv under Weibull distribution and Neuts (1982), who has considered bounds on cv under mixtures of distributions.

The modest aim of this section is to study the approximate bias and mean squared error (MSE) properties of $(cv)^2$ for the general case, that is without imposing any re-

strictions on the form of the distribution of the economic variable under consideration. I have chosen to consider the coefficient of variation because it allows for a relatively simple derivation of its small-sample properties and provides a clear-exposition of the large-n approximation. Although it is not attempted here, the method can be applied to other, more frequently used, inequality measures as well[12] .

The properties of $(cv)^2$, instead of cv, have been considered because comparing $(cv)^2$ across regions or groups is similar to comparing cv and also because of the simplicity in deriving the results; the derivation of sampling properties of cv requires expanding both the numerator and denominator whereas $(cv)^2$ requires the expansion of the denominator only.

The approximate bias and MSE considered in this paper are derived from the large–n asymptotic expansion which is straightforward to obtain and provides simple expressions. Note that the exact results will be extremely difficult to obtain without imposing some specification of the distribution. The results show that the sample cv is biased– that is it under/over-estimates the true cv–and a neat analytical condition under which the bias will be positive or negative is provided. Such an analytical condition, as far as the author is aware, is not available in the literature and it indicates that the sample cv will generally be an underestimate of the true cv for positively-skewed distributions. Since income distributions are generally positively-

[12]Maasoumi and Theil (1979) consider effect of skewness and kurtosis on the population value of Theil's measures of inequality. However, they do not consider the moments of the sample estimates of these inequality measures.

skewed, the author feels that income inequality estimates based on cv are under-reported, especially for small samples. Since income inequality estimates have policy implications, I propose a bias-adjusted estimator of cv which is almost-unbiased up to the order of approximation considered. The plan of this portion of the paper is as follows. In the next section, I present the main results. The results presented in section 3.3.2 are derived in the Appendix A. In section 3.3.3, an application of the result to Chinese and Kenyan data are considered.

3.3.2 Main Results

Let us consider the population mean model as

$$y_i = \mu + u_i \quad , \quad i = 1, ..., n \tag{127}$$

where y_i is the i-th observation on the study variable, μ and σ^2 are the unknown population mean and variance, respectively, and u_i is an unobserved error variable. We assume that the elements u_i are independently and identically distributed such that, for i=1,... ,n,

$$\begin{aligned} Eu_i &= 0, \; Eu_i^2 = \sigma^2, \; Eu_i^3 = \gamma_1\sigma^3 \quad Eu_i^4 = (\gamma_2 + 3)\sigma^4 \\ Eu_i^5 &= (\gamma_3 + 10\,\gamma_1)\sigma^5, \; Eu_i^6 = (\gamma_4 + 10\,\gamma_1^2 + 15\gamma_2 + 15)\sigma^6 \end{aligned} \tag{128}$$

where γ_1 and γ_2 are Pearson's measure of skewness and kurtosis of the distribution. Likewise the quantities γ_3 and γ_4 can also be regarded as measures of deviation

from normality. See Kendall and Stuart (1977, p. 72) for expression of (128) in terms of cumulants. For normal distribution, $\gamma_1, \gamma_2, \gamma_3$, and γ_4 are zero while for a symmetric, non-normal distribution only γ_1 and γ_3 are zero. Thus non-zero values of γ_1 to γ_4 indicate a departure from normality. We also assume that $\mu \neq 0$.

The square of the population coefficient of variation is given by

$$\theta = \frac{\sigma^2}{\mu^2} \tag{129}$$

Further, the square of the sample coefficient of variation is expressed as

$$\hat{\theta} = \frac{s^2}{\bar{y}^2} \tag{130}$$

where $\bar{y}$ and s^2 are the sample mean and variance, respectively, and these are written as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ , \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{131}$$

In what follows, using (131) and (130), I present large-n approximations for the bias and the mean squared error (MSE) of $\hat{\theta}$. I also note that exact analytical results under non-normal distributions are difficult to obtain, and even if one is able to obtain them, they will depend on the specific distribution of u. The results here are for any non-normal distribution having finite moments of at least six.

Proposition 3.1: If the errors follow (128), the bias of $\hat{\theta}$, upto $O(n^{-1})$, and the MSE of $\hat{\theta}$, upto $O(n^{-2})$, are respectively given as:

$$Bias(\hat{\theta}) = \frac{\theta^{3/2}}{n}\left[3\theta^{1/2} - 2\gamma_1\right] \tag{132}$$

and

$$\begin{aligned} \text{MSE}\left(\hat{\theta}\right) &= \frac{\theta^2}{n}\left[\gamma_2 + 4\theta - 4\gamma_1\theta^{1/2} + 2\frac{n}{n-1}\right] \\ &+ \frac{\theta^2}{n^2}\left[\theta(24\gamma_2\frac{n}{n-1} + 20\gamma_1^2\left(\frac{n+1}{n-1}\right) + 20\frac{n}{n-1}) \right. \\ &\left. -4\theta^{1/2}(\gamma_3 + 4\gamma_1\tfrac{n}{n-1}) + 75\theta^2 - 108\theta^{3/2}\gamma_1\right] \end{aligned} \tag{133}$$

The results are derived in the Appendix A. I observe that the approximate results in the above proposition are for normal ($\gamma_1 = \gamma_2 = 0$) as well as for non-normal errors.

From (132) it is clear that up to the order of approximation considered, *Bias* $(\hat{\theta})$ is positive for negatively-skewed distributions ($\gamma_1 < 0$), and it is negative for positively-skewed ($\gamma_1 > 0$) distributions provided

$$\gamma_1 > \frac{3}{2}\,(cv) \tag{134}$$

where $cv = \theta^{1/2}$. That is $\hat{\theta}$ provides an over-estimation of θ for negatively- skewed distribution, and an underestimation for positively-skewed distributions provided (134) holds. When $\gamma_1 = \frac{3}{2}cv$, the bias in $\hat{\theta}$ vanishes. Since income distributions are usually positively skewed, the use of $\hat{\theta}$ to measure income inequality will underestimate the extent of income equality. Furthermore, since this property of $\hat{\theta}$ will also hold true for

cv, use of the coefficient of variation will lead to underestimation of income inequality for positively-skewed income distributions.

For example, when u is lognormally distributed, $\gamma_1 = (e^{\sigma^2} - 1)^{1/2}(e^{\sigma^2} + 2)$ and $cv = (e^{\sigma^2} - 1)^{1/2}$, Kendall and Stuart (1977, p. 181). Thus (134) is satisfied and hence $\hat{\theta}$ will underestimate θ. Similarly for the exponential distribution $\gamma_1 = 2$ and $cv = 1$ and hence $\gamma_1 > \frac{3}{2}(cv)$.

Since the estimates $\hat{\theta}$ may be used in policy evaluations/prescriptions it is useful to develop a bias-corrected estimator. This is given by

$$\tilde{\theta} = \hat{\theta} - \widehat{Bias}(\hat{\theta}) \tag{135}$$

where $\widehat{Bias}(\hat{\theta})$ is the $Bias(\hat{\theta})$ in (132) with θ replaced by $\hat{\theta}$ and γ_1 replaced by

$$\hat{\gamma}_1 = \frac{1}{n}\sum^{n}(y_i - \bar{y})^3 / s^3. \tag{136}$$

The estimator $\tilde{\theta}$ is an almost unbiased estimator of θ in the sense that bias of $\tilde{\theta}$ is zero upto $O(n^{-1})$.

Proposition 3.2: If the errors follow (128), the bias of $\tilde{\theta}$, upto $O(n^{-\frac{3}{2}})$, is zero. That is

$$\text{Bias}\left(\tilde{\theta}\right) = 0 \tag{137}$$

The proof of Proposition 3.2 follows by first noting that, under (128), $\hat{\theta}^r = \theta^r + O_p(1/\sqrt{n})$ for a constant $r > 0$, and $\hat{\gamma}_1 = \gamma_1 + O_p(1/\sqrt{n})$. Substituting these in $\widehat{Bias}(\hat{\theta})$ gives $\widehat{Bias}(\hat{\theta}) = Bias(\hat{\theta}) + O(1/n^2)$, and hence from (2.9) $Bias(\hat{\theta}) = O(1/n^2)$ which proves the result in Proposition 3.2.

Proposition 3.3: If the errors follow (128), the mean squared error of $\tilde{\theta}$, upto $O(n^{-2})$, is

$$\begin{aligned} MSE\left(\tilde{\theta}\right) &= \frac{\theta^2}{n}\left[\gamma_2 + 4\theta - 4\gamma_1\theta^{1/2} + 2\frac{n}{n-1}\right] \\ &+\frac{\theta^{\frac{5}{2}}}{n^2}\left\{8\theta^{\frac{1}{2}}\gamma_2\left(1+\frac{3n}{n-1}\right) - 8\gamma_3 + 6\gamma_1\gamma_2 - 4\theta^{\frac{1}{2}}\gamma_1^2\left(2-5\frac{n}{n-1}\right)\right. \\ &-4\gamma_1\left(6+\frac{n}{n-1}\right) + 48\theta^2\gamma_1 - 108\theta\gamma_1 + 8\theta^{\frac{1}{2}}\gamma_1 - 12\theta^{\frac{3}{2}}\gamma_2 \\ &\left.+\theta^{\frac{1}{2}}\left(\frac{20n}{n-1} - 12\theta^{\frac{1}{2}} + \theta(66 - \frac{24n}{n-1}\right) - 48\theta^2\right\} \end{aligned}$$

The proof is given in Appendix A.

When the distribution is normal ($\gamma_1 = \gamma_2 = \gamma_3 = 0$),

$$MSE\left(\tilde{\theta}\right) - MSE(\hat{\theta}) = -\frac{\theta^{\frac{7}{2}}}{n^2}\left[12 + \theta^{\frac{1}{2}}\left(9 + \frac{24n}{n-1}\right) + 48\theta^{\frac{3}{2}}\right]$$

which is clearly negative. As discussed above, the bias problem is most severe for skewed distributions, however, even in the case of normality there will be bias and the suggested adjusted estimator will reduce bias and is shown to have a lower mean squared error. For non-normal distributions, it will not in general be possible to show that $MSE\left(\tilde{\theta}\right) - MSE(\hat{\theta})$ is negative. It will in fact depend upon the shape of the distribution.

Finally, I note from Propositions 1, 2, and 3 that for large samples ($n \to \infty$), Bias $\hat{\theta} \to 0$ and Bias $\tilde{\theta} \to 0$ and

$$\lim_{n \to \infty} nV(\hat{\theta}) = \lim_{n \to \infty} nV(\tilde{\theta}) = \theta^2 \left[\gamma_2 + 4\theta - 4\gamma_1\theta^{1/2} + 2\right] \qquad (138)$$

which is the asymptotic variance of $\hat{\theta}$ and $\tilde{\theta}$. A consistent estimator of the asymptotic variance can be written by substituting $\hat{\theta}$, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ for θ, γ_1 and γ_2, respectively, in (2.12), where $\hat{\gamma}_2$, is given by: $\hat{\gamma}_2 = (n-1)^{-1} \sum \left(\frac{(y_i - \bar{y})^4}{s^4}\right) - 3$.

3.3.3 Empirical application: Kenya and China

For application of the above results I consider two data sets on income: one from Kenya (1986, Central Bureau of Statistics and the Ministry of National Planning and Development of Kenya) on 2,424 urban households and 1988 Chinese data on 9,009 urban households gathered by a group of six UC-Riverside faculty members along with the Chinese Academy of Social Sciences. For additional information on these data sets see Githinji(1996) and Khan, et. al. (1991).

Table 3.6 gives summary statistics for household and per-capita income as well as the estimates of $\hat{\theta}$ and $\tilde{\theta}$. Per-capita income is computed by equally dividing household income among all members of the household and thus almost certainly adds a downward bias to inequality. Household income uses the household as the main unit of analysis, weighting total household income by household size.

Table 3.6
Results on Inequality Measures

Unit of income	Kenyan Shillings	Chinese Yuan
Exchange rate in $US (at survey date)	16.04 KS=1.00 US$	4.86 CY=1.00 US$
Survey date	1986	1988
Sample Size	2,424	9,009
HOUSHOLD INCOME	**KENYA**	**CHINA**
Sample statistics		
Mean	46,628.95	6,507.26
Median	16,813	5759.00
Average Household Size	3.55	3.53
Variance	25,022,439,640.12	10,873,357
Standard Deviation	158,184.83	3297.48
Skewness coefficient	20.74	2.83
CV	3.39	0.50674
CV squared	11.51	0.25678
Corrected CV	3.47	0.5068
Corrected CV squared	12.01	0.25684
Gini coefficient	0.645	0.238
PER-CAPITA INCOME	**KENYA**	**CHINA**
Sample statistics		
Mean	12,204.6	1,841.95
Median	7,451.61	1,700.00
Variance	2,403,269,188.45	842,322.83
Standard Deviation	49,023.15	917.78
Skewness coefficient	27.45	3.04
CV	4.02	0.498
CV squared	16.13	0.248
Corrected CV	4.16	0.498
Corrected CV squared	17.28	0.248
Gini coefficient	0.652	0.222

It is clear from this table and the nonparametric kernel density estimates (Silverman (1986)) that urban inequality is much larger in Kenya than in China. Comparing the almost-unbiased cv-squared $\tilde{\theta}$ with the sample cv-squared as it is usually calculated, $\hat{\theta}$, the two data sets give quite different results. For the China data, the bias correction gives almost no change in cv-squared. In the Kenya data, however, he usual cv-squared estimator, $\hat{\theta}$, gives an underestimate of inequality. This is because for the Kenya data the inequality $\gamma_1 > \frac{3}{2} c\,v$ in (2.8) is satisfied. The lack of difference between $\hat{\theta}$ the $\tilde{\theta}$ in the China data may be due to two reasons. First, the China data set is almost four times larger than the Kenya one so that the Bias $(\hat{\theta})$ in (132) is almost zero– that is $\hat{\theta}$ and hence $\tilde{\theta}$ are both asymptotically unbiased. The second reason is that the China data has a less–skewed distribution– that is the value of $\hat{\gamma}_1$ is small and close to $\frac{3}{2}$ cv making the bias of $\hat{\theta}$ in (132) near zero.

The above findings indicate that when the sample is small or moderately large, or the skewness in the distribution is either negative or when it is $>\frac{3}{2}$ cv the almost unbiased estimator $\tilde{\theta}$ will be useful for correcting bias in $\hat{\theta}$. Although it is beyond the scope of this paper, similar results could be developed for other inequality measures, see Maasoumi (1991), Sen (1992) and Ravallion (1994).

3.3.4 Simulation

Now I turn to the question of how important the bias in the coefficient of variation is for small samples. I attempt to answer this question through two simulation exercises. First, using the Kenya data as our "population", I drew a random sample (with replacement) of size 100, and calculated both the sample coefficient of variation and the "almost unbiased" CV for this new sample. I then repeated this exercise 1000 times. For the per-capita incomes, on average, the almost-unbiased estimator of CV gave an increase of 17% over the uncorrected cv (calculated in the usual way as the sample standard deviation divided by the sample mean). In several cases, the bias correction changed the uncorrected cv by over 100%. (The maximum increase from $\hat{\theta}$ to $\tilde{\theta}$ was 112%.) Thus, in small sample, CV (uncorrected) may understate the true coefficient of variation by more than half.

When we increase the sample size, the average change from $\hat{\theta}$ to the almost-unbiased estimator becomes smaller, as expected. At a sample size of 200, the average change is 15.8%. At a sample size of 500, the average change is 11%. In the simulation with sample sizes of 200, the maximum change from $\hat{\theta}$ to $\tilde{\theta}$ was 70%; in that with sample size 500, 34%.

In the second simulation exercise, I compare the empirical bias and mean-squared error of $\widehat{cv} = \frac{\sqrt{s^2}}{\bar{y}}$ with $\widetilde{cv} = \sqrt{\tilde{\theta}}$ (see equation (135)) and a leave-one out jackknife estimator, cv_{jkn} calculated in the usual way. (Efron (1982, p. 6)) Again, using the

Kenya sample as my population, I consider seven different sample sizes ranging from n=50 to n=500. Drawing a sample of size n with replacement, I calculate $\widehat{cv}$, $\widetilde{cv}$, and cv_{jkn}. This exercise is repeated 5000 times for each sample size considered. Results are summarized in Table 3.7 and Figures 3.1 and 3.2.

The jackknife estimator has the lowest average bias and the highest mean squared error throughout, as can be seen from Table 3.7 and Figure 3.2. The high variability of the estimator is a well-known problem of the jackknife (see Hinkley (1978), for example). Considering the bias, which is presented in the first part of Table 3.7 and in Figure 3.1, the proposed adjusted cv ($\widetilde{cv}$) dominates cv as it is usually calculated ($\widehat{cv}$) for all sample sizes which were considered. However, at medium and large sample sizes, the adjusted cv has a slightly higher mean squared error than cv. However, compared to the jackknife, the difference is very small (see Figure 3.2). Non-parametric density estimates of the 5000 repetitions for the three different estimators are provided in Figure 3.3 for the case of n=200. The non-parametric densities for the other sample sizes were similar.

Table 3.7
Average Bias and Mean Squared Error
for Three Estimators of the Coefficient of Variation

Simulation based on 5000 repetitions

	Average Bias			Average Mean Squared Error		
n	$\hat{C}V$	$\tilde{C}V$	CV_{jkn}	$\hat{C}V$	$\tilde{C}V$	CV_{jkn}
50	-2.23	-2.2	-1.72	6.160	5.848	6.537
100	-1.75	-1.61	-1.00	4.820	4.600	6.871
150	-1.41	-1.22	-0.68	4.045	4.026	6.467
200	-1.19	-0.97	-0.48	3.327	3.582	5.879
300	-0.94	-0.7	-0.29	2.768	2.970	4.578
400	-0.76	-0.54	-0.20	2.286	2.483	3.559
500	-0.61	-0.39	-0.095	1.901	2.095	2.922

Figure 3.1
Bias for Three Estimators of the Coefficient of Variation

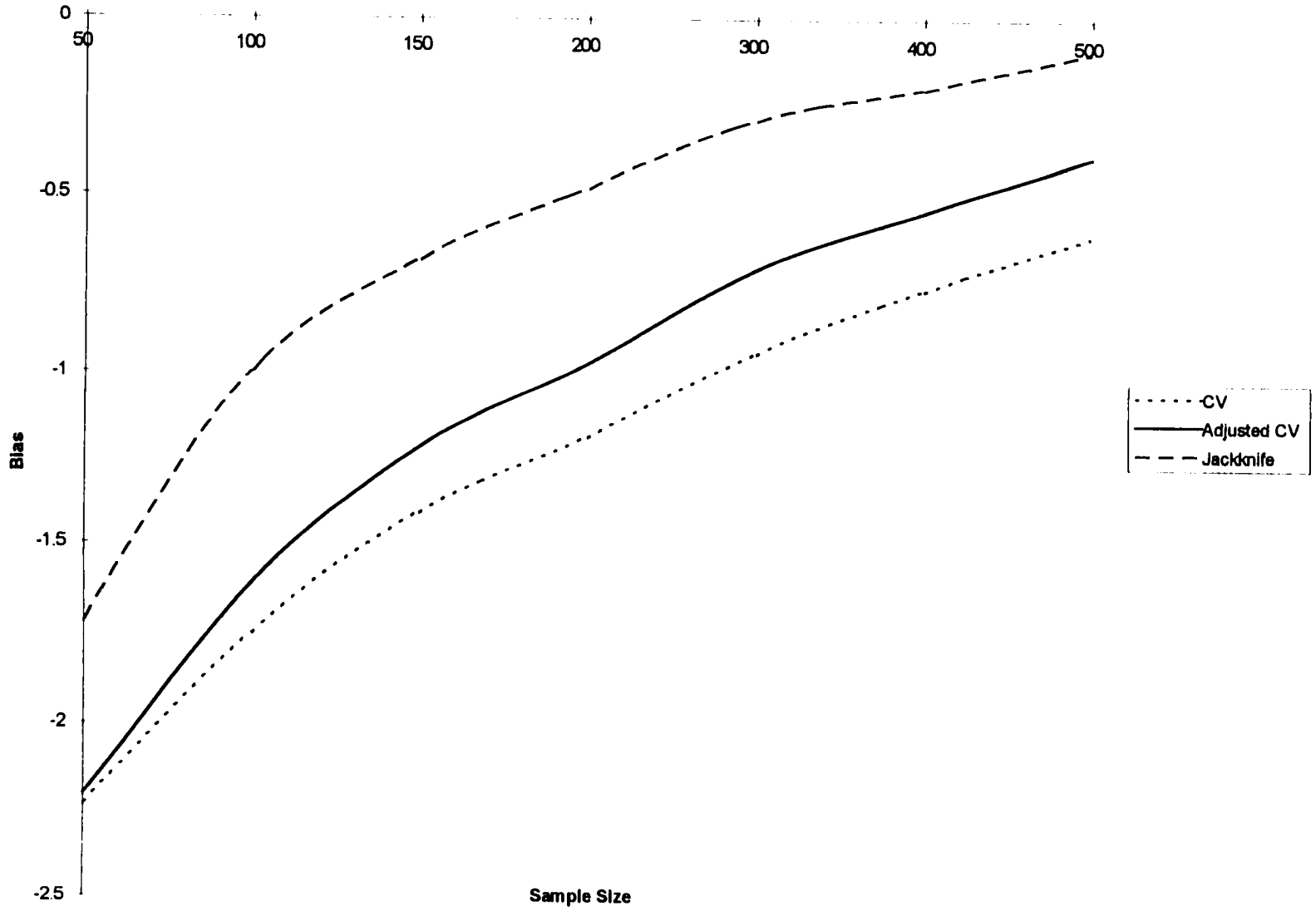


Figure 3.2
Mean Squared Error for Three Estimators of the Coefficient of Variation

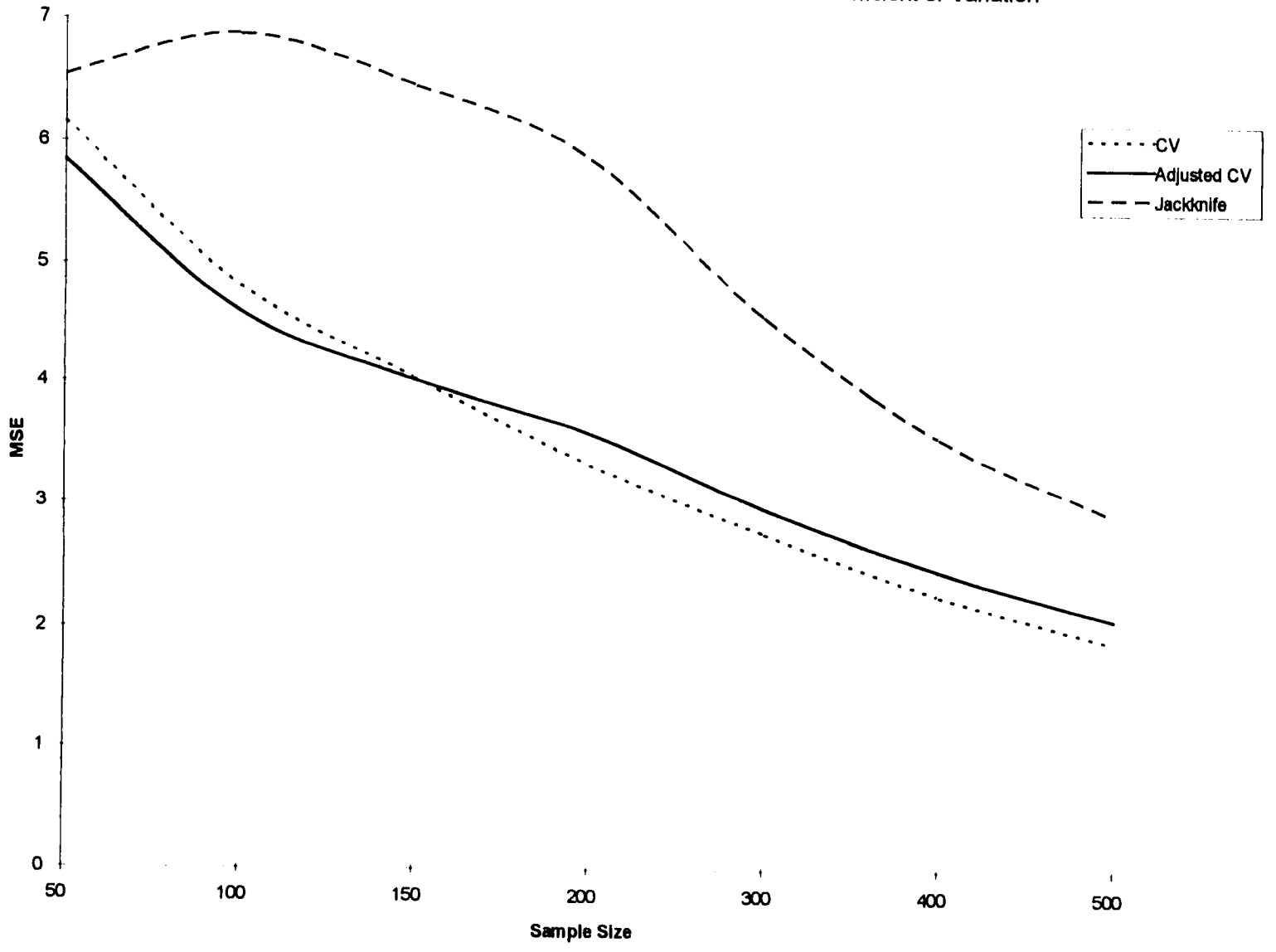
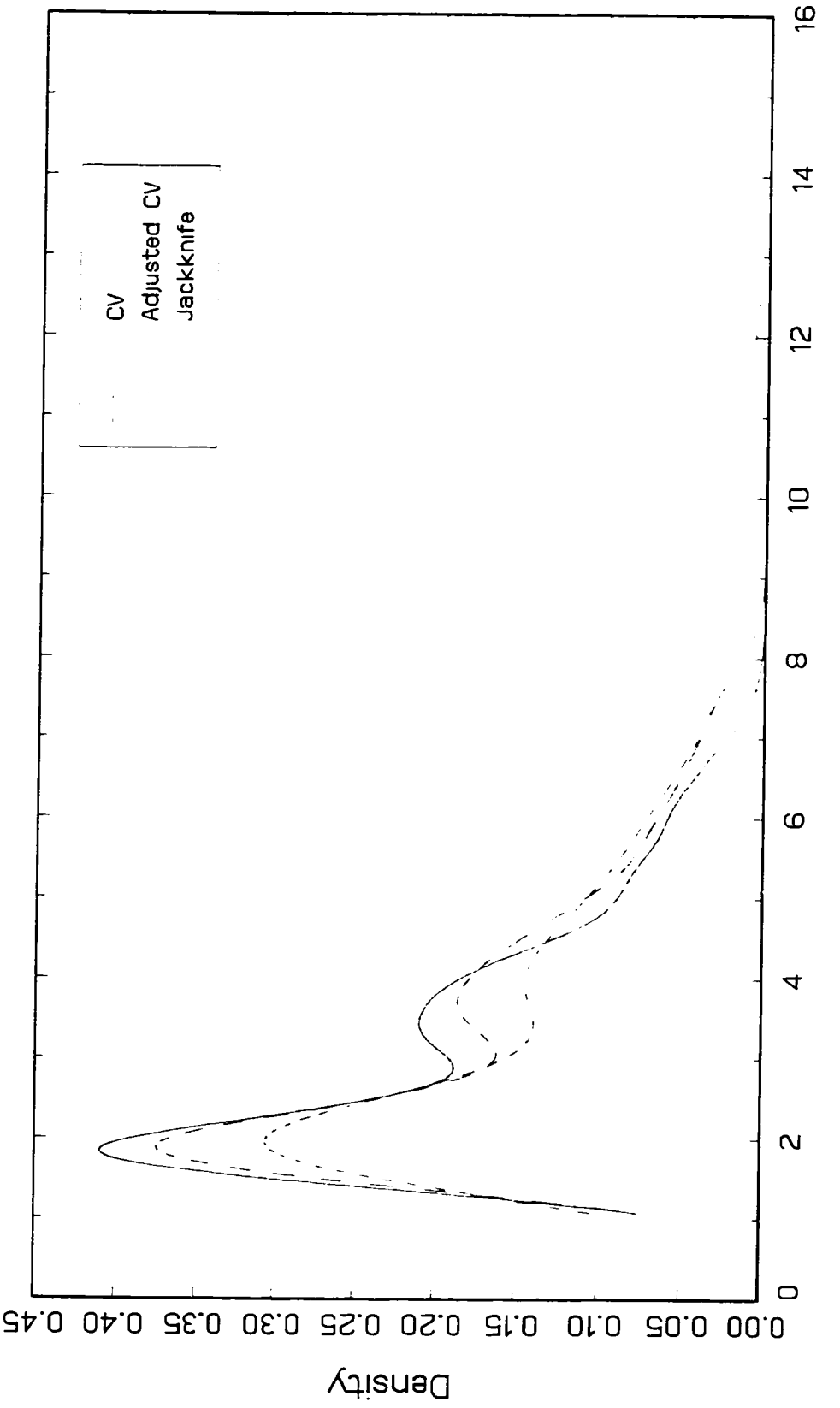


Figure 3.3
Nonparametric Density Estimates of Simulation Results:
Three Estimators of the Coefficient of Variation



3.3.5 Conclusion

$\widetilde{cv}$ is shown to reduce bias compared to $\widehat{cv}$, with only a small loss of mean squared error. In cases where bias is the main concern of the practitioner, $\widetilde{cv}$ is thus preferable. Although the jackknife gives lower bias, the extremely high mean squared error would indicate that one should be suspicious of the results for any given case. $\widetilde{cv}$ thus provides a bias reduction that does not come at the expense of a large increase in variance.

Now let us turn to issues of the sampling structure in estimating inequality measures. Like the mean case, we will have to adjust our standard errors in the case of sampling without replacement and clustering. When we have stratified data, weighting will be required to have unbiased estimators.

3.4 Random Sampling without replacement

When the sampling is without replacement, our estimators from (87) through (92) will remain consistent, though biased in small samples as discussed above. However, inference will be affected when the sampling is without replacement from a finite population.

We saw in section 2 that when the sampling is without replacement (RSWOR) that the variance of the mean, $\overline{y}$ needs to be adjusted by the finite population correction term to adjust for the correlation which is induced by sampling without replacement

$$Var\left(\overline{y}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right). \tag{139}$$

How will the standard errors which we calculated for inequality measures need to be adjusted for the case of sampling without replacement from a finite population?

Consider the example of the coefficient of variation. For moderately large samples, we can use the approximation to the variance of (96)

$$\begin{aligned} Var(\widehat{I}(y)) &= Var(\widehat{g}(y))\left(\frac{\partial I(y)}{\partial \widehat{g}(y)}\right)^2 + 2Cov(\widehat{g}(y),\overline{y})\frac{\partial I(y)}{\partial \widehat{g}(y)}\frac{\partial I(y)}{\partial \overline{y}} \\ &\quad + Var(\overline{y})\left(\frac{\partial I(y)}{\partial \overline{y}}\right)^2 + O\left(\frac{1}{n^{3/2}}\right). \end{aligned} \tag{140}$$

where $Var(\overline{y})$ is now (139), and we can solve for the $Var(\widehat{g}(y))$ and the $Cov(\widehat{g}(y),\overline{y})$ by the same method that was used in section 3.2.

For RSWOR,

$$Var(\hat{\beta}_2) = \frac{1}{n}\left\{(\gamma_2+3)\,\sigma^4 + 4\mu^2\sigma^2 + 4\mu\sigma^3\gamma_1 - \sigma^4\right\} + \frac{4(n-1)}{n}\mu^2\sigma_{12} + \frac{4(n-1)}{n}\mu\sigma_{112} + \frac{(n-1)}{n}\sigma_{1122} \tag{141}$$

and the $Cov(\hat{\beta}_2, \bar{y})$ under RSWOR will be

$$Cov(\hat{\beta}_2, \bar{y}) = \frac{1}{n}\left\{\sigma^3\gamma_1 + 2\mu\sigma^2 + 2(n-1)\mu\sigma_{12} + (n-1)\sigma_{112}\right\} \tag{142}$$

where σ_{12}, σ_{112}, and σ_{1122} are as given in (11). These two results are proven in Appendix B.

Simplifying and rearranging gives

$$\begin{aligned} Var(\hat{\beta}_2) &= \frac{1}{n}\left\{(\gamma_2+3)\,\sigma^4 + 4\mu^2\sigma^2 + 4\mu\sigma^3\gamma_1 - \sigma^4\right\}\left(\frac{N-n}{N-1}\right) \\ &= Var_{RSWR}(\hat{\beta}_2)\left(\frac{N-n}{N-1}\right) \end{aligned} \tag{143}$$

and

$$\begin{aligned} Cov(\hat{\beta}_2, \bar{y}) &= \frac{1}{n}\left[\sigma^3\gamma_1 + 2\mu\sigma^2\right]\left(\frac{N-n}{N-1}\right) \\ &= Cov_{RSWR}(\hat{\beta}_2, \bar{y})\left(\frac{N-n}{N-1}\right) \end{aligned} \tag{144}$$

where $Var_{RSWR}(\hat{\beta}_2)$ and are $Cov_{RSWR}(\hat{\beta}_2, \bar{y})$ the values for the variance and covariance under random sampling with replacement from (102) and (103).

For the case of the coefficient of variation, then, we can simply calculate the variance of CV under RSWR and correct by the finite population correction, which will be the same as the finite population correction (fpc) from the mean case which we saw in section 2. That gives, upto $O(\frac{1}{n})$

$$\begin{aligned} Var_{RSWOR}(\widehat{CV}) &= \frac{1}{n}\left\{(\gamma_2+2)\frac{\sigma^2}{4\mu^2} - \frac{\sigma^3\gamma_1}{\mu^3} + \frac{\sigma^4}{\mu^4}\right\}\left(\frac{N-n}{N-1}\right) \\ &= Var_{RSWR}(\widehat{CV})\left(\frac{N-n}{N-1}\right) \end{aligned} \tag{145}$$

Similar methods can be used to calculate the finite population correction (fpc) for the other inequality measures. We note here that it may not be the case that the fpc is the same as that for the mean case for the other measures.

3.5 Stratification: Inequality estimation and inference

Many of the samples which economists use are gathered through some type of stratification scheme. The population mean model will thus be

$$Y_{hi} = \mu_h + U_{hi}\,,\ h = 1, ..., H,\ i = 1, ..., N_h \tag{146}$$

where Y_{hi} is the *i-th* unit in the *h-th* stratum, μ_h is the mean of the *h-th* stratum and U_{hi} is the error.

The stratified sample observations, generated by RSWOR in each stratum, follow

$$y_{hi} = \mu_h + u_{hi}, \qquad h = 1, ..., H, \qquad i = 1, ..., n_h. \tag{147}$$

As we saw in section 2, an unbiased estimate of the mean can be written as a weighted sum of the stratum-specific means. For all widely-used inequality measures, this will no longer be the case. $\sum \frac{N_h}{N} \hat{I}_h$ will underestimate the population inequality since it takes into account only the within-stratum inequality and not the between-stratum inequality.

If the stratified sample is chosen such that the proportion of elements sampled is equal in all strata, then the sample is self-weighting and we can estimate inequality following (87) through (92). If not, we need to develop some type of weighted estimator in order to have consistent estimation of inequality.

One approach in this case is to apply the literature on decomposable inequality measures. Bourguignon (1979), Shorrocks (1980) and Cowell (1989) give expressions for calculating the inequality measures in (87) through (92) as combinations of within-group and between-group inequality. For example, the estimate of Theil's measure, $\hat{I}_1$, becomes

$$\begin{aligned} & \frac{1}{n} \sum_h \sum_{n_h} \frac{y_{hi}}{\overline{y}} \log \frac{y_{hi}}{\overline{y}} \\ = & \sum_h \frac{n_h \overline{y}_h}{n \overline{y}} \hat{I}_{1,h} + \frac{1}{n} \sum_h n_h \frac{\overline{y}_h}{\overline{y}} \log \frac{\overline{y}_h}{\overline{y}} \end{aligned} \tag{148}$$

when the sampling is proportional. As we can see, this is merely the sum of within-group inequality indices, $\hat{I}_{1,h}$, and another term which provides a measure of between group inequality.

This will be biased whenever the sampling probabilities are unequal between subgroups (strata), analogous to the mean case. It is straightforward to generalize this method to the non-proportional sampling case by writing a weighted version of this estimator

$$\hat{I}_{1,w} = \sum_{h}\sum_{n_h} w_{hi}\frac{y_{hi}}{\overline{y}_w}\log\frac{y_{hi}}{\overline{y}_w}. \tag{149}$$

Making the weights

$$w_{hi} = w_h = \frac{N_h}{Nn_h} \tag{150}$$

will yield

$$\hat{I}_{1,w} = \sum_{h}\frac{N_h\overline{y}_h}{N\overline{y}_w}\hat{I}_{1,h} + \sum_{h}\frac{N_h}{N}\frac{\overline{y}_h}{\overline{y}_w}\log\frac{\overline{y}_h}{\overline{y}_w}. \tag{151}$$

The weights in the proportional case are $w_{hi} = w_h = \frac{1}{n}$.

Showing that Theil's estimator can be decomposed this way in the non-proportional sampling case can be done by following Shorrocks (1980). It should be intuitively clear however, that just as in the mean case, our subgroup measures need to be weighted by population proportions, $\frac{N_h}{N}$, instead of by sampling proportions, $\frac{n_h}{n}$.

Cowell (1989) provides standard errors for inequality measures calculated from subgroup data. Standard errors and confidence intervals which take into account the non-proportional sampling can be formed by following the approach in that paper, keeping the above discussion in mind.

In practice, two problems immediately present themselves. The first is that there may be a very large number of strata and only a few elements sampled in each strata. (Pudney (1989) for example cites a sample from the U.K. with several thousand strata and many strata with only one or two households sampled.) Confidence intervals composed following the method above and Cowell (1989) will be based upon within-strata variances. Calculating these sub-group (stratum-specific) variances on the basis of only two or three observations will give odd results indeed! The second problem for the practitioner is that quite frequently we are given data which has been drawn in a stratified or multi-stratified sample and are given weights to inflate to population totals, but we are not given information on which element of the sample belongs to which stratum. We will thus be unable to implement the above methodology.

Instead of treating each stratum as a subgroup, therefore, we need to find a way to consider a pooled estimate of the data which takes into account the non-proportional sampling. We can generalize our sample estimator of inequality (94) by writing

$$\hat{I}_w(y) = F(\hat{g}_w(y), \overline{y}_w). \tag{152}$$

$\overline{y}_w$ will be as in (46) and $\hat{g}_w(y) = \sum_{i=1}^{n} w_i g(y_i)$. The weights in the case of equal probability sampling are

$$w_i = \frac{1}{n} \tag{153}$$

and are identical for all elements in the sample. As pointed out above, however, equal

probability sampling will be the exception in the household income and expenditure surveys used for inequality analysis.

For unbiased estimation of those inequality measures which can be written in the form (152) the weights should be chosen proportional to the inverse of the sampling probabilities, just as in the mean case

$$w_i \alpha \frac{1}{\pi_i}. \tag{154}$$

The same line of reasoning holds—elements which enter the sample with higher probability need to be deflated with a relatively smaller weight than elements entering the sample with a low probability.

In the case of unequal probability sampling, the Coefficient of Variation (CV), for example, will be estimated as the ratio of the weighted estimates of the standard deviation and the mean

$$\widehat{CV}_w = \frac{\sqrt{s_w^2}}{\bar{y}_w} \tag{155}$$

where s_w^2 is an unbiased estimator of the variance of y.

When we are given weighted data with weights smaller than one (which may be due to the normalization used) we note that $s_w^2 = (w_t - w_s)^{-1} \sum w_i (y_i - \bar{y}_w)^2$ will be unbiased for σ^2 where w_t is the total of the weights, and w_s is the smallest weight. If all weights are greater than or equal to one, $w_s = 1$ will give unbiased estimation of the variance.

We can likewise develop weighted estimators of the other inequality measures

considered in (88) through (92)

$$\widehat{I_{1,w}} = \overline{y_w}^{-1} \sum_{i=1}^{n} w_i y_i \ln(y_i) - \ln(\overline{y}_w), \tag{156}$$

$$\widehat{I_{0,w}} = \ln(\overline{y}_w) - \sum_{i=1}^{n} w_i \ln(y_i); \tag{157}$$

$$\widehat{A_w(\epsilon)} = 1 - \overline{y_w}^{-1} \left(\sum_{i=1}^{n} w_i y_i^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}}. \tag{158}$$

When $\epsilon = 1$ in Atkinson's measure, the weighted estimator maybe written as

$$\widehat{A_w(1)} = 1 - \overline{y_w}^{-1} e^{\sum_{i=1}^{n} w_i \ln(y_i)} \tag{159}$$

$$= 1 - e^{-\widehat{I_{0,w}}}. \tag{160}$$

These are of course general expressions, since setting $w_i = \frac{1}{n}$ for all observations gives the simple random sampling result.

To calculate standard errors, we can use the formulas in (107) to (110) and replace the various moments with their weighted estimators. Thus, an estimate of the variance of the coefficient of variation, for example, becomes

$$\widehat{Var}(\widehat{CV}) = \frac{1}{n} \left\{ (\gamma_{2,w} + 2) \frac{s_w^2}{4\overline{y}_w^2} - \frac{s_w^3 \gamma_{1,w}}{\overline{y}_w^3} + \frac{s_w^4}{\overline{y}_w^4} \right\} \tag{161}$$

where $\overline{y}_w$ and s_w^2 are given above,

$$\gamma_{1,w} = s_w^{-3} \sum w_i (y_i - \overline{y}_w)^3 \tag{162}$$

and

$$\gamma_{2,w} = s_w^{-4} \sum w_i (y_i - \overline{y}_w)^4 - 3. \tag{163}$$

This should provide a slight over-estimation of the variance, since we are ignoring the fact that the sampling is conducted independently across strata. For a simple model, we explore how well this approximation will work through a simulation exercise.

3.5.1 Simulation

For the simulation, we construct the following population models for two strata

$$Y_{1i} = \mu_1 + U_{1i} \tag{164}$$

$$Y_{2i} = \mu_2 + U_{2i} \tag{165}$$

where $\mu_1 = 100$, $\mu_2 = 200$, $\sigma_1^2 = 20$, and $\sigma_2^2 = 20$. This provides a "wealthy" stratum (stratum 2) and a "poor" stratum (stratum one), as well as more income variation in the poorer stratum than in the richer one. The coefficient of variation in stratum 1 will be .2 and in stratum 2, it will equal .1. The disturbance terms, U, for the simulated populations were drawn from normal populations. Each stratum is given total population of 100,000.

Table 3.8 gives the population values for the five inequality indices which we have been considering. Note that in all cases, most of the inequality is coming from the between-group term as opposed to the within-group inequality. This will be typical of the income distributions in many developing nations where large differences in average income exist between rural and urban populations.

Table 3.8
Inequality Measurement under Stratified Sampling:
Population Values for Two Simulated Strata

	Stratum 1	Stratum 2	Total
Population	100,000	100,000	200,000
CV	.2	.1	.359
T(0)	.0216	.0051	.0722
T(1)	.0207	.0050	.0668
A(1)	.0214	.0050	.0697
A(2)	.0446	.0102	.1408

Both strata generated from a normal distribution
σ_1=20 and σ_2=20
μ_1=100 and μ_2=200

We would like to consider the effects of non-proportional sampling from the two strata on our estimates of inequality measures. We will draw unequal sized samples from the two strata and consider the weighted and unweighted estimates of inequality. The samples in the simulation are drawn with replacement, thus we will not have to consider any effects of the finite population when we examine inference under stratification below.

Using the weighted measures of (155) through (160) gives unbiased estimation of inequality in the stratified case as can be seen from Tables 3.9 through 3.13. In the first panel of each table, we present the results for non-proportional sampling when stratum 2 is sampled twice as intensively as stratum 1 (recall that the two strata are the same size.) In the second and third panels of each table, we provide the results when we increase the disproportion in the sampling. The second panel provides the results from the case when three times as many elements are drawn from stratum 2 as from stratum 1 and the third panel looks at a four to one ratio of sampling. In all cases, we are over-sampling the stratum with lower inequality and as we see, when we do not weight the data, we have persistent negative bias in the estimate of inequality. This result holds true for all five indices considered.

We would also like to examine the weighted approximations of the variance by using the asymptotically normal variances provided in section 3.2.2 and their weighted analogs as suggested above. These results are given in Tables 3.14a through 3.18b for

the five inequality measures.

The first result, which is quite surprising, is that these weighted approximations will tend to under-estimate the true variance of the inequality measure. This is surprising, since in the mean case such weighted estimation will generally lead to slight over-estimation of the variance since such a calculation does not take into account the independence of the samples in different strata. The simulation shows that in the inequality case, this problem of underestimation of the variances increases as the dis-proportionality of sampling grows. When stratum 2 is sampled twice as intensively as stratum 1, the weighted variance approximation for CV and I(1) give slight overestimation of the true (simulated variance). Column seven in Tables 3.14a and 3.14b show this. The approximation for the variances of I(0) and A(1) both give a 20% underestimate, as we can see from column seven of Tables 3.15a, 3.15b, 3.17a and 3.17b. This shouldn't be surprising given the relationship between these two measures shown above. From Tables 3.18a and 3.18b, we can see that the variance estimate for A(2) does the worst of the five inequality measures considered, underestimating the simulated variance by almost 30%.

When the sampling disproportion grows to a 4 to 1 ratio, all of the inequality measures have variance approximations which underestimate the true variance. These range from an underestimation of about 10% in the case of the coefficient of variation to 50% underestimation in the case of Atkinson's measure with $\epsilon = 2$.

For all of the inequality measures considered, as sample sizes grow, the estimated

variances do decrease as expected. The estimated variance of I(1) does quite poorly for very small sample sizes, as seen above in the simulation of section 3.2.3.

Recall that the purpose of considering this method of estimating the variance of inequality measures was to deal with the problem of insufficient information about the sample. When full information about the sampling is available, it is certainly more effective to consider each stratum separately and use the results of Cowell (1991). However, what is shown here is that even when we only have information about weights and no information about which element in the sample belongs to which stratum, we can estimate inequality measures unbiasedly and estimate their variances fairly adequately.

The simulation shows that weighted variance estimates for inequality measures should be taken with some caution. When samples are highly non-proportional, these estimates do not perform very well. They may still be taken as lower bounds on the variance, however, as there appears to be a general pattern of under-estimation of variance using weighted approximations.

In section 3.5.2 below, we will apply these results to a stratified data set from Kenya to observe the effects of ignoring the stratification in calculating inequality.

Table 3.9
Inequality Measurement under Stratified Sampling:
Simulation Results for the Coefficient of Variation

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	CV	Bias	CV	Bias
5	10	15	0.3165	-0.0425	0.3673	0.0083
10	20	30	0.3103	-0.0487	0.3614	0.0024
20	40	60	0.3093	-0.0497	0.3608	0.0018
50	100	150	0.3084	-0.0506	0.3602	0.0012
100	200	300	0.3075	-0.0515	0.3591	0.0001
200	400	600	0.3074	-0.0516	0.3592	0.0002
500	1000	1500	0.3074	-0.0516	0.3592	0.0002
1000	2000	3000	0.3073	-0.0517	0.3591	0.0001
2000	4000	6000	0.3072	-0.0518	0.3589	-0.0001
3000	6000	9000	0.3073	-0.0517	0.3591	0.0001
5000	10000	15000	0.3072	-0.0518	0.3590	0.0000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	CV	Bias	CV	Bias
5	15	20	0.2780	-0.0810	0.3637	0.0047
10	30	40	0.2745	-0.0845	0.3602	0.0012
20	60	80	0.2736	-0.0854	0.3599	0.0009
50	150	200	0.2729	-0.0861	0.3593	0.0003
100	300	400	0.2725	-0.0865	0.3587	-0.0003
200	600	800	0.2726	-0.0864	0.3591	0.0001
500	1500	2000	0.2727	-0.0863	0.3593	0.0003
1000	3000	4000	0.2725	-0.0865	0.3590	0.0000
2000	6000	8000	0.2726	-0.0864	0.3592	0.0002
3000	9000	12000	0.2725	-0.0865	0.3590	0.0000
5000	15000	20000	0.2725	-0.0865	0.3590	0.0000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	CV	Bias	CV	Bias
5	20	25	0.2525	-0.1065	0.3624	0.0034
10	40	50	0.2495	-0.1095	0.3593	0.0003
20	80	100	0.2492	-0.1098	0.3596	0.0006
50	200	250	0.2489	-0.1101	0.3596	0.0006
100	400	500	0.2484	-0.1106	0.3588	-0.0002
200	800	1000	0.2485	-0.1105	0.3592	0.0002
500	2000	2500	0.2485	-0.1105	0.3592	0.0002
1000	4000	5000	0.2485	-0.1105	0.3591	0.0001
2000	8000	10000	0.2484	-0.1106	0.3590	0.0000
3000	12000	15000	0.2484	-0.1106	0.3590	0.0000
5000	20000	25000	0.2484	-0.1106	0.3590	0.0000

Table 3.10
Inequality Measurement under Stratified Sampling:
Simulation Results for Theil's Inequality Measure I(0)

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	I(0)	Bias	I(0)	Bias
5	10	15	0.0594	-0.0128	0.0723	0.0001
10	20	30	0.0588	-0.0134	0.0716	-0.0006
20	40	60	0.0592	-0.0130	0.0721	-0.0001
50	100	150	0.0594	-0.0128	0.0723	0.0001
100	200	300	0.0592	-0.0130	0.0720	-0.0002
200	400	600	0.0592	-0.0130	0.0721	-0.0001
500	1000	1500	0.0593	-0.0129	0.0723	0.0001
1000	2000	3000	0.0593	-0.0129	0.0722	0.0000
2000	4000	6000	0.0593	-0.0129	0.0722	0.0000
3000	6000	9000	0.0593	-0.0129	0.0722	0.0000
5000	10000	15000	0.0593	-0.0129	0.0722	0.0000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	I(0)	Bias	I(0)	Bias
5	15	20	0.0487	-0.0235	0.0717	-0.0005
10	30	40	0.0486	-0.0236	0.0715	-0.0007
20	60	80	0.0489	-0.0233	0.0721	-0.0001
50	150	200	0.0489	-0.0233	0.0721	-0.0001
100	300	400	0.0488	-0.0234	0.0719	-0.0003
200	600	800	0.0489	-0.0233	0.0722	0.0000
500	1500	2000	0.0490	-0.0232	0.0723	0.0001
1000	3000	4000	0.0489	-0.0233	0.0722	0.0000
2000	6000	8000	0.0490	-0.0232	0.0723	0.0001
3000	9000	12000	0.0489	-0.0233	0.0722	0.0000
5000	15000	20000	0.0489	-0.0233	0.0722	0.0000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	I(0)	Bias	I(0)	Bias
5	20	25	0.0416	-0.0306	0.0718	-0.0004
10	40	50	0.0412	-0.0310	0.0713	-0.0009
20	80	100	0.0416	-0.0306	0.0721	-0.0001
50	200	250	0.0417	-0.0305	0.0724	0.0002
100	400	500	0.0415	-0.0307	0.0720	-0.0002
200	800	1000	0.0417	-0.0305	0.0723	0.0001
500	2000	2500	0.0417	-0.0305	0.0723	0.0001
1000	4000	5000	0.0417	-0.0305	0.0723	0.0001
2000	8000	10000	0.0416	-0.0306	0.0722	0.0000
3000	12000	15000	0.0416	-0.0306	0.0722	0.0000
5000	20000	25000	0.0416	-0.0306	0.0722	0.0000

Table 3.11
Inequality Measurement under Stratified Sampling:
Simulation Results for Theil's Inequality Measure I(1)

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	I(1)	Bias	I(1)	Bias
5	10	15	0.0519	-0.0149	0.0672	0.0004
10	20	30	0.0514	-0.0154	0.0664	-0.0004
20	40	60	0.0518	-0.0150	0.0668	0.0000
50	100	150	0.0519	-0.0149	0.0670	0.0002
100	200	300	0.0517	-0.0151	0.0667	-0.0001
200	400	600	0.0518	-0.0150	0.0668	0.0000
500	1000	1500	0.0518	-0.0150	0.0669	0.0001
1000	2000	3000	0.0518	-0.0150	0.0668	0.0000
2000	4000	6000	0.0518	-0.0150	0.0668	0.0000
3000	6000	9000	0.0518	-0.0150	0.0668	0.0000
5000	10000	15000	0.0518	-0.0150	0.0668	0.0000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	I(1)	Bias	I(1)	Bias
5	15	20	0.0416	-0.0252	0.0668	0.0000
10	30	40	0.0415	-0.0253	0.0664	-0.0004
20	60	80	0.0417	-0.0251	0.0668	0.0000
50	150	200	0.0417	-0.0251	0.0668	0.0000
100	300	400	0.0416	-0.0252	0.0666	-0.0002
200	600	800	0.0417	-0.0251	0.0668	0.0000
500	1500	2000	0.0418	-0.0250	0.0669	0.0001
1000	3000	4000	0.0417	-0.0251	0.0668	0.0000
2000	6000	8000	0.0418	-0.0250	0.0669	0.0001
3000	9000	12000	0.0417	-0.0251	0.0668	0.0000
5000	15000	20000	0.0417	-0.0251	0.0668	0.0000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	I(1)	Bias	I(1)	Bias
5	20	25	0.0351	-0.0317	0.0669	0.0001
10	40	50	0.0348	-0.0320	0.0663	-0.0005
20	80	100	0.0350	-0.0318	0.0668	0.0000
50	200	250	0.0351	-0.0317	0.0670	0.0002
100	400	500	0.0350	-0.0318	0.0667	-0.0001
200	800	1000	0.0351	-0.0317	0.0669	0.0001
500	2000	2500	0.0351	-0.0317	0.0669	0.0001
1000	4000	5000	0.0351	-0.0317	0.0669	0.0001
2000	8000	10000	0.0351	-0.0317	0.0668	0.0000
3000	12000	15000	0.0351	-0.0317	0.0668	0.0000
5000	20000	25000	0.0351	-0.0317	0.0668	0.0000

Table 3.12
Inequality Measurement under Stratified Sampling:
Simulation Results for Atkinson's Inequality Measure A(1)

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	A(1)	Bias	A(1)	Bias
5	10	15	0.05759	-0.01207	0.06958	-0.00008
10	20	30	0.05702	-0.01264	0.06896	-0.00070
20	40	60	0.05749	-0.01217	0.06954	-0.00012
50	100	150	0.05767	-0.01199	0.06977	0.00011
100	200	300	0.05745	-0.01221	0.06950	-0.00016
200	400	600	0.05752	-0.01214	0.06958	-0.00008
500	1000	1500	0.05760	-0.01206	0.06971	0.00005
1000	2000	3000	0.05756	-0.01210	0.06966	0.00000
2000	4000	6000	0.05753	-0.01213	0.06962	-0.00004
3000	6000	9000	0.05758	-0.01208	0.06969	0.00003
5000	10000	15000	0.05756	-0.01210	0.06966	0.00000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	A(1)	Bias	A(1)	Bias
5	15	20	0.04750	-0.02216	0.06905	-0.00061
10	30	40	0.04737	-0.02229	0.06896	-0.00070
20	60	80	0.04768	-0.02198	0.06954	-0.00012
50	150	200	0.04771	-0.02195	0.06958	-0.00008
100	300	400	0.04760	-0.02206	0.06940	-0.00026
200	600	800	0.04776	-0.02190	0.06966	0.00000
500	1500	2000	0.04781	-0.02185	0.06974	0.00008
1000	3000	4000	0.04775	-0.02191	0.06964	-0.00002
2000	6000	8000	0.04779	-0.02187	0.06971	0.00005
3000	9000	12000	0.04776	-0.02190	0.06966	0.00000
5000	15000	20000	0.04776	-0.02190	0.06966	0.00000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	A(1)	Bias	A(1)	Bias
5	20	25	0.04066	-0.02900	0.06911	-0.00055
10	40	50	0.04037	-0.02929	0.06878	-0.00088
20	80	100	0.04072	-0.02894	0.06954	-0.00012
50	200	250	0.04084	-0.02882	0.06978	0.00012
100	400	500	0.04065	-0.02901	0.06941	-0.00025
200	800	1000	0.04080	-0.02886	0.06973	0.00007
500	2000	2500	0.04080	-0.02886	0.06973	0.00007
1000	4000	5000	0.04080	-0.02886	0.06973	0.00007
2000	8000	10000	0.04078	-0.02888	0.06967	0.00001
3000	12000	15000	0.04077	-0.02889	0.06966	0.00000
5000	20000	25000	0.04078	-0.02888	0.06967	0.00001

Table 3.13
Inequality Measurement under Stratified Sampling:
Simulation Results for Atkinson's Inequality Measure A(2)

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	A(2)	Bias	A(2)	Bias
5	10	15	0.12381	-0.01694	0.13861	-0.00214
10	20	30	0.12314	-0.01761	0.13836	-0.00239
20	40	60	0.12447	-0.01628	0.14008	-0.00067
50	100	150	0.12500	-0.01575	0.14076	0.00001
100	200	300	0.12450	-0.01625	0.14025	-0.00050
200	400	600	0.12467	-0.01608	0.14045	-0.00030
500	1000	1500	0.12494	-0.01581	0.14081	0.00006
1000	2000	3000	0.12487	-0.01588	0.14074	-0.00001
2000	4000	6000	0.12480	-0.01595	0.14066	-0.00009
3000	6000	9000	0.12495	-0.01580	0.14084	0.00009
5000	10000	15000	0.12488	-0.01587	0.14075	0.00000

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	A(2)	Bias	A(2)	Bias
5	15	20	0.10632	-0.03443	0.13739	-0.00336
10	30	40	0.10648	-0.03427	0.13828	-0.00247
20	60	80	0.10753	-0.03322	0.14012	-0.00063
50	150	200	0.10766	-0.03309	0.14039	-0.00036
100	300	400	0.10741	-0.03334	0.14011	-0.00064
200	600	800	0.10782	-0.03293	0.14070	-0.00005
500	1500	2000	0.10794	-0.03281	0.14086	0.00011
1000	3000	4000	0.10781	-0.03294	0.14071	-0.00004
2000	6000	8000	0.10792	-0.03283	0.14086	0.00011
3000	9000	12000	0.10784	-0.03291	0.14076	0.00001
5000	15000	20000	0.10783	-0.03292	0.14074	-0.00001

Sample Sizes			Unweighted estimate		Weighted estimate	
Stratum 1	Stratum 2	Total	A(2)	Bias	A(2)	Bias
5	20	25	0.09329	-0.04746	0.13731	-0.00344
10	40	50	0.09290	-0.04785	0.13768	-0.00307
20	80	100	0.09413	-0.04662	0.14011	-0.00064
50	200	250	0.09447	-0.04628	0.14081	0.00006
100	400	500	0.09394	-0.04681	0.14002	-0.00073
200	800	1000	0.09444	-0.04631	0.14090	0.00015
500	2000	2500	0.09443	-0.04632	0.14090	0.00015
1000	4000	5000	0.09444	-0.04631	0.14093	0.00018
2000	8000	10000	0.09437	-0.04638	0.14081	0.00006
3000	12000	15000	0.09434	-0.04641	0.14075	0.00000
5000	20000	25000	0.09435	-0.04640	0.14077	0.00002

Table 3.14a
Weighted Approximations of Inequality Index Variances under Stratified Sampling:
Simulation results for the Coefficient of Variation

Sample sizes St.1	St.2	Total	CV_w	$Var(CV_w)$	$Var_{sim}(CV_w)$	$Var(CV_w)/Var_{sim}(CV_w)$
5	10	15	0.367	0.00206776	0.00186568	1.1083155
10	20	30	0.361	0.00105542	0.00089923	1.1736934
20	40	60	0.360	0.00054663	0.00042696	1.2802874
50	100	150	0.360	0.00022018	0.00016361	1.3457268
100	200	300	0.359	0.00011025	8.90E-05	1.2388519
200	400	600	0.359	5.52E-05	4.03E-05	1.3700043
500	1000	1500	0.359	2.22E-05	1.80E-05	1.2335304
1000	2000	3000	0.359	1.11E-05	8.62E-06	1.2892905
2000	4000	6000	0.359	5.56E-06	4.25E-06	1.3073473
3000	6000	9000	0.359	3.71E-06	3.01E-06	1.2334416
5000	10000	15000	0.359	2.22E-06	1.72E-06	1.2944365

Sample sizes St.1	St.2	Total	CV_w	$Var(CV_w)$	$Var_{sim}(CV_w)$	$Var(CV_w)/Var_{sim}(CV_w)$
5	15	20	0.3637	0.00152388	0.0015792	0.9649673
10	30	40	0.3602	0.00079305	0.0008355	0.9492084
20	60	80	0.3599	0.00041155	0.0003938	1.0450549
50	150	200	0.3593	0.00016552	0.0001736	0.9537090
100	300	400	0.3587	8.26E-05	7.84E-05	1.0540138
200	600	800	0.3591	4.16E-05	3.93E-05	1.0586132
500	1500	2000	0.3592	1.67E-05	1.73E-05	0.9666327
1000	3000	4000	0.3590	8.33E-06	7.85E-06	1.0613886
2000	6000	8000	0.3591	4.17E-06	4.10E-06	1.0185747
3000	9000	12000	0.3590	2.78E-06	2.78E-06	1.0001088
5000	15000	20000	0.3590	1.67E-06	1.68E-06	0.9907968

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population CV=.359

Table 3.14b
Weighted Approximations of Inequality Index Variances
under Stratified Sampling:
Simulation results for the Coefficient of Variation

Sample sizes						
St.1	St.2	Total	CV_w	$Var(CV_w)$	$Var_{sim}(CV_w)$	$Var(CV_w)/Var_{sim}(CV_w)$
5	20	25	0.3624	0.0012163	0.0017285	0.70367462
10	40	50	0.3593	0.00062769	0.00081978	0.76567722
20	80	100	0.3596	0.00032855	0.00039165	0.83886726
50	200	250	0.3596	0.00013316	0.00016796	0.79278659
100	400	500	0.3587	6.60E-05	8.43E-05	0.78258243
200	800	1000	0.3592	3.33E-05	4.16E-05	0.80151357
500	2000	2500	0.3591	1.34E-05	1.54E-05	0.86584341
1000	4000	5000	0.3591	6.68E-06	7.52E-06	0.88915843
2000	8000	10000	0.3590	3.34E-06	4.39E-06	0.76094926
3000	12000	15000	0.3590	2.22E-06	2.58E-06	0.86148283
5000	20000	25000	0.3590	1.33E-06	1.49E-06	0.89626966

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population CV=.359

Table 3.15a
Weighted Approximations of Inequality Index Variances under Stratified Sampling:
Simulation results for Theil's Measure: I(0)

Sample sizes St.1	St.2	Total	$I_w(0)$	$Var(I_w(0))$	$Var_{sim}(I_w(0))$	$Var(I_w(0))/Var_{sim}(I_w(0))$
5	10	15	0.0723	0.00071592	0.00040703	1.758902
10	20	30	0.0716	0.00024625	0.00019017	1.294863
20	40	60	0.0721	0.00010026	9.17E-05	1.092827
50	100	150	0.0723	3.40E-05	3.57E-05	0.951417
100	200	300	0.0720	1.58E-05	1.94E-05	0.813830
200	400	600	0.0721	7.66E-06	8.81E-06	0.868988
500	1000	1500	0.0723	3.04E-06	3.95E-06	0.770573
1000	2000	3000	0.0722	1.51E-06	1.87E-06	0.810509
2000	4000	6000	0.0722	7.52E-07	9.15E-07	0.822380
3000	6000	9000	0.0722	5.04E-07	6.64E-07	0.759045
5000	10000	15000	0.0722	3.01E-07	3.76E-07	0.799702

Sample sizes St.1	St.2	Total	$I_w(0)$	$Var(I_w(0))$	$Var_{sim}(I_w(0))$	$Var(I_w(0))/Var_{sim}(I_w(0))$
5	15	20	0.0717	0.00040623	0.00032336	1.256281
10	30	40	0.0715	0.00015658	0.00018072	0.866424
20	60	80	0.0721	6.85E-05	8.62E-05	0.793999
50	150	200	0.0721	2.44E-05	3.89E-05	0.626360
100	300	400	0.0719	1.16E-05	1.75E-05	0.663747
200	600	800	0.0722	5.75E-06	8.63E-06	0.666146
500	1500	2000	0.0723	2.27E-06	3.88E-06	0.585441
1000	3000	4000	0.0722	1.13E-06	1.74E-06	0.649193
2000	6000	8000	0.0723	5.66E-07	8.91E-07	0.634742
3000	9000	12000	0.0722	3.77E-07	6.05E-07	0.622750
5000	15000	20000	0.0722	2.25E-07	3.67E-07	0.614369

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population I(0)=.0722

Table 3.15b
Weighted Approximations of Inequality Index Variances under Stratified Sampling:
Simulation results for Theil's Measure: I(0)

Sample sizes						
St.1	St.2	Total	$I_w(0)$	$Var(I_w(0))$	$Var_{sim}(I_w(0))$	$Var(I_w(0))/Var_{sim}(I_w(0))$
5	20	25	0.0718	0.00027848	0.0003583	0.777207
10	40	50	0.0713	0.00011087	0.00017274	0.641868
20	80	100	0.0721	5.22E-05	8.55E-05	0.609910
50	200	250	0.0724	1.92E-05	3.68E-05	0.521435
100	400	500	0.0720	9.07E-06	1.87E-05	0.485504
200	800	1000	0.0723	4.61E-06	9.20E-06	0.500416
500	2000	2500	0.0723	1.82E-06	3.43E-06	0.530992
1000	4000	5000	0.0723	9.11E-07	1.68E-06	0.541968
2000	8000	10000	0.0722	4.54E-07	9.78E-07	0.463685
3000	12000	15000	0.0722	3.01E-07	5.82E-07	0.517260
5000	20000	25000	0.0722	1.80E-07	3.32E-07	0.543724

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population I(0)=.0722

Table 3.16a
Weighted Approximations of Inequality Index Variances
under Stratified Sampling:
Simulation results for Theil's Measure: I(1)

Sample sizes St.1	St.2	Total	$I_w(1)$	$Var(I_w(1))$	$Var_{sim}(I_w(1))$	$Var(I_w(1))/Var_{sim}(I_w(1))$
5	10	15	0.0672	0.01707368	0.00028061	60.843990
10	20	30	0.0664	0.00416773	0.00013196	31.582351
20	40	60	0.0668	0.00106844	6.40E-05	16.681458
50	100	150	0.0670	0.00018564	2.47E-05	7.517245
100	200	300	0.0667	5.25E-05	1.35E-05	3.902910
200	400	600	0.0668	1.63E-05	6.11E-06	2.673534
500	1000	1500	0.0669	4.18E-06	2.74E-06	1.523953
1000	2000	3000	0.0668	1.69E-06	1.30E-06	1.300312
2000	4000	6000	0.0668	7.46E-07	6.41E-07	1.163130
3000	6000	9000	0.0668	4.76E-07	4.59E-07	1.038198
5000	10000	15000	0.0668	2.75E-07	2.61E-07	1.050877

Sample sizes St.1	St.2	Total	$I_w(1)$	$Var(I_w(1))$	$Var_{sim}(I_w(1))$	$Var(I_w(1))/Var_{sim}(I_w(1))$
5	15	20	0.0668	0.00840511	0.0002322	36.197929
10	30	40	0.0664	0.00210066	0.0001256	16.724946
20	60	80	0.0668	0.00054802	5.97E-05	9.177736
50	150	200	0.0668	9.89E-05	2.67E-05	3.709581
100	300	400	0.0666	2.94E-05	1.20E-05	2.454923
200	600	800	0.0668	9.82E-06	5.98E-06	1.641853
500	1500	2000	0.0669	2.74E-06	2.66E-06	1.028842
1000	3000	4000	0.0668	1.17E-06	1.20E-06	0.975247
2000	6000	8000	0.0669	5.36E-07	6.20E-07	0.864122
3000	9000	12000	0.0668	3.46E-07	4.21E-07	0.820804
5000	15000	20000	0.0668	2.02E-07	2.55E-07	0.792093

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population I(1)=.0668

Table 3.16b
Weighted Approximations of Inequality Index Variances under Stratified Sampling:
Simulation results for Theil's Measure: I(1)

Sample sizes						
St.1	St.2	Total	$I_w(1)$	$Var(I_w(1))$	$Var_{sim}(I_w(1))$	$Var(I_w(1))/Var_{sim}(I_w(1))$
5	20	25	0.0669	0.00504927	0.00025756	19.604056
10	40	50	0.0663	0.0012716	0.00012218	10.407619
20	80	100	0.0668	0.00033774	5.92E-05	5.709327
50	200	250	0.0670	6.34E-05	2.56E-05	2.476735
100	400	500	0.0667	1.95E-05	1.29E-05	1.514185
200	800	1000	0.0669	6.87E-06	6.34E-06	1.084180
500	2000	2500	0.0669	2.04E-06	2.36E-06	0.862621
1000	4000	5000	0.0669	8.99E-07	1.15E-06	0.780876
2000	8000	10000	0.0668	4.19E-07	6.73E-07	0.622657
3000	12000	15000	0.0668	2.72E-07	3.96E-07	0.686389
5000	20000	25000	0.0668	1.60E-07	2.28E-07	0.701573

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population I(1)=.0668

Table 3.17a
Weighted Approximations of Inequality Index Variances under Stratified Sampling:
Simulation results for Atkinson's Measure: A(1)

Sample sizes St.1	St.2	Total	A_w (1)	Var(A_w (1))	Var_{sim}(A_w (1))	Var(A_w (1))/ Var_{sim}(A_w (1))
5	10	15	0.0696	0.00060882	0.00034432	1.768188
10	20	30	0.0690	0.00021101	0.00016351	1.290475
20	40	60	0.0695	8.63E-05	7.92E-05	1.089858
50	100	150	0.0698	2.93E-05	3.08E-05	0.949743
100	200	300	0.0695	1.36E-05	1.68E-05	0.813083
200	400	600	0.0696	6.62E-06	7.62E-06	0.869027
500	1000	1500	0.0697	2.63E-06	3.42E-06	0.770422
1000	2000	3000	0.0697	1.31E-06	1.62E-06	0.810348
2000	4000	6000	0.0696	6.51E-07	7.92E-07	0.822411
3000	6000	9000	0.0697	4.36E-07	5.74E-07	0.759108
5000	10000	15000	0.0697	2.60E-07	3.26E-07	0.799711

Sample sizes St.1	St.2	Total	A_w (1)	Var(A_w (1))	Var_{sim}(A_w (1))	Var(A_w (1))/ Var_{sim}(A_w (1))
5	15	20	0.0690	0.00034693	0.00027679	1.253402
10	30	40	0.0690	0.00013425	0.00015543	0.863763
20	60	80	0.0695	5.89E-05	7.44E-05	0.792293
50	150	200	0.0696	2.10E-05	3.36E-05	0.626061
100	300	400	0.0694	1.00E-05	1.51E-05	0.663118
200	600	800	0.0697	4.98E-06	7.47E-06	0.665860
500	1500	2000	0.0697	1.96E-06	3.35E-06	0.585550
1000	3000	4000	0.0696	9.79E-07	1.51E-06	0.648998
2000	6000	8000	0.0697	4.90E-07	7.71E-07	0.634725
3000	9000	12000	0.0697	3.26E-07	5.23E-07	0.622708
5000	15000	20000	0.0697	1.95E-07	3.17E-07	0.614334

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population A(1)=.06966

Table 3.17b
Weighted Approximations of Inequality Index Variances under Stratified Sampling:
Simulation results for Atkinson's Measure: A(1)

Sample sizes						
St.1	St.2	Total	$A_w(1)$	$Var(A_w(1))$	$Var_{sim}(A_w(1))$	$Var(A_w(1))/Var_{sim}(A_w(1))$
5	20	25	0.0691	0.00023717	0.00030636	0.774154
10	40	50	0.0688	9.51E-05	0.00014877	0.639479
20	80	100	0.0695	4.49E-05	7.38E-05	0.608090
50	200	250	0.0698	1.66E-05	3.18E-05	0.520922
100	400	500	0.0694	7.85E-06	1.62E-05	0.484991
200	800	1000	0.0697	3.98E-06	7.96E-06	0.500185
500	2000	2500	0.0697	1.58E-06	2.97E-06	0.530887
1000	4000	5000	0.0697	7.89E-07	1.46E-06	0.541909
2000	8000	10000	0.0697	3.93E-07	8.47E-07	0.463664
3000	12000	15000	0.0697	2.60E-07	5.03E-07	0.517252
5000	20000	25000	0.0697	1.56E-07	2.87E-07	0.543668

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population A(1)=.06966

Table 3.18a
Weighted Approximations of Inequality Index Variances under Stratified Sampling:
Simulation results for Atkinson's Measure: A(2)

Sample sizes St.1	St.2	Total	$A_w(2)$	$Var(A_w(2))$	$Var_{sim}(A_w(2))$	$Var(A_w(2))/Var_{sim}(A_w(2))$
5	10	15	0.1386	0.00104221	0.00146726	0.710310
10	20	30	0.1384	0.00051566	0.0007668	0.672476
20	40	60	0.1401	0.00026928	0.00036336	0.741071
50	100	150	0.1408	0.00011316	0.00015373	0.736078
100	200	300	0.1402	5.36E-05	8.00E-05	0.670044
200	400	600	0.1405	2.72E-05	3.71E-05	0.733294
500	1000	1500	0.1408	1.10E-05	1.62E-05	0.678902
1000	2000	3000	0.1407	5.59E-06	7.80E-06	0.717029
2000	4000	6000	0.1407	2.78E-06	3.75E-06	0.743044
3000	6000	9000	0.1408	1.89E-06	2.82E-06	0.672475
5000	10000	15000	0.1408	1.12E-06	1.56E-06	0.719887

Sample sizes St.1	St.2	Total	$A_w(2)$	$Var(A_w(2))$	$Var_{sim}(A_w(2))$	$Var(A_w(2))/Var_{sim}(A_w(2))$
5	15	20	0.1374	0.00067612	0.00116156	0.582077
10	30	40	0.1383	0.00036387	0.0007006	0.519364
20	60	80	0.1401	0.00019704	0.00034702	0.567808
50	150	200	0.1404	8.02E-05	0.00015993	0.501703
100	300	400	0.1401	4.15E-05	7.54E-05	0.551046
200	600	800	0.1407	2.07E-05	3.58E-05	0.578573
500	1500	2000	0.1409	8.33E-06	1.62E-05	0.515335
1000	3000	4000	0.1407	4.18E-06	7.37E-06	0.567648
2000	6000	8000	0.1409	2.10E-06	3.74E-06	0.560634
3000	9000	12000	0.1408	1.40E-06	2.51E-06	0.558526
5000	15000	20000	0.1407	8.34E-07	1.53E-06	0.546749

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population A(2)=.14075

Table 3.18b
Weighted Approximations of Inequality Index Variances under Stratified Sampling:
Simulation results for Atkinson's Measure: A(2)

Sample sizes						
St.1	St.2	Total	A_w (2)	Var(A_w (2))	$Var_{sim}(A_w$ (2))	Var(A_w (2))/ $Var_{sim}(A_w$ (2))
5	20	25	0.1373	0.00049827	0.00127377	0.391180
10	40	50	0.1377	0.0002651	0.00065137	0.406990
20	80	100	0.1401	0.00015798	0.00034791	0.454092
50	200	250	0.1408	6.43E-05	0.0001486	0.432374
100	400	500	0.1400	3.15E-05	7.63E-05	0.412339
200	800	1000	0.1409	1.72E-05	3.97E-05	0.431634
500	2000	2500	0.1409	6.68E-06	1.45E-05	0.461660
1000	4000	5000	0.1409	3.40E-06	7.24E-06	0.470090
2000	8000	10000	0.1408	1.70E-06	4.13E-06	0.412113
3000	12000	15000	0.1407	1.13E-06	2.51E-06	0.448263
5000	20000	25000	0.1408	6.74E-07	1.40E-06	0.482495

Two strata
$\sigma_1 = \sigma_2 = 20$
$\mu_1 = 100$; $\mu_2 = 200$
Stratum 2 (the less unequal stratum) is over-sampled
Population A(2)=.14075

3.5.2 Example: Kenya

Below we give a simple example of the large changes in estimated inequality when the sampling structure is ignored using an example of urban income data from Kenya.[13]

This sample is stratified by geographical region, and then sub-stratified within regions. Though information on the exact nature of the stratification scheme is not available, we are given weights which will inflate the sample to population totals. Using this information, we can calculate the five measures of inequality for this data, comparing the inequality measures calculated with and without weights. The upper and lower bounds are the approximate, asymptotically normal 95% confidence bounds given by the weighted process described above, which are based upon the assumption of sampling with replacement from an infinite population. (In other words, we have not used any finite population correction.) We saw the performance of this method of calculating standard errors in the simulation above. Since the sampling proportions are not extremely different between observations, we have some confidence that the weighted estimates of the variances will give a reasonable approximation. The 95% upper and lower bounds are calculated as the estimated inequality measure $\pm 1.96\sqrt{est.\ variance}$.

Up to this point we have abstracted from any problems of defining household as opposed to individual income. The survey of Kenyan urban households includes mar-

[13]This data is originally from the Central Bureau of Statistics and the Ministry of National Planning and Development of Kenya. I am grateful to Mwangi wa Githinji for making this data available.

ket activities as well as imputed income from non-market and household production. It lumps all of a household's income in one group, without attempting to allocate income (or consumption) to individuals within the household. To measure inequality, three options present themselves. The first is to simply take the household as the unit of analysis and estimate income inequality across households. The problem with this approach is that we do not take into account household size. One alternate option is to simply divide household income by the number of individuals in the household, use the household inflation weights, and calculate inequality across per-capita household income. Another way to account for household size is to use some type of household equivalence scale. Because of economies of scale in cooking and lodging, it may take less per person income to support four people than one, for example. Also, household with many members may include young children or elderly family members whose consumption needs are smaller than adults. Since we do not have information on the composition of the household, a simple approximation may be to count the first two members of the household as full units and count additional members as fractional units. Below we present results using a household equivalent of .8 and one of .6 for each member greater than two in a household. This is admittedly a very rough approximation, but may give a better indication of welfare and inequality. A final alternative is to divide the income of the household across all individuals in the household and then calculate inequality over individuals. In as much as poorer households are larger, this will give the appearance of greater inequality. It may also not

be representative of welfare for the same reasons that equivalence scales are useful.

Tables 3.19 through 3.23 present these five different ways of attributing income to the members of the household for the five inequality measures considered. In all of the tables, we compare the unweighted and the weighted calculation of inequality under the various definitions of income and highlight the change in measured inequality for the weighted calculation. The first thing to notice is that the different inequality measures are affected differently by the weighting. For per-capita household income (Table 3.20), for example, the coefficient of variation and Theil's second measure (I(1)) decrease sixteen and eighteen percent, respectively.

However, Atkinson's measure with $\epsilon = 2$, ($A(2)$), hardly shows any change. In this case, the weighting doesn't seem to matter much for this choice of ϵ. However, if we set $\epsilon = 1$, the resulting change in the measure when we account for the sample structure is similar to Theil's measure, I(0). As ϵ increases, the social welfare function implied by Atkinson's measure become Rawlsian, and thus the only incomes that matter are those at the bottom of the distribution. In this sample, there are several very large incomes which have relatively small weights, and $\widehat{A(2)}$ is discounting these incomes for both the weighted and unweighted case. These incomes at the top end of the distribution are having a disproportionately large effect on the other measures, an effect that is being corrected when we weight the measures.

We also notice the effect of the various ways of considering household income. As expected, dividing up all of the income equally amongst individuals in the household

and then measuring either per-capita household income or inequality over individual incomes gives the highest values for inequality. This probably overstates inequality (at least our intuitive understanding of inequality) because it does not consider the economies of scale in consumption for larger households, nor the smaller needs of young children. In Table 3.19, we consider household income without adjusting for household size at all. This gives the lowest values for income inequality, not surprising if larger households tend to be poorer. These measures probably over-state inequality for that reason. Tables 3.21 and 3.22 present two different versions of the crude equivalence scale. Despite the naive application of this idea, these probably give the statement of inequality which most matches out intuitive understanding of inequality.

As we have seen here, ignoring the stratification in the data leads to biased estimation of inequality. As in the mean case, it is straightforward to correct our inequality indices using a scheme which assigns weights that are inversely proportional to sample inclusion probabilities. Likewise, we can calculate standard errors which are adjusted for the weighting scheme. These estimated variances will tend to give an underestimate of the true variance of the inequality measure. However, this effect is minor provided that the sampling disproportions are not too large. Full information about the sampling scheme is needed to calculate unbiased standard errors. It may also be possible to conduct a weighted bootstrap calculation of standard errors which would be more accurate than the standard errors of the simulation above. This section thus provides several interesting areas of future research.

Table 3.19
Inequality in the Kenyan Urban Income Distribution:
Household Income

	lower 95% bound	Estimate	lower 95% bound
CV	3.0834	4.3005	5.5177
CV_w	2.1180	3.3924	4.6668
change		**-21%**	
I(0)	0.7491	0.9042	1.0594
$I_w(0)$	0.6741	0.7943	0.9144
change		**-12%**	
I(1)	1.0076	1.3574	1.7072
$I_w(1)$	0.7540	1.0490	1.3440
change		**-23%**	
A(1)	0.5323	0.5951	0.6580
$A_w(1)$	0.4938	0.5481	0.6024
change		**-8%**	
A(2)	0.7387	0.7796	0.8205
$A_w(2)$	0.7335	0.7689	0.8043
change		**-1%**	

Table 3.20
Inequality in the Kenyan Urban Income Distribution:
Per-capita Household Income

	lower 95% bound	Estimate	lower 95% bound
CV	3.8644	4.9094	5.9544
CV_w	2.8146	4.1463	5.4781
change		**-16%**	
I(0)	0.8061	0.9863	1.1665
$I_w(0)$	0.7348	0.8858	1.0368
change		**-10%**	
I(1)	1.1365	1.5221	1.9078
$I_w(1)$	0.8873	1.2459	1.6046
change		**-18%**	
A(1)	0.5598	0.6271	0.6943
$A_w(1)$	0.5253	0.5876	0.6499
change		**-6%**	
A(2)	0.7989	0.8366	0.8743
$A_w(2)$	0.7939	0.8287	0.8635
change		**-1%**	

Table 3.21
Inequality in the Kenyan Urban Income Distribution:
Household Income (Equivalent Scales = .8)

	lower 95% bound	Estimate	lower 95% bound
CV	3.2148	4.5708	5.9269
CV_w	2.1597	3.6568	5.1540
change		**-20%**	
I(0)	0.7234	0.8900	1.0566
$I_w(0)$	0.6517	0.7831	0.9145
change		**-12%**	
I(1)	1.0198	1.3997	1.7795
$I_w(1)$	0.7628	1.0904	1.4180
change		**-22%**	
A(1)	0.5209	0.5893	0.6578
$A_w(1)$	0.4830	0.5430	0.6031
change		**-8%**	
A(2)	0.7497	0.7932	0.8367
$A_w(2)$	0.7451	0.7836	0.8222
change		**-1%**	

Household size is corrected using equivalent scale of .8 for all members of the household after first two. (e.g., a household with 4 people is treated as 3.6 people.)

Table 3.22
Inequality in the Kenyan Urban Income Distribution:
Household Income (Equivalent Scales = .6)

	lower 95% bound	Estimate	lower 95% bound
CV	3.1523	4.5460	5.9397
CV_w	2.0847	3.6079	5.1310
change		**-21%**	
I(0)	0.7137	0.8793	1.0449
$I_w(0)$	0.6417	0.7711	0.9006
change		**-12%**	
I(1)	1.0039	1.3858	1.7677
$I_w(1)$	0.7461	1.0721	1.3981
change		**-23%**	
A(1)	0.5162	0.5849	0.6537
$A_w(1)$	0.4776	0.5375	0.5974
change		**-8%**	
A(2)	0.7406	0.7849	0.8291
$A_w(2)$	0.7356	0.7746	0.8136
change		**-1%**	

Household size is corrected using equivalent scale of .6 for all members of the household after first two. (e.g., a household with 4 people is treated as 3.2 people.)

Table 3.23
Inequality in the Kenyan Urban Income Distribution:
Individual Income

	lower 95% bound	Estimate	lower 95% bound
CV	3.2703	4.9770	6.6836
CV_w	1.9825	4.0168	6.0511
change		**-19%**	
I(0)	0.6873	0.9679	1.2485
$I_w(0)$	0.7100	0.8573	1.0047
change		**-11%**	
I(1)	0.9788	1.4327	1.8866
$I_w(1)$	0.7244	1.1251	1.5257
change		**-21%**	
A(1)	0.5135	0.6201	0.7267
$A_w(1)$	0.5132	0.5757	0.6382
change		**-7%**	
A(2)	0.8046	0.8410	0.8775
$A_w(2)$	0.7965	0.8296	0.8628
change		**-1%**	

3.6 Clustering

As shown in section 2.4 above, when estimating the mean, we know that if we estimate the variance of the estimator $\overline{y}$ using the formula for RSWR, that we need to inflate that estimate by the design effect, d, (equation (68) above) to get an unbiased approximation of the true variance. The same is true for inequality measures, though the design effect in this case will most likely take a different form than d and will be different for different inequality measures. For clarification, we will refer to d in equation (68) as $d_{\overline{y}}$.

As a preliminary exercise, we conduct a simulation to examine the design effect for three of the inequality measures considered above. The purpose of the simulation is to get a rough idea of the impact of clustering on the standard errors of these inequality measures and to compare the relative impact of clustering on estimation of standard errors for mean and inequality measures. Below, we will demonstrate how to develop the exact design effect for the different estimators.

3.6.1 Preliminary simulation results: clustering

The point of the simulation exercise is to demonstrate the effect on inequality measurement of cluster sampling of the type usually found in economic data. We assume the same structure for the data as in section 2.4, equations (63) and (64) above. For the simulation, we set $\mu = 100$ and $\sigma_u^2 = 900$. The cluster size was set equal

to ten, and the cluster effects α were generated from a N(0,σ_α^2) distribution. The idiosyncratic errors, ϵ, were generated from a N(0,σ_ϵ^2) distribution. Based upon the model of (63) and (64), $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$, so by properly choosing σ_α^2 and σ_ϵ^2 such that $\sigma_\alpha^2 + \sigma_\epsilon^2 = \sigma_u^2 = 900$, it is possible to generate different values of ρ.

Here we report the results for a total sample size of 100: 10 clusters with 10 elements in each cluster. The sample is generated by choosing a draw for σ_α^2 for each cluster and a draw of σ_u^2 for each element. We then estimate ρ, the design effect for the mean case, d, and the $\widehat{Var}(\widehat{CV})$ using Kakwani's asymptotic method. This is compared to the variance in simulation over 1000 repetitions of the experiment. We are then able to calculate a simulated design-effect, d_{sim} for each inequality measure by dividing the simulated approximation to the true variance by the value calculated using the asymptotic method (which is based upon independent draws of the sample.) Average results over the 1000 repetitions are presented in Table 3.24 below.

The estimated design effect for the mean is given by $\hat{d}$ (first ρ is estimated using (70) and this value is then plugged into the expression for the design effect (67).) The design effect in simulation for the particular inequality measure is given by $d_{sim,(.)}$. The design effects for the inequality measures in all cases are less than for the mean case. However, they are much greater than one. This indicates that asymptotically normal approximates to inequality measure confidence intervals will be strongly under-estimated.

Table 3.24

Preliminary Simulation Results

Inequality Measurement in Clustered Samples

Coefficient of Variation:

$\hat{\rho}$	$\hat{d}$	$\widehat{CV}$	$\widehat{Var}(\widehat{CV})$	$\text{Var}_{sim}(\widehat{CV})$	$d_{sim,CV}$
.08	1.72	.2975	.00051	.00061	1.20
.17	2.53	.2971	.00050	.00084	1.68
.45	5.05	.2924	.00047	.00177	3.77
.54	5.86	.2889	.00046	.00217	4.72

Theil's Measure:

$\hat{\rho}$	$\hat{d}$	$\widehat{I(1)}$	$\widehat{Var}(\widehat{I(1)})$	$\text{Var}_{sim}(\widehat{I(1)})$	$d_{sim,I(1)}$
.08	1.72	.0466	.000061	.00007	1.15
.17	2.53	.0465	.000060	.00010	1.67
.45	5.05	.0455	.000057	.00021	3.68
.54	5.86	.0447	.000056	.00026	4.64

Atkinson's Measure:

$\hat{\rho}$	$\hat{d}$	$\widehat{A(2)}$	$\widehat{Var}(\widehat{A(2)})$	$\text{Var}_{sim}(\widehat{A(2)})$	$d_{sim,A(2)}$
.08	1.72	.1359	.00336	.00573	1.71
.17	2.53	.1318	.00281	.00568	2.02
.45	5.05	.1235	.00191	.00685	3.59
.54	5.86	.1221	.00198	.00748	3.78

3.6.2 Design effects for inequality measures: coefficient of variation

Just as we are able to calculate a design effect for the case of the mean model, we can do so for inequality measures as well. The variance of our inequality measure under clustered sampling will take the form $Var(\widehat{I}) \cdot d_{\widehat{I}}$. From the simulation above, we expect that $d_{\widehat{I}}$ will be positive, but less than $d_{\overline{y}}$. We will want to consider whether this is a general rule or dependent upon functional forms.

As an example, we will derive the variance for the coefficient of variation under clustered sampling, since we have considered the $Var(\widehat{cv})$ in the random sampling without and with replacement cases. Applying this process for deriving the variances of the other inequality measures considered above in the case of clustered sampling is straightforward.

The sample model is

$$y_{ci} = \mu + u_{ci}, \qquad c = 1, ..., C, \qquad i = 1, ..., M_c \tag{166}$$

where we assume, as in section 2.5 above, that the sampling is random sampling with replacement (or from an infinite population), thus allowing us to ignore any finite population corrections. As is standard in this literature, we assume the that the sample obeys the following

$$\begin{aligned} E\,u_{ci} &= 0,\ E\,u_{ci}^2 = \sigma^2,\ E\,u_{ci}u_{cj} = \rho\sigma^2,\ i \neq j \\ E\,u_{ci}\,u_{c'j} &= 0,\ c \neq c'. \end{aligned} \tag{167}$$

Recall that ρ is the intra-cluster correlation coefficient. Again, we assume that elements within clusters are correlated, but across clusters the elements are independent. The assumptions in (167) are sufficient to determine the size of the design effect in the mean case, however, for the case of inequality measures we will have to consider higher-order correlations as well. Therefore, we will suppose that the following two assumptions also hold

$$\begin{aligned} E\,u_{ci}^2 u_{c'j} &= \sigma_{112,w} \quad , i \neq j \text{ and } c = c' \\ &= 0, \qquad\quad c \neq c' \\ E\,u_{ci}^2 u_{cj}^2 &= \sigma_{1122,w} \quad , i \neq j \text{ and } c = c' \\ &= 0, \qquad\quad c \neq c'. \end{aligned} \tag{168}$$

Following this notation, we can also write $\rho\sigma^2 = \sigma_{12,w}$. Note that $\sigma_{12,w}$, $\sigma_{112,w}$, $\sigma_{1122,w}$ are the same as in equations (10) and (11), but are within-cluster moments only. Here we are implicitly imposing the assumption that $\sigma_{12,w}$, $\sigma_{112,w}$, $\sigma_{1122,w}$ are constant across clusters. This is done both for convenience, as well as for practicality. When we consider sample estimates of these moments, assuming constancy across clusters allows us to pool all of the data to estimate these values. This also mimics the usual statistics literature, where $\sigma_{12,w}$ is assumed constant across clusters.

The total sample size is $n = \sum_c M_c$ and the form of the covariance-variance matrix of the sample data is as in 65 above.

Applying the method used above from (96), the variance of the coefficient of variation will be

$$\begin{aligned} Var(\widehat{CV}) &= Var(\hat{\beta}_2)\left(\frac{\partial CV}{\partial \hat{\beta}_2}\right)^2 + 2Cov(\hat{\beta}_2, \overline{y})\frac{\partial CV}{\partial \hat{\beta}_2}\frac{\partial CV}{\partial \overline{y}} \\ &\quad + Var(\overline{y})\left(\frac{\partial CV}{\partial \overline{y}}\right)^2 + O\left(\frac{1}{n^{3/2}}\right). \end{aligned} \tag{169}$$

where $\hat{\beta}_2 = \frac{1}{n}\sum y_i^2$. Under cluster sampling, we can write

$$\begin{aligned} Var_{clust}(\widehat{CV}) &= Var_{clust}(\hat{\beta}_2)\left(\frac{\partial CV}{\partial \hat{\beta}_2}\right)^2 + 2Cov_{clust}(\hat{\beta}_2, \overline{y})\frac{\partial CV}{\partial \hat{\beta}_2}\frac{\partial CV}{\partial \overline{y}} \\ &\quad + Var_{clust}(\overline{y})\left(\frac{\partial CV}{\partial \overline{y}}\right)^2 + O\left(\frac{1}{n^{3/2}}\right). \end{aligned} \tag{170}$$

where the "clust" subscript refers to the variance of these functions under the assumption of clustered sampling.

As seen in (67), when the sample is clustered, the mean has variance

$$V_{clust}(\overline{y}) = \frac{\sigma_u^2}{n}[1 + (\overline{M} - 1)\rho] \tag{171}$$

where $\overline{M}$ is the sum of cluster sizes weighted by the cluster's proportion of the sample, $\frac{1}{n}\sum_{c=1}^{C} M_c^2$. In Appendix B, we derive

$$\begin{aligned} Var_{clust}(\hat{\beta}_2) &= \frac{1}{n}\left[\left(\gamma_2 + 3 - \overline{M}\right)\sigma^4 + 4\mu\gamma_1\sigma^3 + 4\mu^2\sigma^2\right] \\ &\quad + \frac{4(\overline{M} - 1)}{n}\left[\frac{1}{4}\sigma_{1122,w} + \mu\sigma_{112,w} + \mu^2\sigma_{12,w}\right]. \end{aligned} \tag{172}$$

and

$$Cov_{clust}(\hat{\beta}_2, \overline{y}) = \frac{\gamma_1\sigma^3}{n} + \frac{2\mu\sigma^2}{n} + \frac{\overline{M} - 1}{n}\left[2\mu\sigma_{12,w} + \sigma_{112,w}\right]. \tag{173}$$

Combining the results in (98), (99), (170), (171), (172) and (173) gives the variance of the coefficient of variation under clustered sampling

$$\begin{aligned} Var_{clust}(\widehat{CV}) &= \frac{1}{n}\left\{\left(\gamma_2+3-\overline{M}\right)\frac{\sigma^2}{4\mu^2}-\frac{\sigma^3\gamma_1}{\mu^3}+\frac{\sigma^4}{\mu^4}\right\} \\ &\quad +\frac{\overline{M}-1}{n}\frac{1}{\mu^2}\left\{\frac{1}{4}\frac{\sigma_{1112,w}}{\sigma^2}-\frac{\sigma_{112,w}}{\mu}+\frac{\sigma^2\sigma_{12,w}}{\mu^2}\right\}. \end{aligned} \tag{174}$$

We can also write this as

$$\begin{aligned} nVar_{clust}(\widehat{CV}) &= \frac{1}{4\sigma^2\mu^2}\left[\sigma^4\left(\gamma_2+3-\overline{M}\right)+\sigma_{1112,w}\left(\overline{M}-1\right)\right] \\ &\quad -\frac{1}{\mu^3}\left[\sigma^3\gamma_1+\sigma_{112,w}\left(\overline{M}-1\right)\right]+\frac{\sigma^4}{\mu^4}(1+\rho\left(\overline{M}-1\right)). \end{aligned} \tag{175}$$

In practice we can calculate $\rho=\sigma^2\sigma_{12,w}$ using (70),

$$\sigma_{112,w}=\frac{\sum_{c=1}^{C}\sum_{i=1}^{M_c}\sum_{j\neq j}^{M_c}\hat{u}_{cj}^2\,\hat{u}_{cj'}}{n(\overline{M}-1)} \tag{176}$$

and

$$\sigma_{1112,w}=\frac{\sum_{c=1}^{C}\sum_{i=1}^{M_c}\sum_{j\neq j}^{M_c}\hat{u}_{cj}^2\,\hat{u}_{cj'}^2}{n(\overline{M}-1)} \tag{177}$$

It is perhaps easier to see what is going on if we express this as

$$Var_{clust}(\widehat{CV})=Var_{RSWR}(\widehat{CV})+d_{CV}^{*} \tag{178}$$

where

$$d_{CV}^{*}=\frac{\overline{M}-1}{n\mu^2}\left[\frac{1}{4\sigma^2}\left(\sigma_{1122,w}-\sigma^4\right)-\frac{\sigma_{112,w}}{\mu}+\sigma_{12,w}\frac{\sigma^2}{\mu^2}\right]. \tag{179}$$

In practice, it is simple to calculate d_{CV}^{*} by replacing μ with $\overline{y}$, replacing σ^2, $\sigma_{1122,w}$, $\sigma_{112,w}$, $\sigma_{12,w}$ with their sample estimates, and adding this (additive) design effect to the

usual calculation (wrongly assuming RSWR) of the variance of CV. We can see that $\sigma_{12,w}$ will be positive when ρ is positive (which will normally be the case for clustered data) and that $(\sigma_{1122,w} - \sigma^4)$ will be positive. $\sigma_{1122,w}$ reaches its minimum of σ^4 when the error terms within the same cluster are independent. The only indeterminate term is $\sigma_{112,w}$. The design effect will be positive whenever $\frac{1}{4\sigma^2}(\sigma_{1122,w} - \sigma^4) + \sigma_{12,w}\frac{\sigma^2}{\mu^2} > \frac{\sigma_{112,w}}{\mu}$. We would generally expect this to be the case, as seen from the preliminary simulation.

In the mean case, the design effect was expressed as a multiplicative design effect. We also provide the design term in this way by calculating

$$d_{CV} = \frac{Var_{clust}(\widehat{CV})}{Var_{RSWR}(\widehat{CV})}. \tag{180}$$

The result

$$d_{CV} = 1 + \frac{\left(\overline{M} - 1\right)}{\sigma^2} \frac{\mu^2\left(\sigma_{1122,w} - \sigma^4\right) - 4\mu\sigma^2\sigma_{112,w} + 4\sigma^4\sigma_{12,w}}{\mu^2\sigma^2\left(\gamma_2 + 2\right) - 4\mu\sigma^3\gamma_1 + 4\sigma^4} \tag{181}$$

is quite easy to interpret. First of all, if there is only one cluster, then $\overline{M}$ will equal one and the design effect will be one. If the data are independent, then $\sigma_{112,w} = \sigma_{12,w} = 0$ and $\sigma_{1122,w} = \sigma^4$, and again the design effect will be one. This is as we expected.

A parallel approach can be taken to derive design effects for the other inequality measures. Testing the importance of the design effect in calculating standard errors for inequality measurement remains an important project. Unfortunately, as practitioners we often do not have information on which element belongs to which cluster,

making it impossible to calculate $\sigma_{1122,w}$ and $\sigma_{112,w}$, even when information on the size of ρ is available.

By conducting a simulation study of the properties of CV under clustering, we can determine whether or not this method of estimating the design effect works well in practice. With this goal in mind, we conducted a detailed simulation study of the effects of clustered sampling on the coefficient of variation, and compared the estimated variance with the simulated variance, both ignoring and taking into account the clustering of the data. The results are reported in the next section.

3.6.3 Simulation: coefficient of variation under clustered sampling

To conduct the simulation, we write the clustered sample as an error components model

$$y_{ci} = \mu + \alpha_c + \epsilon_{ci} \tag{182}$$

or

$$y_{ci} = \mu + u_{ci}. \tag{183}$$

As before we assume

$$\begin{aligned} E\,u_{ci} &= 0,\ E\,u_{ci}^2 = \sigma^2,\ E\,u_{ci}u_{cj} = \rho\sigma^2,\ i \neq j \\ E\,u_{ci}\,u_{c'j} &= 0,\ c \neq c'; \end{aligned} \tag{184}$$

but now we also assume

$$\begin{aligned} E\,u_{ci}^2 u_{cj} &= \sigma_{112,w}\,,\ i \neq j \\ E\,u_{ci}^2 u_{cj} &= 0,\ c \neq c'; \\ E\,u_{ci}^2 u_{cj}^2 &= \sigma_{1122,w}\,,\ i \neq j \\ E\,u_{ci}^2 u_{cj}^2 &= 0,\ c \neq c'. \end{aligned} \tag{185}$$

Given (182) through (185), the additive design effect for the coefficient of variation will take the form

$$d_{CV}^{*} = \frac{\overline{M}-1}{n\mu^2}\left[\frac{1}{4\sigma_u^2}\left(\gamma_{2,\alpha}+2\sigma_\alpha^4\right) - \frac{\gamma_{1,\alpha}}{\mu} + \sigma_\alpha^2\frac{\sigma_u^2}{\mu^2}\right]. \tag{186}$$

In terms of creating the simulation, note that the skewness and kurtosis of the idiosyncratic error terms, ϵ, will have no effect on the design effect of CV. Thus, the design effect term can be controlled simply by choosing the form of the cluster-specific error terms, α_c, and its skewness, $\gamma_{1,\alpha}$, and kurtosis, $\gamma_{2,\alpha}$. The effect of ρ on the design effect of CV comes through σ_α^2 which will equal $\rho\sigma_u^2$ for the specification in (182). We can re-write (186) exclusively in terms of ρ, $\gamma_{1,\alpha}$, $\gamma_{2,\alpha}$, and $\gamma_{2,\alpha}$

$$d_{CV}^{*} = \frac{\rho\left(\overline{M}-1\right)}{n\mu^2}\left[\frac{1}{4}\rho\sigma_u^2\left(\gamma_{2,\alpha}+2\right) - \rho^{\frac{1}{2}}\sigma_u^3\frac{\gamma_{1,\alpha}}{\mu} + \frac{\sigma_u^4}{\mu^2}\right]. \tag{187}$$

When all of the error terms are normally distributed, this becomes

$$d_{CV,normal}^{*} = \frac{\rho\left(\overline{M}-1\right)}{n\mu^2}\left[\frac{1}{2}\rho\sigma_u^2 + \frac{\sigma_u^4}{\mu^2}\right]. \tag{188}$$

Under normally distributed cluster-specific errors, therefore, there is still a positive design effect for the coefficient of variation, as was seen in the preliminary simulation.

Tables 3.25 through 3.33 present a selection of results from the simulation exercise. The first three tables, Tables 3.25, 3.26 and 3.27 present simulation results for normally distributed error terms. The idiosyncratic error terms, ϵ_{ci}, are distributed $N(0, 900(1-\rho))$ and the cluster-specific error terms, α_c, are distributed $N(0, 900\rho)$. This provides a distribution of the overall error term of $N(0, 900)$, a correlation coefficient of ρ, and a CV of .3, chosen to match that of the preliminary simulation where the design effect was left uncorrected. The sample size is 100, with 10 clusters of 10 elements. The results in the tables are averages from 1000 repetitions of the simulation.

In Table 3.25, we present the estimated $Var(\widehat{CV})$ calculated using the formula of Cramer from equation (107) and correct it for the clustering using (179). We present the corrected and uncorrected variances in columns 5 and 3, respectively. Column 10 shows the ratio of the estimated $Var(\widehat{CV})$ corrected for clustering to the simulated $Var(\widehat{CV})$. In the case of normally distributed errors, the corrected estimate of the variance tends to slightly over-estimate the true variance by 10-80%. Column 9 provides the ratio of the uncorrected estimated variance to the true (simulated) variance. Here we can see that by ignoring the clustering, we will under-estimate the variance of CV by 300% for even moderate values of ρ ($\rho = .4$).

In Table 3.26 we present the estimated $Var(\widehat{CV})$ calculated using equation (106) which is asymptotically equivalent to (107), but will give slightly different results in practice. As we can see, the results are roughly similar. The corrected variance does

tend to over-estimate the actual variance somewhat, however, the damage done by leaving the variance estimated uncorrected for clustering is much worse.

Table 3.27 provides the expected d_{CV} calculated from (186) using the actual parameter values which were used to generate the data: $\gamma_{1,a} = \gamma_{2,a} = 0$, and ρ varying accordingly. In the second column, we provide $\widehat{d}_{CV}$ which is calculated as the design effect based on the averages of $\gamma_{1,a}$, $\gamma_{2,a}$, etc. over the 1000 simulations. As we can see from column 3, these two numbers match up quite closely. $\widehat{d}^{*}_{CV}$ are the design effects which were calculated from each sample of 100, averaged over the 1000 repetitions. The fact that this does not match with $\widehat{d}_{CV}$ very well indicates that in small samples, the estimated values of $\sigma_{1122,w}$, $\sigma_{112,w}$, and ρ are highly variable and biased. This bias is averaged out over 1000 repetitions as we see in column 2. This would indicate that caution should be used in applying the above results in the small-sample case.

Table 3.28 through 3.30 provide the same exercise with the same parameter values for a larger sample size. The average cluster size was kept at 10, but the number of clusters was increased to 100, giving a total sample size of 1000. As can be seen immediately from column 10 of Tables 3.28 and 3.29, the degree of over-estimation of the variance using the cluster-corrected estimate derived in this paper is much smaller than that in Tables 3.25 and 3.26 for the smaller sample size. This is quite encouraging.

The design effect derived in (186) and (188) allow us to explore the impact of

non-normality in combination with clustering for the coefficient of variation and its variance. Recall that income distributions are highly non-normal and positively-skewed. If the underlying model which we have in mind is that of (182), and we believe that the non-normality comes from cluster effects, then we should be concerned about such effects on computation of the variance and confidence intervals for our inequality measures.

It seems quite reasonable that the skewness in the income distributions comes from the cluster effects more than from the idiosyncratic error terms. If we think of clusters as geographical areas or neighborhoods, we can imagine a wealthy neighborhood with high α_c and incomes roughly normally distributed around this high cluster mean. Likewise, for a poor rural community, the cluster mean is low, with incomes roughly normally distributed around that mean. It therefore seems quite reasonable to be concerned about non-normality in the cluster-specific portion of the model.

We undertook an extensive simulation of non-normal error terms and their effect on the standard error of the coefficient of variation. For small sample sizes (200 or less) the proposed design effect works rather poorly, since the components of d^*_{cv} are difficult to measure in small samples and the skewness and kurtosis terms are highly variable. In larger samples, the design effect correction works quite well. One such example is provided in Tables 3.31 to 3.33.

The layout of these tables is as before, with CV=.3 being chosen as the convenient normalization allowing comparison with what has gone before. Here we choose

element-specific errors, ϵ_{ci}, to be normally distributed, but the cluster-specific error terms are log-normally distributed with $E\alpha_c = 0$, $\gamma_{1,\alpha} = 4.2$, and $\gamma_{2,\alpha} = 24.4$. The variance is as before in order to generate an overall average CV of .3

As can be readily seen in the tables, the corrected $\widehat{Var}_{clust}(\widehat{CV})$ matches the simulated variance much better than the uncorrected estimate. In the case of lognormally distributed data, however, the $\widehat{Var}_{clust}(\widehat{CV})$ tends to provide an under-estimate of the true (simulated) variance (unlike the case of normally distributed errors where $\widehat{Var}_{clust}(\widehat{CV})$ provided an over estimate of the simulated variance.)[14]

Overall, the simulation provides quite positive support for the use of the design effect derived above for estimating the variance of the coefficient of variation. Caution should be exercised in small samples, but even in the case of n=100, the corrected $\widehat{Var}_{clust}(\widehat{CV})$, though biased, provided a much better approximation of the true variance than the uncorrected estimate.

[14]This feature of the simulation–that $\widehat{Var}_{clust}(\widehat{CV})$ under-estimated the true variance when the cluster-specific errors were lognormal held true for many different parametrizations of the model. Even changing the skewness and kurtosis dramatically, as well as increasing or decreasing the sample size, provided no change in this area. Under lognormality, some portion of the design effect term, d^*_{CV}, is being systematically under-estimated, just as in the normal case it is being systematically over-estimated. Like in the normality case however, this bias does decrease as $n \to \infty$.

Table 3.25
Estimating Var($\hat{CV}$) (Eq. (107)) under Clustered Sampling: Normal Errors, n=100

ρ	$\hat{\rho}$	$\hat{CV}$	est. Var($\hat{CV}$) equation (107)	est. d_{CV}* eq. (179)	estimated $Var_{clust}(\hat{CV})$	$Var_{sim}(\hat{CV})$	$d_{sim,CV}$	estimated Var($\hat{CV}$) / $Var_{sim}(\hat{CV})$	estimated $Var_{clust}(\hat{CV})$ / $Var_{sim}(\hat{CV})$
.00	0.002	0.3005	0.000525	-0.000004	0.000521	0.000563	0.000038	0.932715	0.924872
.05	0.052	0.2988	0.000512	0.000155	0.000667	0.000608	0.000096	0.842385	1.097459
.10	0.095	0.2981	0.000515	0.000331	0.000846	0.000673	0.000158	0.765620	1.257812
.15	0.150	0.2992	0.000505	0.000567	0.001071	0.000690	0.000185	0.731741	1.553572
.20	0.195	0.2985	0.000506	0.000770	0.001276	0.000797	0.000291	0.635257	1.601208
.25	0.241	0.2963	0.000506	0.001032	0.001538	0.000837	0.000331	0.604587	1.838303
.30	0.298	0.2968	0.000509	0.001287	0.001796	0.001088	0.000580	0.467325	1.650155
.35	0.350	0.2947	0.000490	0.001612	0.002102	0.001252	0.000762	0.391482	1.679190
.40	0.402	0.2949	0.000493	0.001998	0.002491	0.001453	0.000960	0.339457	1.714293
.45	0.442	0.2924	0.000481	0.002145	0.002626	0.001435	0.000954	0.335093	1.829307
.50	0.500	0.2921	0.000478	0.002598	0.003076	0.001855	0.001377	0.257853	1.658209
.55	0.550	0.2895	0.000470	0.002928	0.003398	0.001979	0.001510	0.237296	1.716801
.60	0.612	0.2916	0.000464	0.003391	0.003855	0.002274	0.001809	0.204165	1.695617
.65	0.672	0.2911	0.000455	0.003763	0.004218	0.002459	0.002004	0.185200	1.715348
.70	0.717	0.2863	0.000446	0.004047	0.004493	0.002625	0.002179	0.170005	1.711683
.75	0.786	0.2873	0.000442	0.004456	0.004899	0.003041	0.002599	0.145453	1.610797
.80	0.856	0.2864	0.000430	0.005194	0.005624	0.003619	0.003189	0.118739	1.553893
.85	0.906	0.2826	0.000409	0.005266	0.005675	0.003678	0.003269	0.111198	1.543005
.90	0.969	0.2848	0.000426	0.006421	0.006848	0.004342	0.003916	0.098185	1.576887
.95	1.050	0.2842	0.000413	0.007032	0.007445	0.005309	0.004895	0.077888	1.402462

estimated Var($\hat{CV}$) = $\dfrac{\bar{y}^2\left(\hat{\beta}_4 - \hat{\beta}_2^2\right) - 4\bar{y}\hat{\beta}_2\hat{\beta}_3 + 4\hat{\beta}_2^3}{4n\bar{y}^4\hat{\beta}_2}$ from equation (107)

Table 3.26

Estimating Var($\hat{CV}$) (Eq. (105)) under Clustered Sampling: Normal Errors, n=100

ρ	$\hat{\rho}$	$\hat{CV}$	est. Var($\hat{CV}$) equation (105)	est. d_{CV}* eq. (179)	estimated $Var_{clust}(\hat{CV})$	$Var_{sim}(\hat{CV})$	$d_{sim,CV}$	estimated Var($\hat{CV}$) / $Var_{sim}(\hat{CV})$	estimated $Var_{clust}(\hat{CV})$ / $Var_{sim}(\hat{CV})$
.00	0.002	0.3005	0.000540	-0.000004	0.000536	0.000563	0.000022	0.960150	0.952307
.05	0.052	0.2988	0.000534	0.000155	0.000689	0.000608	0.000074	0.878205	1.133280
.10	0.095	0.2981	0.000548	0.000331	0.000879	0.000673	0.000125	0.814187	1.306379
.15	0.150	0.2992	0.000554	0.000567	0.001120	0.000690	0.000136	0.802900	1.624731
.20	0.195	0.2985	0.000559	0.000770	0.001329	0.000797	0.000238	0.701803	1.667755
.25	0.241	0.2963	0.000568	0.001032	0.001600	0.000837	0.000269	0.678302	1.912018
.30	0.298	0.2968	0.000583	0.001287	0.001870	0.001088	0.000506	0.535308	1.718138
.35	0.350	0.2947	0.000577	0.001612	0.002189	0.001252	0.000674	0.461308	1.749016
.40	0.402	0.2949	0.000602	0.001998	0.002600	0.001453	0.000851	0.414426	1.789262
.45	0.442	0.2924	0.000582	0.002145	0.002727	0.001435	0.000853	0.405450	1.899663
.50	0.500	0.2921	0.000596	0.002598	0.003194	0.001855	0.001259	0.321436	1.721792
.55	0.550	0.2895	0.000590	0.002928	0.003519	0.001979	0.001389	0.298179	1.777684
.60	0.612	0.2916	0.000603	0.003391	0.003994	0.002274	0.001671	0.265034	1.756485
.65	0.672	0.2911	0.000597	0.003763	0.004360	0.002459	0.001862	0.242879	1.773028
.70	0.717	0.2863	0.000597	0.004047	0.004643	0.002625	0.002028	0.227274	1.768953
.75	0.786	0.2873	0.000590	0.004456	0.005046	0.003041	0.002451	0.194030	1.659375
.80	0.856	0.2864	0.000613	0.005194	0.005807	0.003619	0.003006	0.169345	1.604499
.85	0.906	0.2826	0.000571	0.005266	0.005837	0.003678	0.003107	0.155335	1.587142
.90	0.969	0.2848	0.000639	0.006421	0.007060	0.004342	0.003703	0.147169	1.625871
.95	1.050	0.2842	0.000639	0.007032	0.007671	0.005309	0.004670	0.120316	1.444890

$$\text{estimated Var}(\hat{CV}) = \frac{(\hat{CV})^2}{4n}\left[\frac{\hat{\gamma}_2 + 2}{4} - \hat{\gamma}_1(\hat{CV}) + (\hat{CV})^2\right] \quad \text{from equation (105)}$$

Table 3.27
Design Effect for Coefficient of Variation under Clustered Sampling: Normal Errors, n=100

d_{CV}	$\hat{d}_{CV}$	$\hat{d}_{CV}/d_{CV}$	$\hat{d}^*_{CV}$	$\hat{d}^*_{CV}/d_{CV}$
0.000000	0.000001	–	-0.000004	–
0.000047	0.000046	0.99	0.000155	3.327966
0.000113	0.000103	0.91	0.000331	2.918871
0.000200	0.000182	0.91	0.000567	2.828283
0.000308	0.000284	0.92	0.000770	2.501624
0.000435	0.000376	0.86	0.001032	2.370370
0.000583	0.000543	0.93	0.001287	2.206790
0.000751	0.000696	0.93	0.001612	2.145686
0.000940	0.000893	0.95	0.001998	2.126437
0.001148	0.001033	0.90	0.002145	1.868182
0.001377	0.001250	0.91	0.002598	1.886710
0.001626	0.001451	0.89	0.002928	1.800655
0.001895	0.001766	0.93	0.003391	1.789068
0.002185	0.002031	0.93	0.003763	1.722217
0.002495	0.002234	0.90	0.004047	1.622174
0.002825	0.002695	0.95	0.004456	1.577415
0.003175	0.003223	1.01	0.005194	1.635802
0.003546	0.003291	0.93	0.005266	1.485148
0.003937	0.003610	0.92	0.006421	1.631103
0.004348	0.004207	0.97	0.007032	1.617416

d_{CV} is calculated as $d_{CV} = \frac{\bar{M}-1}{n\mu^2}\left[\frac{1}{4\sigma_u^2}(\gamma_{2.\alpha}+2)\sigma_\alpha^4 - \frac{\gamma_{1.\alpha}\sigma_\alpha^3}{\mu} + \rho\frac{\sigma_u^4}{\mu^2}\right]$ using the true parameters from the simulation.

$\hat{d}_{CV}$ is calculated using the estimated averages of the above parameters over the 1000 repetitions of the simulation

$\hat{d}^*_{CV}$ is calculated as in equation (179) for each sample of size 100. The number presented in column four is the average of those estimates over the 1000 simulation repetitions.

Table 3.28
Estimating Var($\hat{CV}$) (Eq. (107)) under Clustered Sampling: Normal Errors, n=1000

ρ	$\hat{\rho}$	$\hat{CV}$	est. Var($\hat{CV}$) equation (107)	est. d_{CV}* eq. (179)	estimated $Var_{clust}(\hat{CV})$	$Var_{sim}(\hat{CV})$	$d_{sim,CV}$	estimated Var($\hat{CV}$) / $Var_{sim}(\hat{CV})$	estimated $Var_{clust}(\hat{CV})$ / $Var_{sim}(\hat{CV})$
.00	0.000	0.300	0.0000526	-0.0000006	0.0000520	0.0000536	0.0000010	0.981	0.970
.05	0.051	0.300	0.0000529	0.0000139	0.0000668	0.0000570	0.0000041	0.929	1.172
.10	0.100	0.300	0.0000531	0.0000293	0.0000824	0.0000608	0.0000077	0.874	1.356
.15	0.149	0.300	0.0000529	0.0000472	0.0001002	0.0000676	0.0000147	0.782	1.481
.20	0.199	0.299	0.0000523	0.0000642	0.0001165	0.0000817	0.0000293	0.641	1.427
.25	0.248	0.299	0.0000525	0.0000855	0.0001379	0.0000946	0.0000421	0.555	1.459
.30	0.298	0.299	0.0000527	0.0001098	0.0001625	0.0001183	0.0000657	0.445	1.373
.35	0.350	0.299	0.0000527	0.0001375	0.0001902	0.0001252	0.0000724	0.421	1.520
.40	0.401	0.300	0.0000528	0.0001689	0.0002217	0.0001475	0.0000947	0.358	1.503
.45	0.449	0.299	0.0000526	0.0001924	0.0002450	0.0001765	0.0001239	0.298	1.388
.50	0.499	0.299	0.0000527	0.0002289	0.0002815	0.0001965	0.0001439	0.268	1.433
.55	0.548	0.299	0.0000522	0.0002574	0.0003096	0.0002170	0.0001647	0.241	1.427
.60	0.600	0.299	0.0000525	0.0002958	0.0003483	0.0002497	0.0001972	0.210	1.395
.65	0.650	0.299	0.0000522	0.0003322	0.0003844	0.0002526	0.0002004	0.207	1.521
.70	0.699	0.298	0.0000517	0.0003636	0.0004153	0.0003016	0.0002499	0.171	1.377
.75	0.752	0.299	0.0000521	0.0004172	0.0004694	0.0003127	0.0002606	0.167	1.501
.80	0.802	0.298	0.0000518	0.0004555	0.0005073	0.0003553	0.0003035	0.146	1.428
.85	0.854	0.298	0.0000519	0.0005020	0.0005540	0.0004105	0.0003585	0.127	1.350
.90	0.906	0.298	0.0000510	0.0005433	0.0005943	0.0004452	0.0003942	0.115	1.335
.95	0.957	0.298	0.0000519	0.0005999	0.0006518	0.0005224	0.0004704	0.099	1.248

$$\text{estimated Var}(\hat{CV}) = \frac{\bar{y}^{-2}\left(\hat{\beta}_4 - \hat{\beta}_2^2\right) - 4\bar{y}\hat{\beta}_2\hat{\beta}_3 + 4\hat{\beta}_2^3}{4n\bar{y}^{-4}\hat{\beta}_2} \quad \text{from equation (107)}$$

Table 3.29
Estimating Var($\hat{CV}$) (Eq. (105)) under Clustered Sampling: Normal Errors, n=1000

ρ	$\hat{\rho}$	$\hat{CV}$	est. Var($\hat{CV}$) equation (105)	est. d_{CV}* eq. (179)	estimated $Var_{clust}(\hat{CV})$	$Var_{sim}(\hat{CV})$	$d_{sim,CV}$	estimated Var($\hat{CV}$) / $Var_{sim}(\hat{CV})$	estimated $Var_{clust}(\hat{CV})$ / $Var_{sim}(\hat{CV})$
.00	0.000	0.299	0.000053	-0.000001	0.000052	0.000054	0.000001	0.983	0.971
.05	0.051	0.300	0.000053	0.000014	0.000067	0.000057	0.000004	0.934	1.177
.10	0.100	0.299	0.000053	0.000029	0.000083	0.000061	0.000007	0.880	1.362
.15	0.149	0.299	0.000053	0.000047	0.000101	0.000068	0.000014	0.791	1.489
.20	0.199	0.299	0.000053	0.000064	0.000117	0.000082	0.000029	0.645	1.431
.25	0.248	0.299	0.000053	0.000085	0.000138	0.000095	0.000042	0.559	1.462
.30	0.298	0.299	0.000053	0.000110	0.000163	0.000118	0.000065	0.451	1.379
.35	0.350	0.299	0.000053	0.000138	0.000191	0.000125	0.000072	0.427	1.526
.40	0.401	0.299	0.000054	0.000169	0.000223	0.000147	0.000093	0.366	1.511
.45	0.449	0.299	0.000053	0.000192	0.000246	0.000177	0.000123	0.302	1.392
.50	0.499	0.299	0.000054	0.000229	0.000283	0.000197	0.000142	0.275	1.440
.55	0.548	0.298	0.000054	0.000257	0.000311	0.000217	0.000163	0.247	1.433
.60	0.600	0.299	0.000054	0.000296	0.000350	0.000250	0.000196	0.216	1.400
.65	0.650	0.298	0.000054	0.000332	0.000386	0.000253	0.000199	0.213	1.528
.70	0.699	0.297	0.000053	0.000364	0.000416	0.000302	0.000249	0.175	1.380
.75	0.752	0.298	0.000054	0.000417	0.000471	0.000313	0.000259	0.173	1.507
.80	0.802	0.298	0.000053	0.000455	0.000509	0.000355	0.000302	0.150	1.432
.85	0.854	0.298	0.000054	0.000502	0.000556	0.000410	0.000357	0.130	1.354
.90	0.906	0.297	0.000053	0.000543	0.000596	0.000445	0.000392	0.119	1.339
.95	0.957	0.297	0.000053	0.000600	0.000653	0.000522	0.000469	0.102	1.251

$$\text{estimated Var}(\hat{CV}) = \frac{(\hat{CV})^2}{4n}\left[\frac{\hat{\gamma}_2 + 2}{4} - \hat{\gamma}_1(\hat{CV}) + (\hat{CV})^2\right] \quad \text{from equation (105)}$$

Table 3.30
Design Effect for Coefficient of Variation under Clustered Sampling
Normal Errors, n=1000

d_{CV}	$\hat{d}_{CV}$	$\hat{d}_{CV}/d_{CV}$	$\hat{d}^*_{CV}$	$\hat{d}^*_{CV}/d_{CV}$
0.00000000	0.00000000	0.000	-0.00000063	0.000
0.00000466	0.00000469	1.007	0.00001387	2.978
0.00001134	0.00001131	0.997	0.00002934	2.587
0.00002005	0.00001970	0.983	0.00004724	2.357
0.00003078	0.00002995	0.973	0.00006419	2.085
0.00004354	0.00004230	0.972	0.00008545	1.963
0.00005832	0.00005621	0.964	0.00010981	1.883
0.00007513	0.00007439	0.990	0.00013750	1.830
0.00009396	0.00009442	1.005	0.00016891	1.798
0.00011482	0.00011129	0.969	0.00019238	1.676
0.00013770	0.00013757	0.999	0.00022887	1.662
0.00016261	0.00015866	0.976	0.00025739	1.583
0.00018954	0.00018576	0.980	0.00029578	1.560
0.00021850	0.00021580	0.988	0.00033219	1.520
0.00024948	0.00024017	0.963	0.00036358	1.457
0.00028249	0.00028038	0.993	0.00041722	1.477
0.00031752	0.00031073	0.979	0.00045547	1.434
0.00035458	0.00034679	0.978	0.00050204	1.416
0.00039366	0.00038104	0.968	0.00054329	1.380
0.00043477	0.00042590	0.980	0.00059985	1.380

d_{CV} is calculated as $d_{CV} = \frac{\overline{M}-1}{n\mu^2}\left[\frac{1}{4\sigma_u^2}(\gamma_{2,\alpha}+2)\sigma_\alpha^4 - \frac{\gamma_{1,\alpha}\sigma_\alpha^3}{\mu} + \rho\frac{\sigma_u^4}{\mu^2}\right]$ using the true parameters from the simulation.

$\hat{d}_{CV}$ is calculated using the estimated averages of the above parameters over the 1000 repetitions of the simulation

$\hat{d}^*_{CV}$ is calculated as in equation (179) for each sample of size 100. The number presented in column four is the average of those estimates over the 1000 simulation repetitions

Table 3.31
Estimating Var($\hat{CV}$) (Eq. (107)) under Clustered Sampling: Lognormal Errors, n=1000

ρ	$\hat{\rho}$	$\hat{CV}$	est. Var($\hat{CV}$) equation (107)	est. d_{CV}* eq. (179)	estimated $Var_{clust}(\hat{CV})$	$Var_{sim}(\hat{CV})$	$d_{sim,cv}$	estimated Var($\hat{CV}$) / $Var_{sim}(\hat{CV})$	estimated $Var_{clust}(\hat{CV})$ / $Var_{sim}(\hat{CV})$
.00	-0.001	0.300	0.0000530	0.0000001	0.0000530	0.0000559	0.0000029	0.947	0.948
.05	0.043	0.300	0.0000625	0.0001065	0.0001690	0.0002173	0.0001548	0.288	0.778
.10	0.081	0.298	0.0000767	0.0002606	0.0003373	0.0004866	0.0004099	0.158	0.693
.15	0.113	0.296	0.0000808	0.0003197	0.0004004	0.0005738	0.0004930	0.141	0.698
.20	0.152	0.294	0.0000964	0.0004906	0.0005869	0.0008719	0.0007755	0.111	0.673
.25	0.198	0.295	0.0001468	0.0009846	0.0011314	0.0018128	0.0016661	0.081	0.624
.30	0.227	0.293	0.0001825	0.0013434	0.0015259	0.0027750	0.0025925	0.066	0.550
.35	0.268	0.293	0.0002348	0.0018613	0.0020961	0.0039664	0.0037316	0.059	0.528
.40	0.294	0.286	0.0001976	0.0015184	0.0017160	0.0040401	0.0038425	0.049	0.425
.45	0.342	0.289	0.0003106	0.0026142	0.0029247	0.0067171	0.0064066	0.046	0.435
.50	0.378	0.281	0.0002428	0.0020151	0.0022579	0.0042351	0.0039923	0.057	0.533
.55	0.426	0.283	0.0003430	0.0029725	0.0033156	0.0072418	0.0068988	0.047	0.458
.60	0.459	0.277	0.0003328	0.0029170	0.0032498	0.0067010	0.0063682	0.050	0.485
.65	0.520	0.281	0.0004471	0.0039891	0.0044362	0.0112522	0.0108051	0.040	0.394
.70	0.560	0.274	0.0004610	0.0040488	0.0045098	0.0143114	0.0138503	0.032	0.315
.75	0.614	0.269	0.0004744	0.0043211	0.0047955	0.0104816	0.0100071	0.045	0.458
.80	0.676	0.268	0.0005434	0.0049042	0.0054476	0.0159266	0.0153832	0.034	0.342
.85	0.742	0.264	0.0005765	0.0053012	0.0058777	0.0153526	0.0147761	0.038	0.383
.90	0.820	0.267	0.0007381	0.0068139	0.0075520	0.0235716	0.0228335	0.031	0.320
.95	0.908	0.258	0.0006978	0.0063257	0.0070235	0.0231002	0.0224023	0.030	0.304

estimated $\widehat{Var}(\hat{CV}) = \dfrac{\bar{y}^2\left(\hat{\beta}_4 - \hat{\beta}_2^2\right) - 4\bar{y}\hat{\beta}_2\hat{\beta}_3 + 4\hat{\beta}_2^3}{4n\bar{y}^4\hat{\beta}_2}$ from equation (107)

Table 3.32
Estimating Var($\hat{CV}$) (Eq. (105)) under Clustered Sampling: Lognormal Errors, n=1000

ρ	$\hat{\rho}$	$\hat{CV}$	est. Var($\hat{CV}$) equation (105)	est. d_{CV}* eq. (179)	estimated $Var_{clust}(\hat{CV})$	$Var_{slm}(\hat{CV})$	$d_{slm,CV}$	estimated Var($\hat{CV}$) / $Var_{slm}(\hat{CV})$	estimated $Var_{clust}(\hat{CV})$ / $Var_{slm}(\hat{CV})$
.00	-0.001	0.300	0.0000532	0.0000001	0.0000532	0.0000559	0.0000027	0.951	0.952
.05	0.043	0.300	0.0000632	0.0001065	0.0001697	0.0002173	0.0001541	0.290	0.781
.10	0.081	0.298	0.0000780	0.0002606	0.0003386	0.0004866	0.0004086	0.160	0.696
.15	0.113	0.296	0.0000824	0.0003197	0.0004021	0.0005738	0.0004913	0.143	0.701
.20	0.152	0.294	0.0000988	0.0004906	0.0005894	0.0008719	0.0007730	0.113	0.676
.25	0.198	0.295	0.0001510	0.0009846	0.0011356	0.0018128	0.0016619	0.083	0.626
.30	0.227	0.293	0.0001881	0.0013434	0.0015315	0.0027750	0.0025869	0.067	0.552
.35	0.268	0.293	0.0002422	0.0018613	0.0021035	0.0039664	0.0037242	0.061	0.530
.40	0.294	0.286	0.0002040	0.0015184	0.0017225	0.0040401	0.0038360	0.050	0.426
.45	0.342	0.289	0.0003208	0.0026142	0.0029350	0.0067171	0.0063963	0.047	0.437
.50	0.378	0.281	0.0002506	0.0020151	0.0022657	0.0042351	0.0039845	0.059	0.535
.55	0.426	0.283	0.0003544	0.0029725	0.0033269	0.0072418	0.0068875	0.048	0.459
.60	0.459	0.277	0.0003435	0.0029170	0.0032605	0.0067010	0.0063575	0.051	0.487
.65	0.520	0.281	0.0004629	0.0039891	0.0044520	0.0112522	0.0107893	0.041	0.396
.70	0.560	0.274	0.0004761	0.0040488	0.0045249	0.0143114	0.0138352	0.033	0.316
.75	0.614	0.269	0.0004883	0.0043211	0.0048094	0.0104816	0.0099933	0.046	0.459
.80	0.676	0.268	0.0005592	0.0049042	0.0054634	0.0159266	0.0153674	0.035	0.343
.85	0.742	0.264	0.0005942	0.0053012	0.0058954	0.0153526	0.0147584	0.038	0.384
.90	0.820	0.267	0.0007620	0.0068139	0.0075759	0.0235716	0.0228096	0.032	0.321
.95	0.908	0.258	0.0007168	0.0063257	0.0070425	0.0231002	0.0223834	0.031	0.305

$$\text{estimated Var}(\hat{CV}) = \frac{(\hat{CV})^2}{4n}\left[\frac{\hat{\gamma}_2 + 2}{4} - \hat{\gamma}_1(\hat{CV}) + (\hat{CV})^2\right] \quad \text{from equation (105)}$$

Table 3.33
Design Effect for Coefficient of Variation under Clustered Sampling
Lognormal Errors, n=1000

ρ	d_{CV}	$\hat{d}_{CV}$	$\hat{d}_{CV}/d_{CV}$	$\hat{d}^*_{CV}$	$\hat{d}^*_{CV}/d_{CV}$
.00	0.0000000	0.0000000		0.0000001	
.05	0.0000053	0.0000057	1.072	0.0001065	19.995
.10	0.0000277	0.0000257	0.928	0.0002606	9.405
.15	0.0000705	0.0000502	0.712	0.0003197	4.533
.20	0.0001350	0.0001032	0.765	0.0004906	3.635
.25	0.0002217	0.0002434	1.098	0.0009846	4.441
.30	0.0003313	0.0003307	0.998	0.0013434	4.055
.35	0.0004640	0.0005170	1.114	0.0018613	4.011
.40	0.0006202	0.0005600	0.903	0.0015184	2.448
.45	0.0007999	0.0009249	1.156	0.0026142	3.268
.50	0.0010035	0.0008459	0.843	0.0020151	2.008
.55	0.0012311	0.0012877	1.046	0.0029725	2.415
.60	0.0014827	0.0013189	0.890	0.0029170	1.967
.65	0.0017585	0.0020378	1.159	0.0039891	2.268
.70	0.0020586	0.0023353	1.134	0.0040488	1.967
.75	0.0023831	0.0022836	0.958	0.0043211	1.813
.80	0.0027321	0.0030115	1.102	0.0049042	1.795
.85	0.0031056	0.0032662	1.052	0.0053012	1.707
.90	0.0035037	0.0042230	1.205	0.0068139	1.945
.95	0.0039265	0.0044333	1.129	0.0063257	1.611

d_{CV} is calculated as $d_{CV} = \frac{\overline{M}-1}{n\mu^2}\left[\frac{1}{4\sigma_u^2}\left(\gamma_{2,a}+2\right)\sigma_a^4 - \frac{\gamma_{1,a}\sigma_a^3}{\mu} + \rho\frac{\sigma_u^4}{\mu^2}\right]$ using the true parameters from the simulation.

$\hat{d}_{CV}$ is calculated using the estimated averages of the above parameters over the 1000 repetitions of the simulation

$\hat{d}^*_{CV}$ is calculated as in equation (179) for each sample of size 100. The number presented in column four is the average of those estimates over the 1000 simulation repetitions

3.7 Stratification and Clustering: Inequality Measurement from Complex Samples

3.7.1 Example: Mexico

Now we consider an example where both stratification and clustering are present. We correct for the stratification by weighting the inequality estimator as discussed in section 3.5. We will use the simulated design effect to correct for the standard errors. The data are 1989 household expenditure data provided by the Mexican National Statistics Bureau. For this sample, it is known that $\rho = .3$ and that the average cluster size is 12. Information about the exact nature of the clustering, however, is not available because of confidentiality reasons.

In Table 3.34, 3.35 and 3.36, we present the weighted and unweighted estimators of the three inequality measures considered. Again, we provide five different methods of attributing household income among the members of the household. One thing we notice is that the degree of bias from the sample design is much less than in the case of the Kenyan data which we considered in section 3.5.2. These tables present standard errors which are not corrected for the clustering in the data.

Given the information from the simulation, we can expect that for the CV, Theil's measure and Atkinson's measure that the confidence intervals will be approximately 3 times too small for this data since they are obtained by ignoring the correlation induced by the sampling scheme. The confidence intervals presented in Tables 3.34 through 3.36 should thus be treated with caution. The estimated variance of the

coefficient of variation is .0367426. Based upon the simulation which we conducted in section 3.6.3, the true variance may be from 3 times larger, if we assume normally distributed error terms, to 20 times larger if we assume even a moderate degree of skewness and kurtosis! That would imply a confidence interval for the CV of either (1.1874,2.4888) or (the more likely scenario given that the income distribution does exhibit skewness and kurtosis), a confidence interval of (0,3.5183). Inference from these two different estimates of the variance will provide quite different answers than using the confidence intervals reported in the tables.

As Deaton (1997) shows, correlation coefficients between .4 and .6 are not at all uncommon in cross-sectional data from developing countries, implying an even greater design effect in many income surveys.

As in the mean case, failure to correctly specify and correct for the sampling design leads to highly misleading estimates and inference. Estimates of inequality measures can be biased upward or downward by 20% (or more) and standard errors for inequality measures will be many times too small in most cases. The above example from Mexico combined with the simulation suggests that confidence intervals need to be increased 300% to correctly conduct hypothesis tests in this case! In general, the estimated asymptotic confidence intervals for inequality measures may be even more misleading as clustering increases.

Table 3.34
Inequality in Mexico:
(a) Household Income

	lower 95% bound	Estimate	lower 95% bound
CV	1.2233	1.3861	1.5488
CV_w	1.2631	1.4242	1.5853
change		**2.75%**	
I(1)	0.4077	0.4436	0.4794
$I_w(1)$	0.4200	0.4568	0.4936
change		**2.99%**	
A(2)	0.5139	0.5261	0.5384
$A_w(2)$	0.5177	0.5301	0.5425
change		**0.75%**	

(b) Per-capita Household Income

	lower 95% bound	Estimate	lower 95% bound
CV	1.4625	1.8420	2.2215
CV_w	1.4623	1.8381	2.2138
change		**-0.21%**	
I(1)	0.5156	0.5796	0.6435
$I_w(1)$	0.5347	0.5964	0.6580
change		**2.89%**	
A(2)	0.5755	0.5894	0.6034
$A_w(2)$	0.5852	0.5986	0.6120
change		**1.55%**	

Table 3.35
Inequality in Mexico:
Household Income (Equivalent Scales = .8)

	lower 95% bound	Estimate	lower 95% bound
CV	1.3060	1.6622	2.0184
CV_w	1.3445	1.7177	2.0910
change		**3.34%**	
I(1)	0.4621	0.5177	0.5732
$I_w(1)$	0.4832	0.5414	0.5995
change		**4.57%**	
A(2)	0.5455	0.5590	0.5724
$A_w(2)$	0.5562	0.5696	0.5830
change		**1.91%**	

Household size is corrected using equivalent scale of .8 for all members of the household after first two. (e.g., a household with 4 people is treated as 3.6 people.)

Household Income (Equivalent Scales = .6)

	lower 95% bound	Estimate	lower 95% bound
CV	1.2858	1.5936	1.9013
CV_w	1.3228	1.6437	1.9646
change		**3.14%**	
I(1)	0.4461	0.4970	0.5478
$I_w(1)$	0.4663	0.5192	0.5721
change		**4.48%**	
A(2)	0.5318	0.5451	0.5583
$A_w(2)$	0.5419	0.5551	0.5682
change		**1.83%**	

Household size is corrected using equivalent scale of .6 for all members of the household after first two. (e.g., a household with 4 people is treated as 3.2 people.)

Table 3.36
Inequality in Mexico:
Individual Income

	lower 95% bound	Estimate	lower 95% bound
CV	1.2700	1.6852	2.1003
CV_w	1.3109	1.7341	2.1573
change		**2.91%**	
I(1)	0.4722	0.5295	0.5867
$I_w(1)$	0.4982	0.5563	0.6145
change		**5.07%**	
A(2)	0.5522	0.5657	0.5792
$A_w(2)$	0.5609	0.5743	0.5876
change		**1.51%**	

3.8 Conclusion

The measurement and estimation of inequality is an important task for economists and one which has widespread usefulness as a tool of policy evaluation and analysis. Much attention has been paid in the literature to deriving inequality measures which obey axioms corresponding to our intuitive understanding of inequality. This is an important exercise and one which has been widely undertaken.

Another important exercise, and one which has quite unfortunately been almost totally ignored, is the calculation of standard errors and confidence intervals for inequality measures. Since most inequality research involves comparisons over time or across regions, statistical testing of changes and their significance would seem to be of utmost importance. Furthermore, since most surveys of income and expenditure follow the stratified and clustered models outlined in this paper, it seems obvious that estimation and inference should be conducted on inequality measures taking into account the structure of the survey sample. This is something which the literature has ignored quite completely.

Intuitively, results from the mean case regarding bias under stratification and misleadingly small standard errors under clustering should apply in the case of inequality measurement. In this section, we have shown that that intuition is quite correct. Inequality indices will be biased if stratification is ignored. Standard errors must be

corrected if proper inference is to be conducted. In this section, we have shown how to do both. We have derived a design effect for the coefficient of variation which can be estimated quite easily from sample values, something previously unavailable in the literature. Through simulation we have shown the potential of the introduced methods to correct problems arriving from survey data.

We have also provided some previously unknown results about small-sample bias in inequality measurement and linked the shape of the income distribution (specifically the skewness and kurtosis) to both the bias in small samples and the design effect for estimating unbiased variances in the case of clustered sampling. These results are particularly important given the high degree of non-normality usually found in income and expenditure data.

4 NON-PARAMETRIC DENSITY ESTIMATION

The problem of estimating the density of a random variable given a sample of data has long been of interest to economists. When examining variables such as income distribution, density estimation allows the comparison of distributions across regions and over time. Density estimation can also be a useful descriptive tool to give a visual picture of data. Other applications include mapping the distribution of environmental variables like river flow or rainfall, or applications in medical and health-related areas such as treatment lengths hospital stay. For an example applied to the changing distribution of country incomes, see Jones (1997).

Traditionally, density estimation involved the specification of a particular functional form combined with estimates of parameters through maximum likelihood or some other method. For example, if an underlying distribution is known to be normal, it suffices to estimate the mean and standard deviation by usual methods to fully characterize the density. Log-normal, exponential, Pareto, and Gamma distributions have all been used as parametric specifications for examining the distribution of income. As Kakwani (1980) and others have shown, such parameterizations may match one particular distribution of income at a given time, but no single known distribution can characterize all income distributions or even the distribution from one country over time.

The other great drawback of parametric methods is that they require knowledge of the true, underlying density. This is rarely, if ever, known. In examining income distribution changes over time, different parametric densities may show different changes in distribution, thus making analysis sensitive to the parametrization chosen. Inaccurate specification of the distribution can also lead to very misleading results for index inequality measurement or stochastic dominance testing.

Non-parametric density estimation, on the other hand, allows for estimation of univariate and multi-variate densities without the imposition of particular distributional forms. The histogram is a popular non-parametric method of analyzing densities. However, it is not a smooth representation of data (which in addition to being aesthetically desirable is important when representing continuous distributions) and its many other mathematical disadvantages have lead to an extensive literature on smoothed, non-parametric estimates of density beginning with the important contributions of Rosenblatt (1956) and Parzen (1962). For excellent summaries of the non-parametric literature and subsequent developments, see Silverman (1986), Hardle (1990), Pagan and Ullah (1997).

Many different smoothing methods have been considered in the non-parametric literature. Perhaps the most commonly-used is the kernel method of estimating density non-parametrically. To understand how the kernel method works, consider first

a simple histogram constructed from a sample of data of size n

$$y_i \qquad i = 1, \ldots, n. \tag{189}$$

We first choose some binwidth, h, and then construct bins from 0 to h, from h to $2h$, etc. The histogram density estimate of the density in the range $0 - h$ is then

$$\frac{\sum I(y_i)}{n} \tag{190}$$

where $I(y_i)$ is an indicator function which returns value one when y_i falls within the $0-h$ range and zero otherwise. As mentioned above, this technique for estimating the density has several problems, including discontinuity. Different choices of binwidth can produce dramatically different looking distributions. Also, despite the existence of certain rules of thumb, there is no agreed upon method to choose the number of binwidths or starting and ending points.

One generalization of the histogram method is the naive non-parametric estimator or the local histogram. Instead of creating binwidths to span the entire population, we can think of drawing a local histogram at every point of the data. Again, we will need to pick some range, h. We will count the number of observations from the sample which fall within $\frac{h}{2}$ of the point at which we are estimating the density. If we think of the density at y as being,

$$f(y) = \lim_{h \to 0} \frac{1}{2h} P(y - h < Y < y + h). \tag{191}$$

then we can estimate $f(y)$ by picking a small value for h and estimating

$$\hat{f}_N(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{y - Y_i}{h}\right) \tag{192}$$

where $K(\cdot)$ is a weight function (or indicator function)

$$= \frac{1}{2} \qquad when\ \ Y_i\ falls\ in\ \ [y-h, y+h]$$

$$0 \qquad otherwise.$$

and the subscript N stands for "naive" estimate of $f(y)$. When n becomes very large, the window with, h, should become very small. In that case, this will approach the true density of the variable under consideration.

This naive non-parametric density will give a smoother representation of the data than the histogram, but it still suffers from several problems. Firstly, there is no easy way to choose h except for the rule that it should decrease as n increases. Secondly, since the indicator function is not continuous, the estimate of the density of y will not be continuous. If we are interested in estimating derivatives of $f(y)$ for example, we will not be able to do it using this naive estimator. Thirdly, the graph of the density which may be created using (192) will be rough and may demonstrate spurious variation.

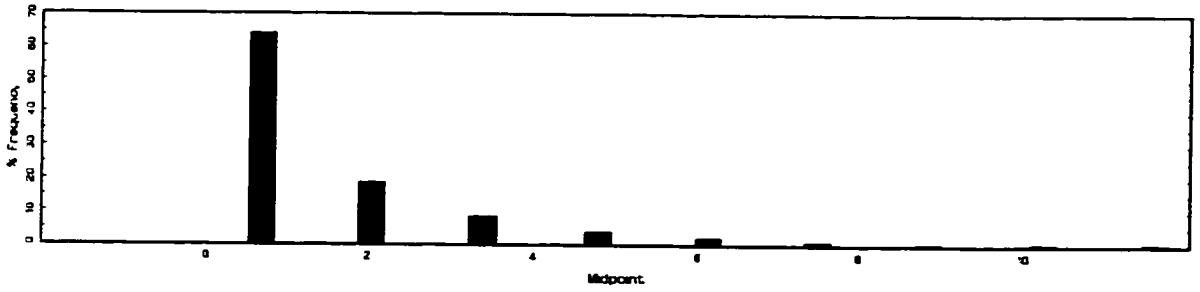
Kernel density estimation is a smooth analog of (192). We replace the indicator function in (192) with a smooth function, called a kernel, which gives small values when the sample value y_i is far from the y for which we are estimating the density.

When y_i is far from the y it will give larger values. A common choice for kernel is the normal kernel

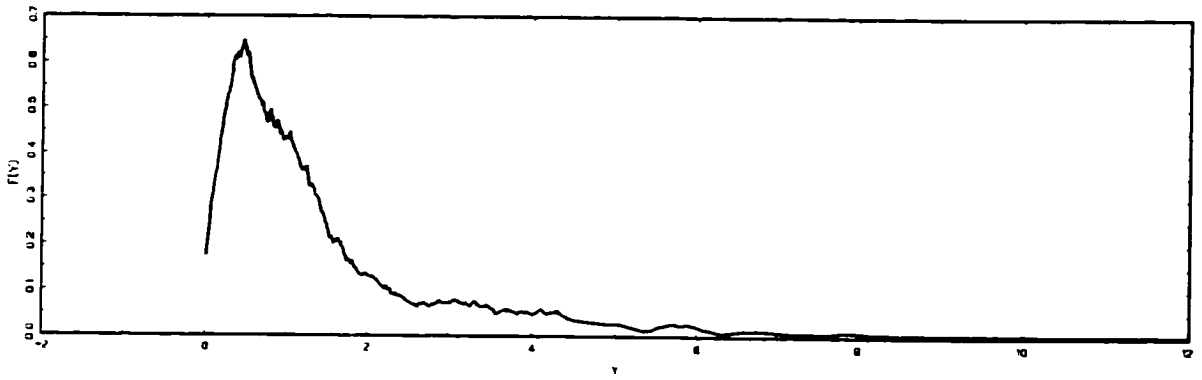
$$K\left(\frac{y-y_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-y_i}{h}\right)} \tag{193}$$

which meets this criteria. Since $K(\cdot)$ is now smooth, we are able to calculate derivatives from this estimate of the density. We can also calculate optimal choices for the kernel and for the window width. To see the improvement in the "picture" of the data created by using the smoother method, see Figure 4.0, where we compare the histogram, the local histogram and the kernel-smoothed density estimates.

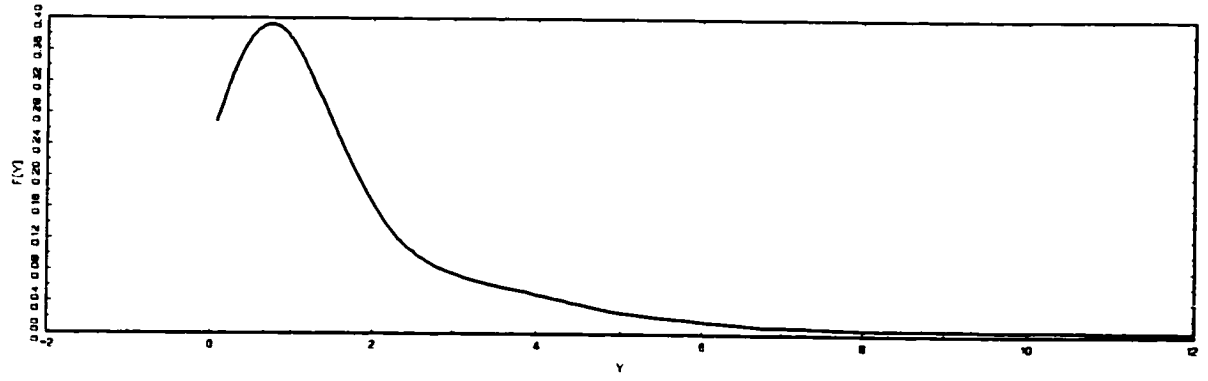
Figure 4.1
Comparison of Three Nonparametric Methods
(a) Histogram



(b) Local Histogram approach



(c) Nonparametric Kernel Density Estimation



The histogram fails to pick up the fact that this data set is not monotonically decreasing. It also gives a discontinuous picture of the data. In the second part of the graph, we see that the naive/local histogram approach is continuous, but is filled with jagged bumps. Also, it is hard to tell whether the apparent increase in the density around $y = 6$ is significant or not. In the bottom part of the graph, we see the smoothed kernel estimate of the density. It is clearly superior aesthetically to the naive/local histogram picture.

The technique of non-parametric density estimation using kernel methods for data which is independently and identically distributed (i.i.d.) is well-developed. Again, Silverman (1986), Hardle (1990) and Pagan and Ullah (1997) provide extensive coverage of this literature. This literature has always assumed (usually implicitly) that the sample data in question has been gathered as a random sample with replacement (RSWR). There has not been any work on density estimation from the kind of stratified and clustered samples which we have seen are quite common in economic analysis. It is also not clear generally how these results are affected when the i.i.d. assumption is violated in other ways, although some work has examined density estimation under weakly dependent time series observations. (See, for example, Hall, Lahiri, and Polzehl (1995) and Herrmann, Gasser, and Kneip (1992).) In this section, we will consider the problems which arise from estimating density from stratified and clustered samples using kernel methods.

In particular, we will consider three deviations from the assumption of indepen-

dently and identically distributed data and simple random sampling with replacement. In section 4.1, we discuss the implications for density estimation of random sampling without replacement (RSWOR) from a finite population. In section 4.2, we consider kernel density estimation from stratified samples with both equal and unequal probability sampling. We will develop the technique of weighted, kernel density estimation to correct for the unequal probability sampling found in most surveys. In section 4.3, we consider a particular deviation from the i.i.d. case–specifically that of non-independence of the data created by clustering of the form frequently found in survey data. For both the case of stratification and the case of clustering, we will consider the problem of choosing the window width, h. Usual techniques for choosing h in the i.i.d./simple random sampling case will no longer give an optimal choice of window width. Below we provide data-based methods for choosing h for both stratified and clustered data. We conclude by indicating how these results may be unified for surveys which combine stratification and clustering.

4.1 Finite Population/Random Sampling Without Replacement

Consider a finite population of size N for some economic variable Y, with population mean μ, population variance σ^2, and probability density f. This will provide a population model like (1) above.

$$Y_i = \mu + U_i \, , \; i = 1, ..., N \tag{194}$$

where the population parameters μ and σ^2 are defined by

$$\mu = \frac{1}{N} \sum_{i=1}^{N} Y_i \, , \qquad \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2 \tag{195}$$

For finite N, $f(Y)$ is often defined as

$$f(Y) = \frac{\sum_{i=1}^{N} I(y_i = Y)}{N} \tag{196}$$

where $I(\cdot)$ is an indicator function giving value 1 when the statement is true and 0 when false. In some cases, particularly when N is large, it may be more useful to define the probability density function of Y as

$$f_h(Y) = \frac{1}{N h_N} \sum_{i=1}^{N} K(\frac{Y_i - Y}{h_N}) = \frac{1}{N} \sum_{i=1}^{N} K_i \tag{197}$$

where $K_i = h_N^{-1} K(\frac{Y_i - Y}{h_N})$. In this case, we write $f(Y)$ with the subscript h to denote the dependence of $f(Y)$ on the choice of h. $f_h(Y)$ is a transformation of the discontinuous, finite probability distribution $f(Y)$ into a continuous distribution. h_N acts as a

smoothing parameter, and as denoted by the subscript N, choice of h should depend upon the size of the finite population. Specifically, h should be inversely related to N. Usual methods for choosing h in the case of density estimation, such as minimizing the integrated mean squared error, will not apply in this case, as $f_h(Y)$ simply defines a continuous version of $f(Y)$. Different h provide different continuous variations on $f(Y)$. In practice, $h = cn^{-\frac{1}{5}}$ with $c = 1.06\sigma$ or $c = .79R$ (where R is the inter-quartile range), may work well.[15]

Now consider a sample of size n drawn without replacement from the finite population, f. (Referred to in the statistics literature as a simple random sample (SRS).) We can write the sample elements as $y_i, i = 1, ..., n$. Let d_i, $i = 1, ..., N$ be the dummy random variables such that $d_i = 1$ if Y_i is selected in the sample and $d_i = 0$ otherwise. Note that the d_i's are not independent and the $E\, d_i^p = n/N$ and $E\, d_i^p\, d_j^q = \frac{n(n-1)}{N(N-1)}$ for any $p > 0$, $q > 0$. Further, we can write the probability of inclusion of any element in the sample as π_i. For SRS, $\pi_i = \pi = \frac{n}{N}$. In the case of Rosenblatt's (1956) kernel estimator, based on y_i, is then

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} k_i = \frac{1}{N\lambda}\sum_{i=1}^{N} d_i K_i \tag{198}$$

where $k_i = h_n^{-1} K(\frac{y_i - y}{h_n})$ and $\lambda = n/N$.

[15]Silverman (1986) demonstrates the optimality of this choice of c for the problem of density estimation of an infinite population when the sample is drawn with replacement. This concept of optimality is not relevant here. These choices are simply suggested as giving an agreeable amount of smoothness.

The finite population density estimation problem reduces to the problem of estimating the population mean discussed in section 2. Provided that h_n satisfies $\lim_{n \to N} h_n = h_N$, it therefore follows from the results there that

$$\begin{aligned} E\hat{f}(y) &\rightarrow f_h(Y) \\ V(\hat{f}(y)) &= (\frac{1}{n} - \frac{1}{N})S_k^2. \end{aligned} \tag{199}$$

where $S_k^2 = (N-1)^{-1}\sum_1^N (K_i - \bar{K})^2$. However, if we treat the finite population Y_i, $i = 1, ..., N$ itself as an i.i.d. random observation from an infinite (super) population with density f^* then as $(n, N) \rightarrow \infty$ such that $\lambda \rightarrow c > 0$,

$$\begin{aligned} E\hat{f}(y) &= (N\lambda)^{-1} N(E\,d_1)(E\frac{1}{h_N} K(\frac{Y_1 - Y}{h_N})) \\ &\rightarrow f^* \end{aligned} \tag{200}$$

and

$$(nh)^{1/2} V(\hat{f}(y)) \rightarrow f^* \int K^2(\psi)\, d\psi. \tag{201}$$

The details of (51) and (52) can be worked out by following Rosenblatt (1956) and Pagan and Ullah (1997).

4.2 Stratified Sampling

Now let us consider density estimation for data chosen under stratified sampling. Consider the following population model,

$$Y_{ij}, \quad i = 1, \ldots, M \quad j = 1, \ldots, N_i. \tag{202}$$

The total number of elements in the population is $\sum N_i = N$ and the proportion of elements in each stratum, i, is $\theta_i = \frac{N_i}{N}$. Within each stratum the data are characterized by some distribution, g_i, with mean μ_i and variance σ_i^2. We will only restrict the strata densities by the requirement that the first two moments exist and are finite for each stratum. We will consider the case where each stratum is an i.i.d. draw from some super-population g_i^*.

Now consider a sample, where n_i elements are drawn from each stratum (i.e. a stratified sample). The total sample size is $\sum n_i = n$. The n_i may or may not be equal. Since both the n_i and the θ_i may vary, the sample inclusion probabilities are no longer equal for all elements in the sample. They will however, be equal for all elements in the same stratum. We will have the probability that the j-th element in the i-th stratum is included in the sample: $\pi_{ij} = \pi_i = \frac{n_i}{N_i}$.

The parameter of interest is the overall distribution of data in the population,

$$f(Y) = \sum \theta_i g_i. \tag{203}$$

The usual Rosenblatt Kernel estimator for the density at a point y is

$$\hat{f}(y) = \frac{1}{nh_n}\sum_{l=1}^{n} K^{*}(\frac{y_l - y}{h_n}) = \frac{1}{N\lambda h_n}\sum_{l=1}^{N} d_l K^{*}_l \tag{204}$$

where h is the window width which is assumed to satisfy

(A1) (i) $h \to 0$

(ii) $nh \to \infty$

as $n \to \infty$.

for sample of size n, and the kernel $K^{*}(\cdot)$ is a symmetric function which satisfies:

(A2) (i) $\int K^{*}(\psi)d\psi = 1$

(ii) $\int \psi K^{*}(\psi)d\psi = 0$

(iii) $\int \psi^2 K^{*}(\psi)d\psi = \gamma_2 < \infty$

We can re-write (204) as

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{M}\sum_{j=1}^{n_i} K^{*}\left(\frac{y_{ij} - y}{h}\right) = \sum_{i=1}^{M}\frac{n_i}{n}\hat{g}_i(y) \tag{205}$$

where $\hat{g}_i(y)$ is the estimate of the density for stratum i at the point y. This estimate of the population density, $f(y)$, is thus a sample-weighted average of the density estimates for each stratum. It is then clear that the density estimate, $\hat{f}(y)$, will only be asymptotically unbiased for (203) when

$$(i) \qquad \frac{n_i}{N_i} = \frac{n}{N}$$

or

$$(ii) \qquad g_i = g \qquad \forall i.$$

These are the same conditions required for unbiased estimation of the mean model from (45) above. These conditions are unlikely to be met in most surveys. It is a common feature of surveys that sampling is disproportionate, violating condition (i). Even when the original survey design is such that the sample inclusion probabilities are equal in all strata, varying rates of non-response and other factors usually force us to re-weight the sample to make it representative. This re-weighting will have the same effect on estimation as having an initial sample design that incorporates unequal probability sampling. Unequal probability sampling is often a desired trait when particular populations of interest are sampled more heavily relative to the rest of the population (SIPP of the US Census, for example) or when cost restricts sampling. (i.e. the case of LSMS data from the World Bank where lower cost of sampling in urban areas leads to higher sampling proportions in these areas.) Though we are interested in an overall estimate of the density, it is problematic to assume that variables of interest will be identically distributed in different strata. Ignoring either this dis-proportionality in the survey design or the differences between strata will lead to biased estimation, even in the simple case of non-parametric kernel density estimation.

The solution is a weighted estimator

$$\hat{f}_w(y) = \frac{1}{h\sum w_i} \sum_{i=1}^{M} \sum_{j=1}^{n_i} w_i K\left(\frac{y_{ij} - y}{h}\right) \tag{206}$$

where $w_i \alpha \frac{N_i}{n_i}$. Again, we set the weights proportional to the inverse of the selection

probabilities. If we further require that $\sum w_i = 1$, then $w_i = \frac{N_i}{N n_i}$. Then

$$\hat{f}_w(y) = \sum_{i=1}^{M} \frac{N_i}{N} \hat{g}_i(y) = \sum_{i=1}^{M} \theta_i \hat{g}_i(y) \tag{207}$$

This takes care of the problem of asymptotic bias. However this is still not unbiased for (203) since $\hat{g}_i(y)$ is not unbiased for g_i. This bias will depend upon the choice of window width, h. We can write

$$\hat{f}_w(y) = \sum_{i=1}^{M} \theta_i (g_i + bias_i) \tag{208}$$

and

$$bias\left(\hat{f}_w(y)\right) = \sum_{i=1}^{M} \theta_i bias_i$$

where the typical bias term upto $O(h^2)$ will be

$$bias_i = \frac{h^2}{2} g_i'' \gamma_2. \tag{209}$$

Assuming that the sampling is independent between strata (which is usually the case), we can also write

$$Var\left(\hat{f}_w(y)\right) = \theta_1^2 var\left(\hat{g}_1\right) + \theta_2^2 var\left(\hat{g}_2\right) + ... + \theta_M^2 var\left(\hat{g}_M\right) \tag{210}$$

and upto $O(\frac{1}{nh})$

$$Var\left(\hat{f}_w(y)\right) = \frac{1}{h}\left[\int \left(K(\psi)\right)^2 d\psi\right] \sum_{i=1}^{M} \frac{\theta_i^2}{n_i}. \tag{211}$$

Silverman (1986) provides details of the non-stratified case for sampling with replacement. If we consider each stratum as such a sample, it is then straightforward to work out (208) through (211).

Proposition 4.1: If the densities of strata 1 through M are given as g_1 through g_M, the population density $f(y)$ is estimated using a kernel density satisfying (A2), and a stratified sample of data is drawn independently in each stratum, then the window width which minimizes the mean-squared error of $\hat{f}(y)$ will be

$$h_{st} = \left(\gamma_2^2\right)^{-\frac{1}{5}} \left(\left[\int_w \left(K^{''}(\psi)\right)^2 d\psi \right] \right)^{\frac{1}{5}} \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i} \right)^{\frac{1}{5}} \left(\int_x \left[\sum_{i=1}^{M} \theta_i \left(g_i^{''}\right) \right]^2 dx \right)^{-\frac{1}{5}}. \qquad (212)$$

Proof: using (208) through (211) write $IMSE\left(\hat{f}_w(y)\right) = Var\left(\hat{f}_w(y)\right) + \left(bias\left(\hat{f}_w(y)\right)\right)^2$ and minimize with respect to h.

Corollary 4.1: If g_1 through g_M are normally distributed with mean μ_M and variance σ_M^2 and the density is estimated using a standard normal kernel, then the optimal window width (in the mean squared error sense) will be

$$h_{st} = 0.87 \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i} \right)^{\frac{1}{5}} \left(\lambda_1 + \lambda_2\right)^{-\frac{1}{5}} \qquad (213)$$

where λ_1 is a weighted sum of stratum-specific standard deviations

$$\lambda_1 = \frac{3}{8} \sum_{i=1}^{M} \theta_i^2 \sigma_i^{-5}$$

and λ_2 is a weighted sum of a function of the distance between stratum means

$$\lambda_2 = \sum_{i=1}^{M} \sum_{l \neq i}^{M} \theta_i \theta_l \frac{\left(\sigma_i^2 + \sigma_l^2\right)^{-\frac{5}{2}}}{\sqrt{2}} \left\{ 3 - 6\frac{\left(\mu_i - \mu_l\right)^2}{\left(\sigma_i^2 + \sigma_l^2\right)} + \frac{\left(\mu_i - \mu_l\right)^4}{\left(\sigma_i^2 + \sigma_l^2\right)^2} \right\} e^{-\frac{1}{2}\frac{\left(\mu_i - \mu_l\right)^2}{\left(\sigma_i^2 + \sigma_l^2\right)}}$$

Proof: For the case of a standard normal kernel $\gamma_2^2 = 1$ and $\int_w (K(\psi))^2 d\psi = \frac{1}{2\sqrt{\pi}}$. We can write

$$\int_y \left[\sum_{i=1}^{M} \theta_i (g_i'')\right]^2 dy = \int_y \sum_{i=1}^{M} \theta_i^2 (g_i'')^2 dy + \int_y \sum_{i=1}^{M} \sum_{l \neq i}^{M} \theta_l \theta_i (g_i'')(g_l'') dy$$

and for normal densities replace g_i'' with $\frac{1}{\sigma_i^5\sqrt{2\pi}}\left[1 - \left(\frac{y-\mu_i}{\sigma_i}\right)^2\right] e^{-\frac{1}{2}\left(\frac{y-\mu_i}{\sigma_i}\right)^2}$. Then the first term, $\int_y \sum_{i=1}^{M} \theta_i (g_i'')^2$, becomes $\frac{3}{8\sqrt{\pi}} \sum_{i=1}^{M} \sigma_i^{-5} \theta_i^2$. The second term can be calculated by integrating the product of g_i'' and g_j''. Using these results, calculate λ_1 and λ_2 and replace in the formula for h_{st}. This proves Corollary 4.1.

We note that the optimal window width is inversely proportional to a weighted sum of the strata sample sizes, n_i. In the case where $n_i = \frac{n}{M}$ and $\theta_i = \frac{1}{M}$, then $\sum_{i=1}^{M} \frac{\theta_i^2}{n_i} = n$ and the window width will be proportional to $n^{-\frac{1}{5}}$ as in the non-stratified case, but the proportionality constant will differ from the usual 1.06σ. When strata share common means and variances, and the population and sample proportions are equal in all strata, this result collapses to the usual optimal window width for normal density: $h^*=1.06\sigma n^{-\frac{1}{5}}$. Note that however if $\sigma_i = \sigma$ for all strata, that the population standard deviation may still differ from σ and $h_{st} \neq 1.06\sigma n^{-\frac{1}{5}}$.

When strata share common means and variances, $h_{st} = 1.06\sigma \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i}\right)^{\frac{1}{5}}$. Recall that the covariance between strata within sample is zero, thus even in the case of homogeneous populations in all strata, the variance of the density estimate is less than in the non-stratified case and the optimal window width is likewise different.

This mimics the case of estimation of the mean under stratified sampling, where even when all strata have identical means, a stratified sample reduces the variance of the estimator, $\overline{y}$.

In practice we can replace σ_i with some consistent estimator like $\sqrt{s_i^2}$ and μ_i with its estimate, $\overline{y}$.

In figure 4.1, we consider the effects on the window width when there are two strata, each distributed normally, with samples selected with proportional sampling. In figure 4.1a, we can see that as the difference between strata means increases, h^* grows without bound. h_{st} on the other hand, increases for a period, but then decreases. Intuitively, the optimal estimator, h_{st}, increases when the combined strata are unimodal, but once the means are far enough apart for the density to exhibit bi-modality, h_{st} begins to decrease. The effect of this is that the density estimation is essentially being conducted separately on each stratum, the small window width giving near zero weight to comparisons between elements in different strata. Figure 4.1b provides the same illustration for strata with identical means, but increasingly different standard deviations. Again, h_{st}, is not going to grow without bound because it takes into account the fact that the increasing sample variation is the result of two strata with two different underlying distributions.

Figure 4.2a
h^* and h_{st} for Two Strata with $\sigma_1=\sigma_2=1$
n=1000

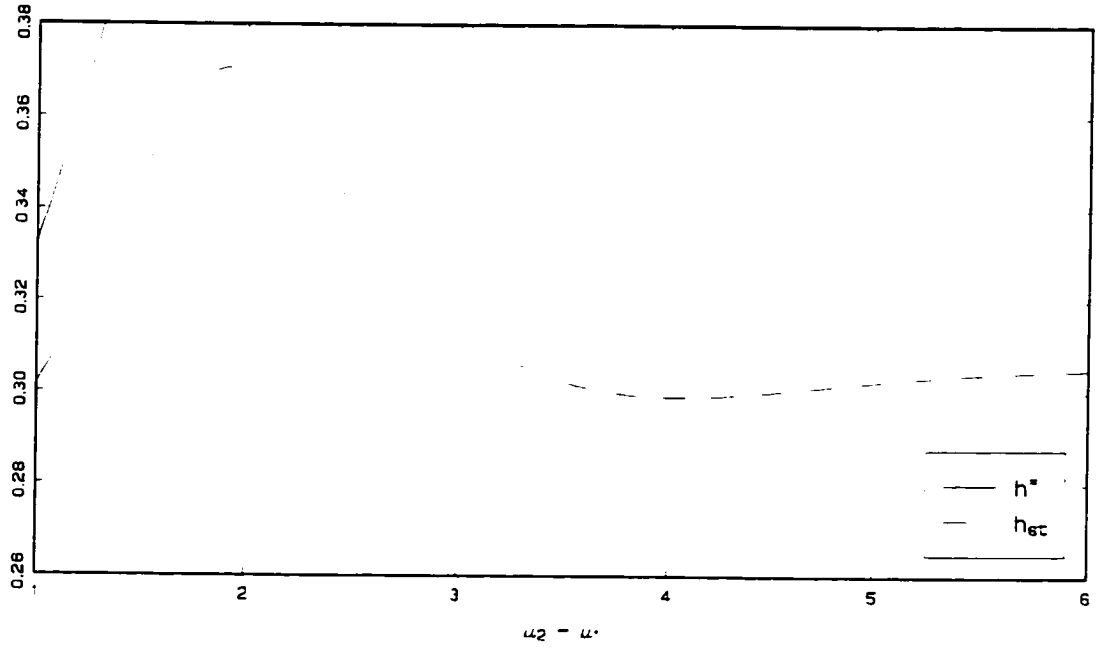
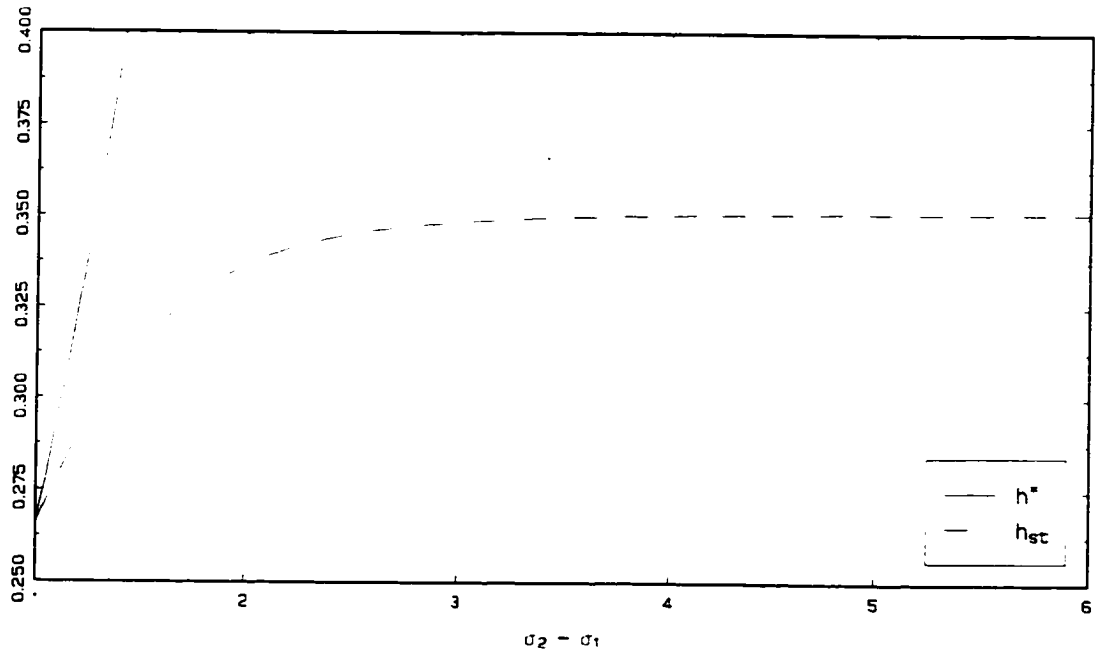


Figure 4.2b
h^* and h_{st} for Two Strata with $\sigma_1=\sigma_2=0$
n=1000



4.2.1 Simulation study: stratified sampling

In a simulation study we consider the proposed optimal window width, h_{st} versus $h^* = 1.06\sigma n^{-\frac{1}{5}}$ and $h_a = .9 * MIN(\sigma, \frac{inter-quartile\ range}{1.34})$. h_a has been shown to be superior to h^* for mixtures of normals and bimodal densities, see Silverman (1986). For clarity, we consider sampling from two strata, where the population in each strata is equal and the underlying densities are normal with mean μ_i and standard deviation σ_i. For the simulation, we fix $n_1 = 1000$, $\mu_1 = 0$ and $\sigma_1 = 1$ while varying the sample size, the mean, and the standard deviation of stratum 2 only. Proportional sampling thus implies $\frac{n_2}{n_1} = 1$, otherwise the sampling is disproportionate.

For proportional sampling, we consider the benchmark case when there is no difference in mean or standard deviation between the two strata. We then consider how estimation changes using the proposed h_{st} as the difference between the two strata means increases, as the difference between the standard deviations increases, and as both change. We then consider the same cases for disproportionate sampling.

We conduct 1000 repetitions for each case. Each repetition involves drawing a sample from the two strata, estimating the three candidate window widths (h^*, h_a, and h_{st}) based upon the formulas given above (with σ replaced by s, and μ replaced by $\bar{y}$), and estimating the non-parametric density at 200 points. The figures give the average estimate of the density over the 1000 repetitions. Table 4.1 presents

the average calculated window widths and the various combinations which have been considered in the simulation exercise.

We first consider the case of proportional sampling ($n_1 = n_2 = 1000$) when both strata have mean zero and variance one. As can be seen in figure 4.2 and from the corresponding row in Table 4.1, the average h_{st} is the same as the average h^* and the two average density estimates are identical. This is as expected given the discussion above. In this case, the stratification is spurious since the two strata are exactly identical. Note that in practice, however, the density estimate using h^* will be superior to that using h_{st} since calculation of h_{st} involves computing two strata means and two strata standard deviations instead of one total sample standard deviation. The estimation of four quantities instead of one introduces more variability into the estimate of h_{st} than h^*.

In figures 4.3 and 4.4, we compare density estimation when both strata have standard deviation equal to one, but have different means. (This gives a population which is a mixture of normals.) When the difference in strata means is such that the overall population density remains unimodal (Figure 4.3), h_{st} performs better than either h^* or h_a. The integrated mean squared error using h_{st} is 27% less than that using h^*. (See Figure 4.11 and Table 4.2 below.) Figure 4.3 presents the average density estimates for the three candidate window widths.

The improvement provided by using h_{st} as opposed to h^* or h_a is dramatic when the difference between the strata means grows and the overall density becomes bi-

modal. As can be seen in Figure 4.4, h^* tends to oversmooth the peaks. h_a gives improved performance and reduces this over-smoothing, but h_{st} can be seen to match the peaks even better than either h^* or h_a.

When means between strata are equal, but variances differ, the same results holds. h_a improves performance over h^*, but h_{st} matches the density better than either. This case is considered in Figure 4.5. Figure 4.6 presents the density estimates when both strata means and strata deviations differ, but sampling is proportional. Using h_{st} provides a much better match of the true underlying density, since it takes into account the different strata-specific distributions.

As noted above, proportional sampling will tend to be the exception in most cross-sectional data sets used by economists. The proposed optimal window width, h_{st} combined with the weighted density estimator of (206), proves to be a very powerful tool for non-proportional sampling. This is examined in the remaining figures.

When the sampling is not proportionate and the strata differ in either means or variances, the unweighted estimator will be biased as discussed above. This is clear from Figures 4.7 and 4.8 where we compare density estimation for two strata with equal standard deviations but different means. In both cases, the weighted estimator using h_{st} clearly outperforms unweighted estimation with any window width. Here stratum 2 is sampled twice as intensively as stratum 1, thus the elements from stratum 2 receive a weight that is half that of elements in stratum 1. This is not a particularly large difference in weights. In many cases, the sampling disproportion is greater than

10 between certain strata, so the results from ignoring the weighting in this case will be even more dramatic with even larger resulting bias.

Figures 4.9 and 4.10 illustrate the case of equal strata means and different variances and the case of variation between strata of both means and standard deviations. Again, the same results hold. Large bias is incurred by ignoring the structure of the sampling.

In Table 4.2, we have calculated the approximate (upto $O(\frac{1}{nh})$) IMSE using h_{st} and h^* with a standard normal kernel and we compare the ratio of these two as a measure of the efficiency loss of using h^*. The top half of the table compares the loss of efficiency for two strata with equal standard deviations as the difference between strata means increases. In the bottom half of the table, the means are held constant while strata standard deviations vary. The results from the table are presented in a more user-friendly format in Figures 4.11 and 4.12.

As can be seen in Figure 4.11, for the case of proportional sampling, when the difference between means is greater than 2, using h^* results in very large efficiency losses compared to using h_{st}. This corresponds to the simulation results presented in Figures 4.3 and 4.4. Comparing Figure 4.12 with Figure 4.11, we see that for the case of dis-proportionate sampling, the relative loss of IMSE is not much different than in the case of proportional sampling. Recall from the simulation, however, that the bias is much greater using h^*. Since the efficiency measure considered here includes integrated bias, it is perhaps not a good measure of the pointwise bias from using h^*.

However, it is quite clear from the figures presented in the simulation exercise that this pointwise bias will be unacceptably large.

In the next section, we consider the effect of clustering on non-parametric density estimation. Some of the same issues which were raised in this section will arise there the usual optimal window width h^* will no longer give best results. However, we will also have to contend with the correlation which is induced in the sample through the method of cluster sampling. It is to that issue that we now turn.

Table 4.1
Weighted Non-Parametric Density Estimation for Stratified Samples:
Results of Simulation Exercise

Average results of 1000 repetitions

Stratum 1 values fix $n_1=1000$
$\sigma_1=1$
$\mu_1=0$

	μ_2	σ_2	n_2	h_{st}	h_a	h^*	Figure	
Proportional sampling:	0	1	1000	0.23159	0.19553	0.23138	2	No difference between strata
	2	1	1000	0.32764	0.27818	0.32071	3	Strata differ only by mean
	3	1	1000	0.41773	0.35468	0.27351	4	Strata differ only by mean
	0	3	1000	0.51800	0.31662	0.30318	5	Strata differ only by standard deviation
	3	3	1000	0.62402	0.49159	0.30584	6	Both means and standard deviations differ
Non-proportional sampling:	2	1	2000	0.29363	0.24931	0.30272	7	Strata differ only by mean
	3	1	2000	0.37006	0.31420	0.25824	8	Strata differ only by mean
	0	3	2000	0.53760	0.36005	0.28605	9	Strata differ only by standard deviation
	3	3	2000	0.61692	0.52379	0.28921	10	Both means and standard deviations differ

Figure 4.3a
Unweighted Estimate Using h^*
Proportional Sampling

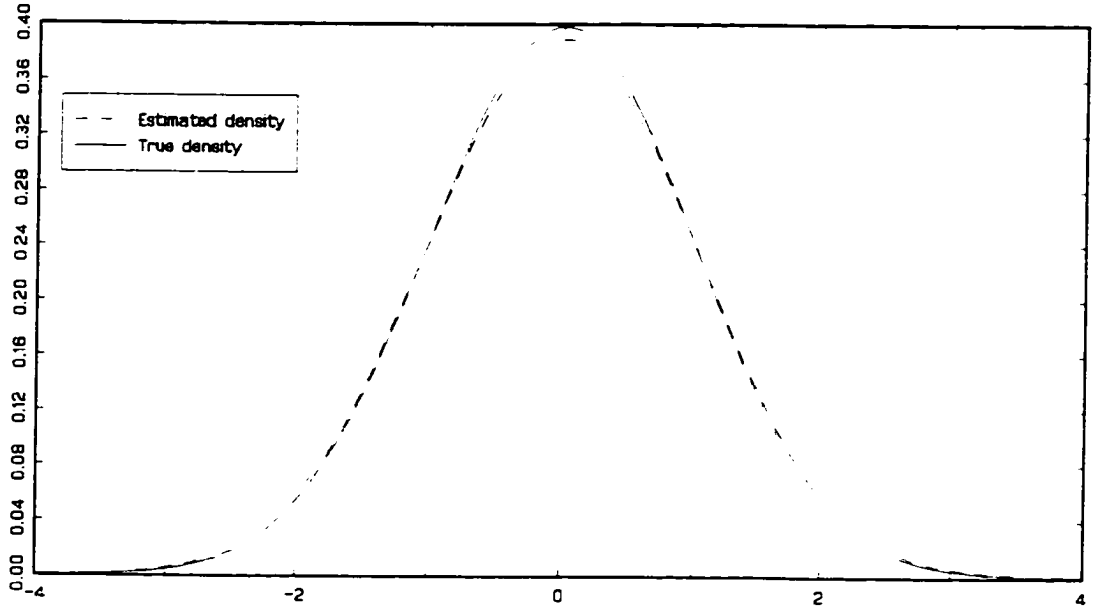
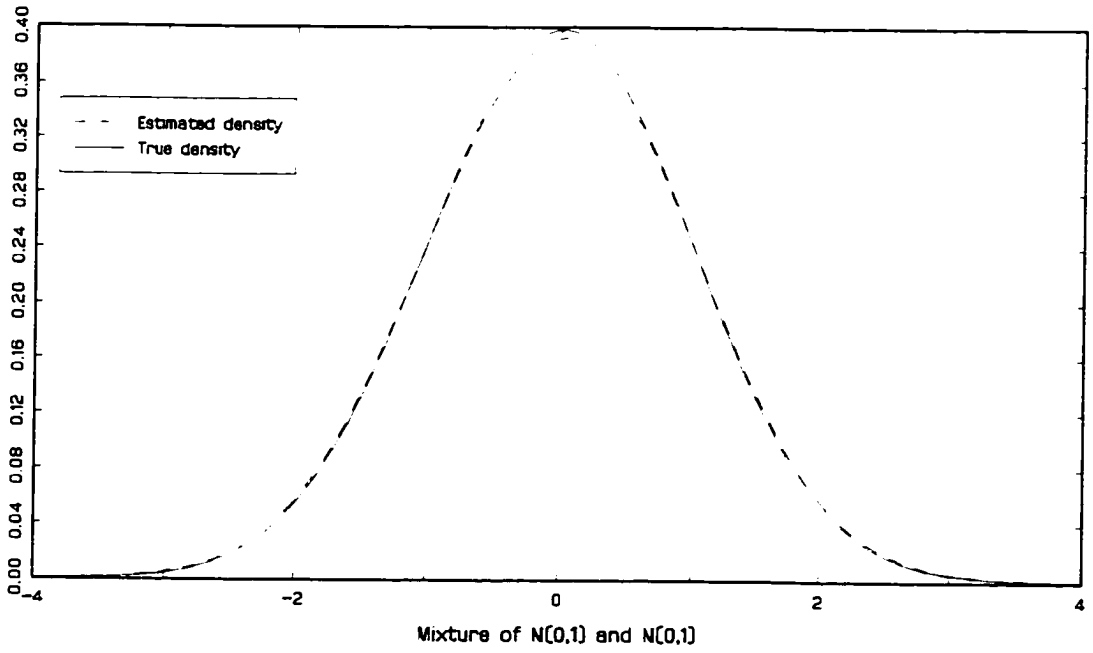


Figure 4.3b
Unweighted Estimate Using h_a
Proportional Sampling



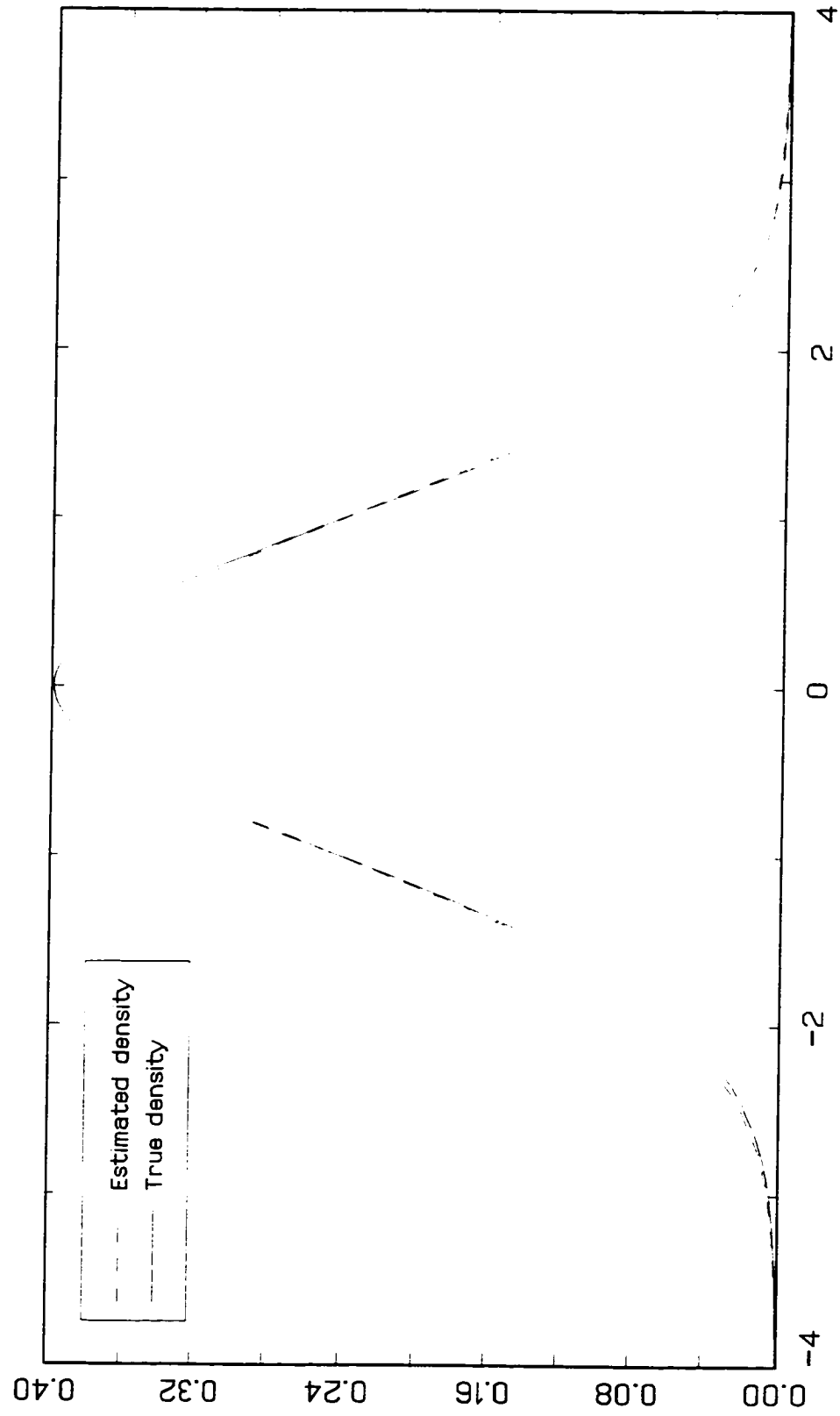


Figure 4.3c
Weighted Estimate Using h_{st}
Proportional Sampling

Figure 4.4a
Unweighted Estimate Using h^*
Proportional Sampling

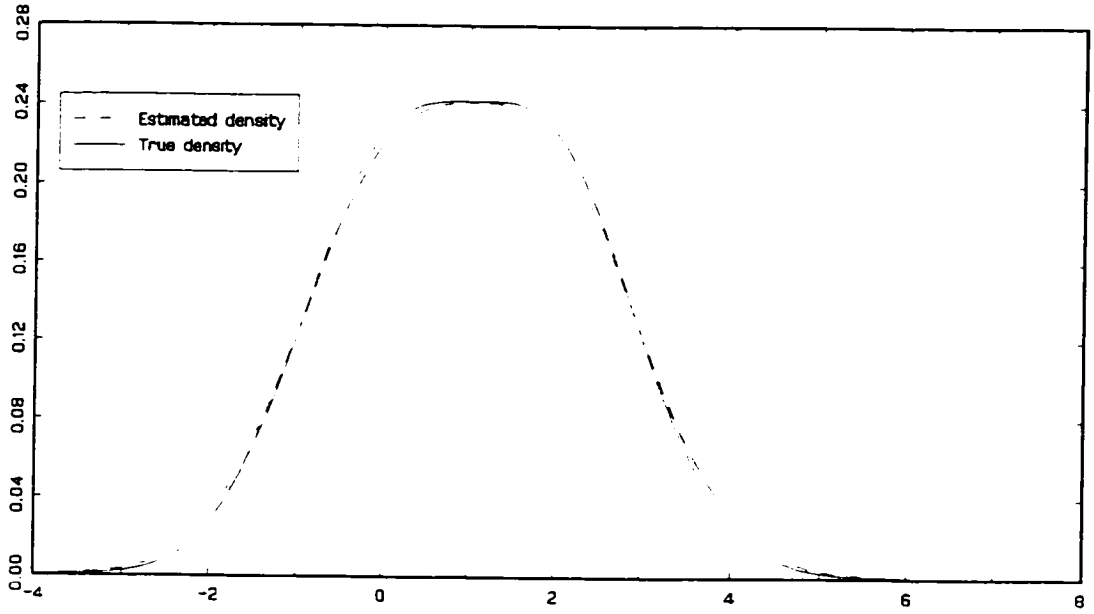
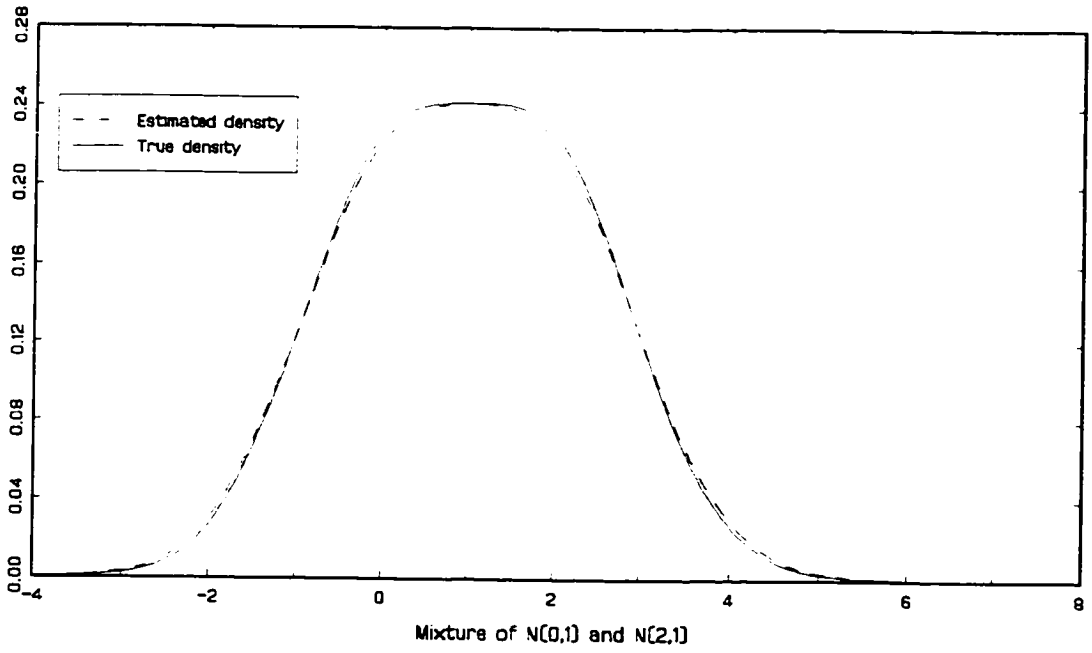


Figure 4.4b
Unweighted Estimate Using h_a
Proportional Sampling



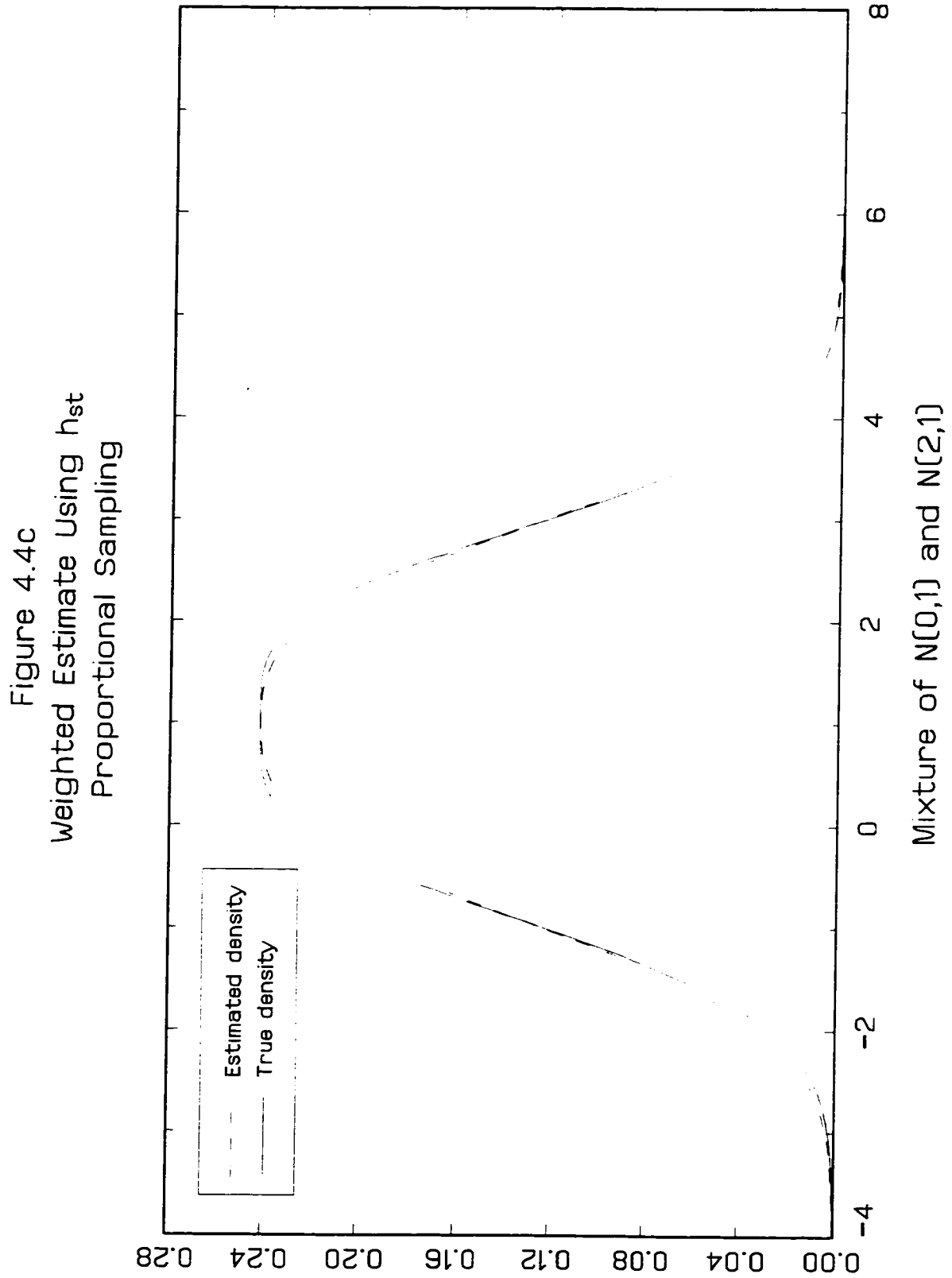


Figure 4.4c
Weighted Estimate Using h_{st}
Proportional Sampling

Figure 4.5a
Unweighted Estimate Using h^*
Proportional Sampling

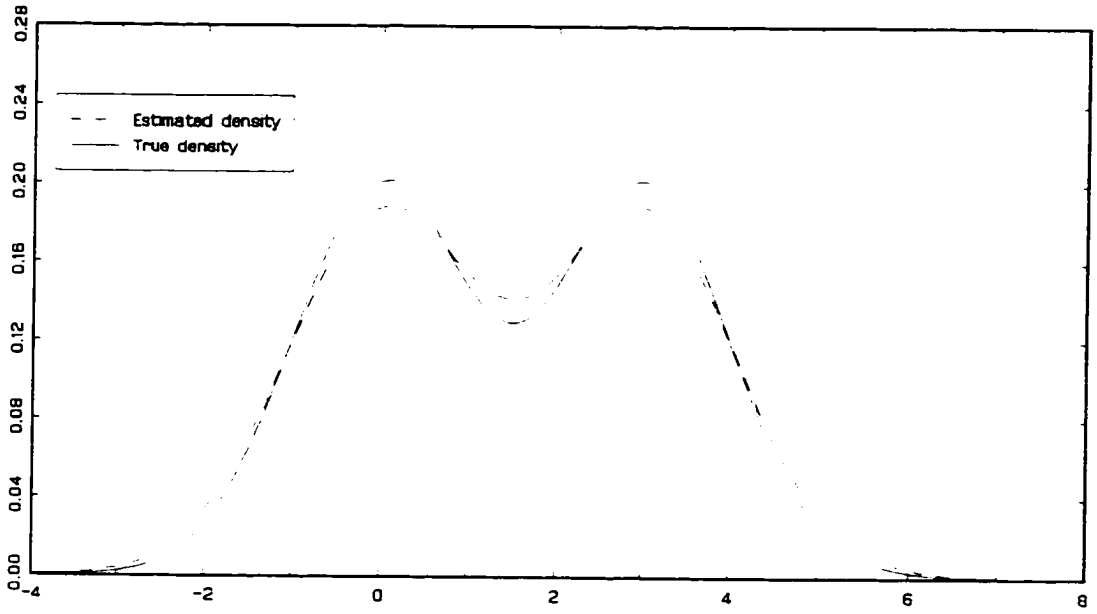


Figure 4.5b
Unweighted Estimate Using h_a
Proportional Sampling

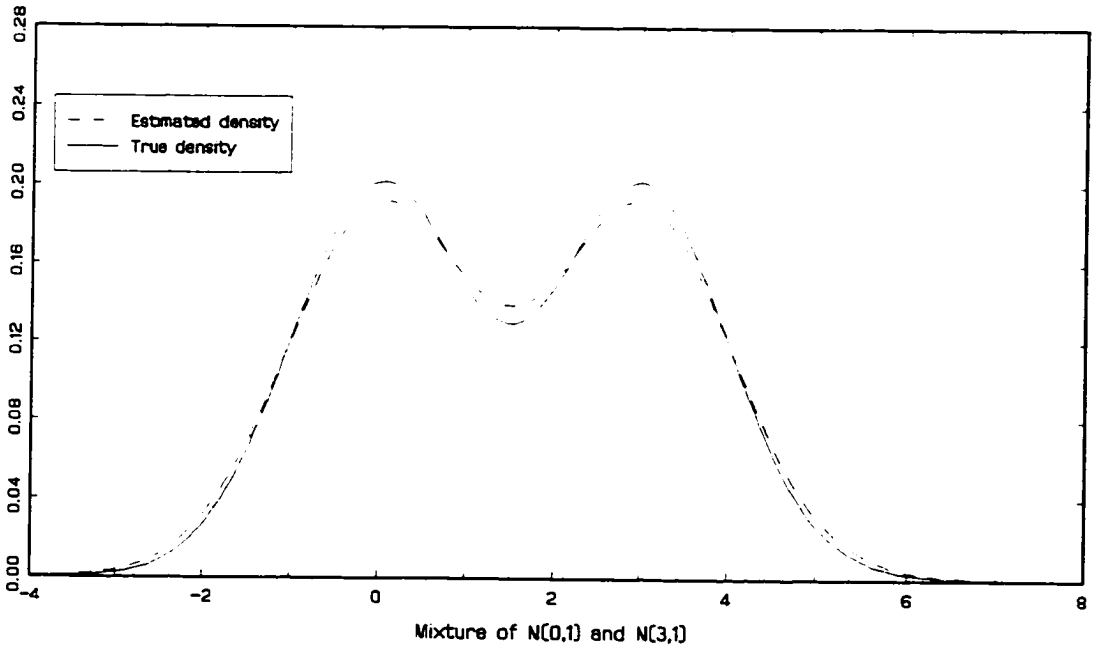


Figure 4.5c
Weighted Estimate Using h_{st}
Proportional Sampling

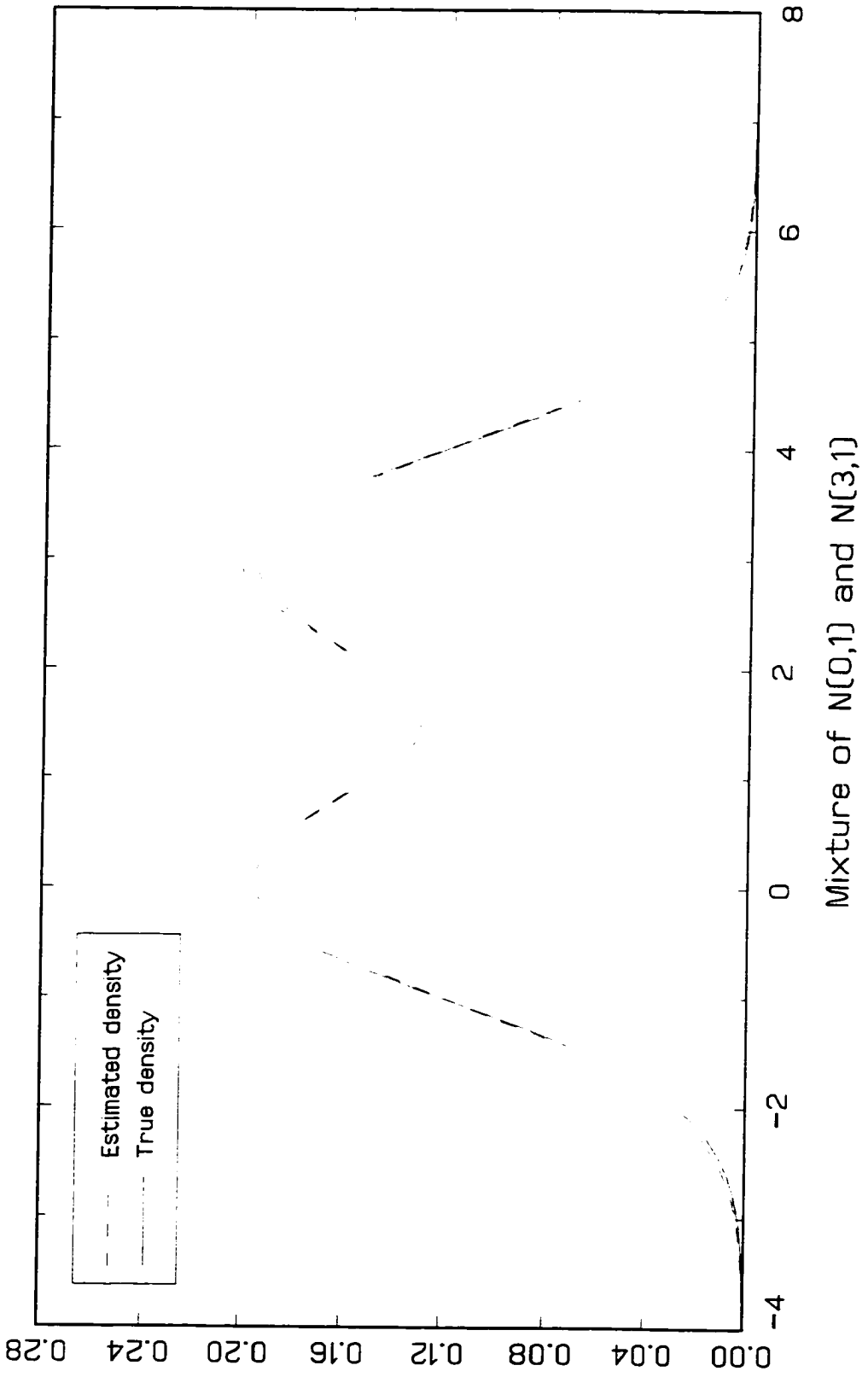


Figure 4.6a
Unweighted Estimate Using h^*
Proportional Sampling

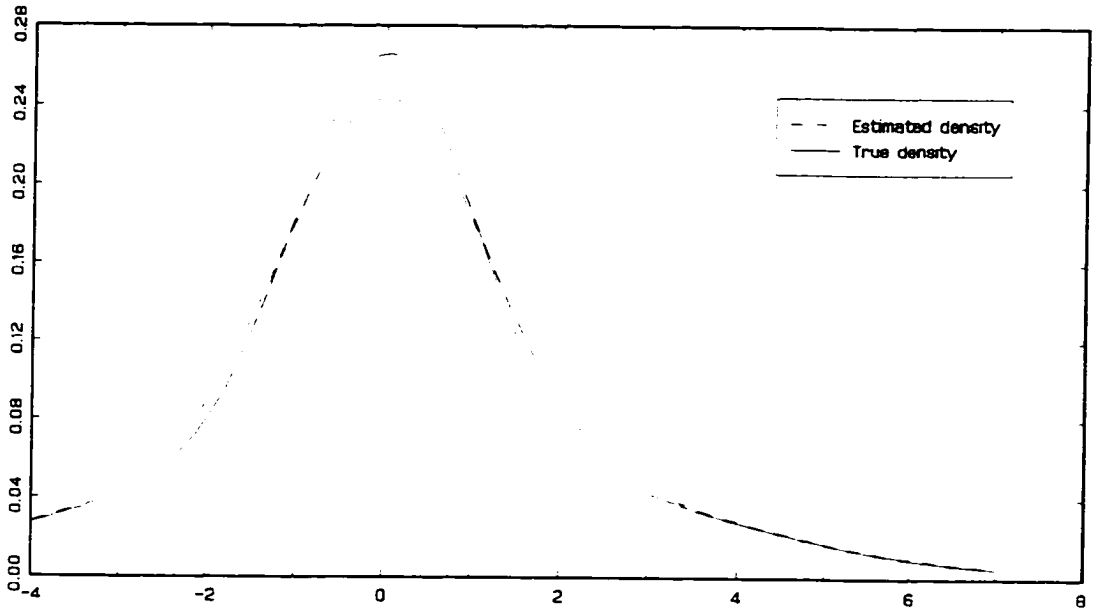


Figure 4.6b
Unweighted Estimate Using h_a
Proportional Sampling

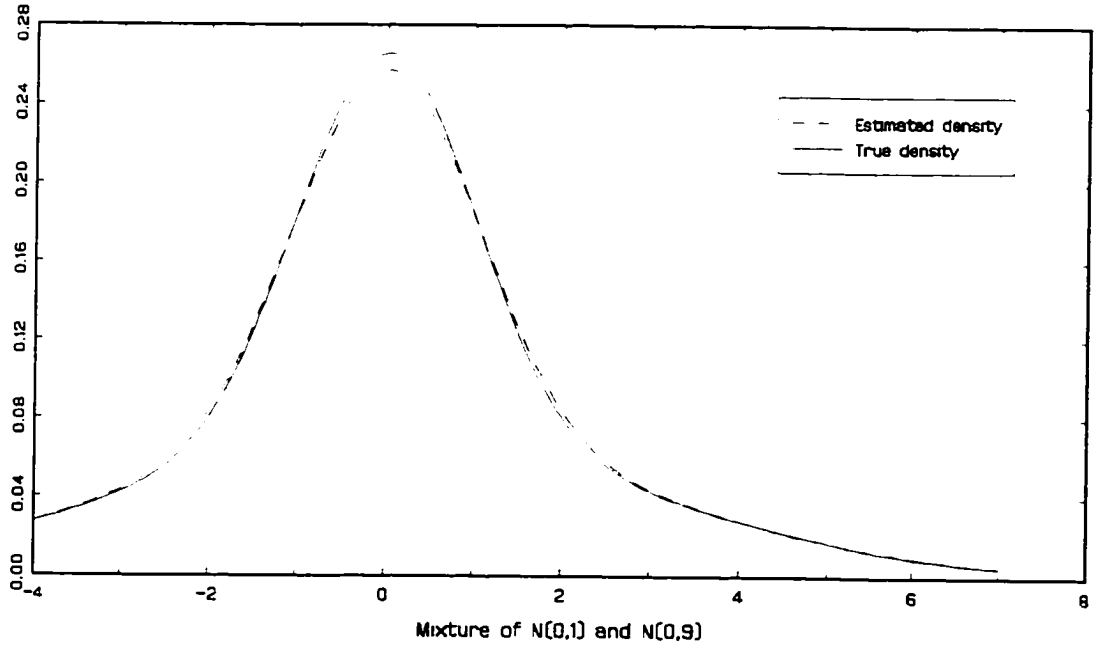


Figure 4.6c
Weighted Estimate Using h_{st}
Proportional Sampling

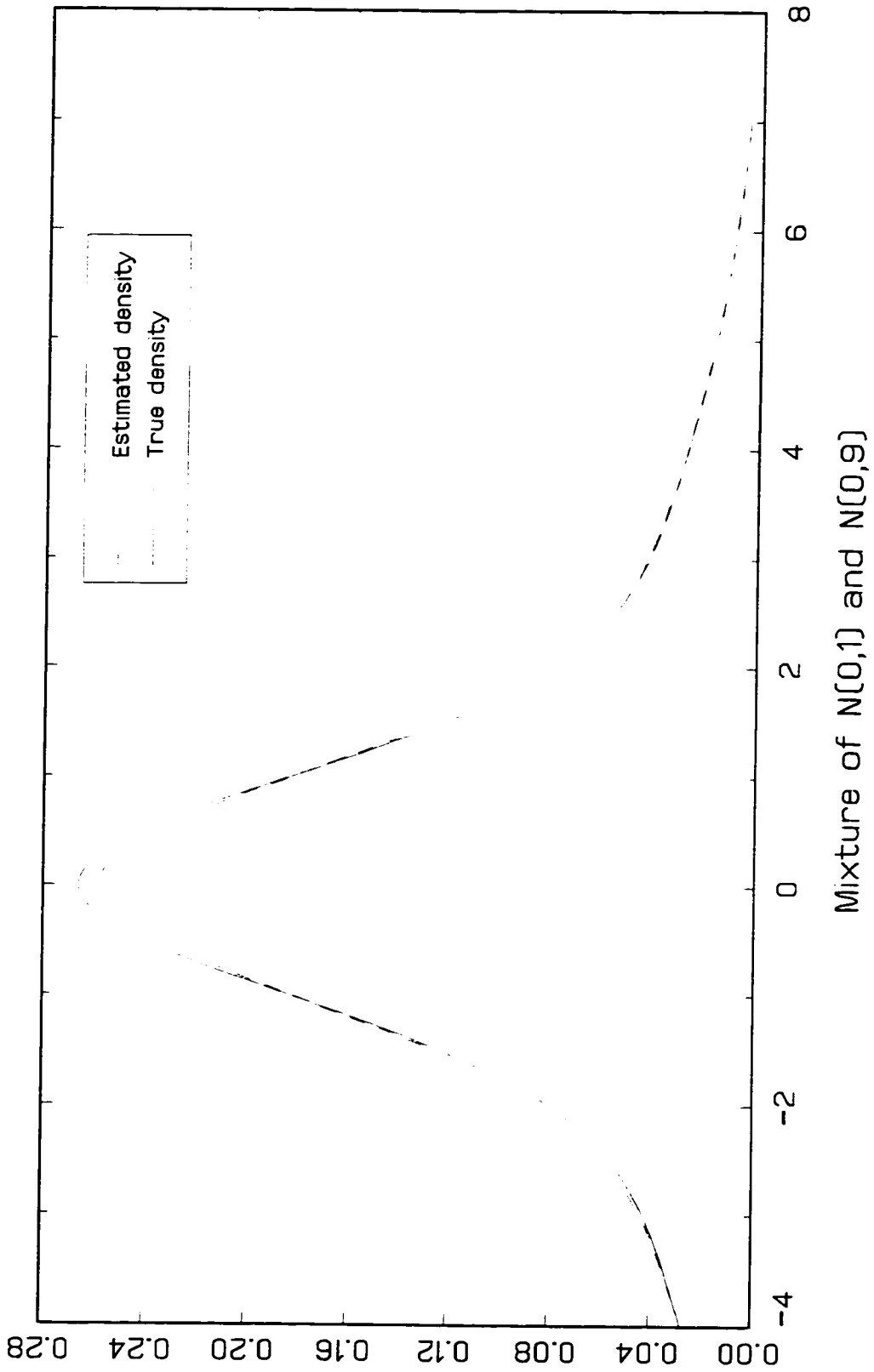


Figure 4.7a
Unweighted Estimate Using h^*
Proportional Sampling

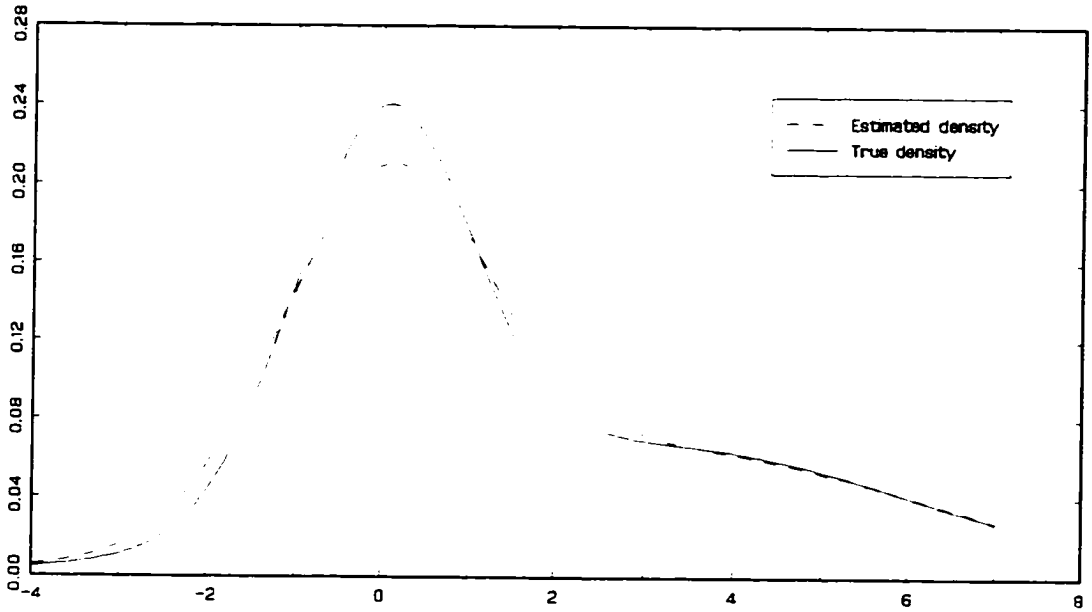


Figure 4.7b
Unweighted Estimate Using h_a
Proportional Sampling

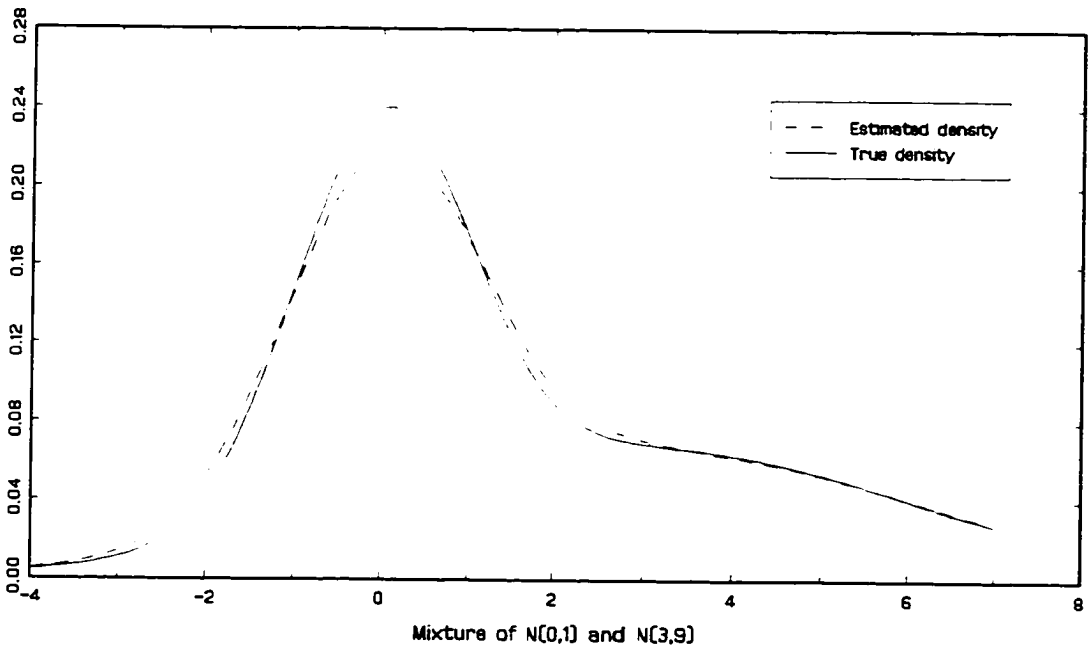


Figure 4.7c
Weighted Estimate Using h_{st}
Proportional Sampling

Figure 4.8a
Unweighted Estimate Using h*
Non-Proportional Sampling n2/n1=2

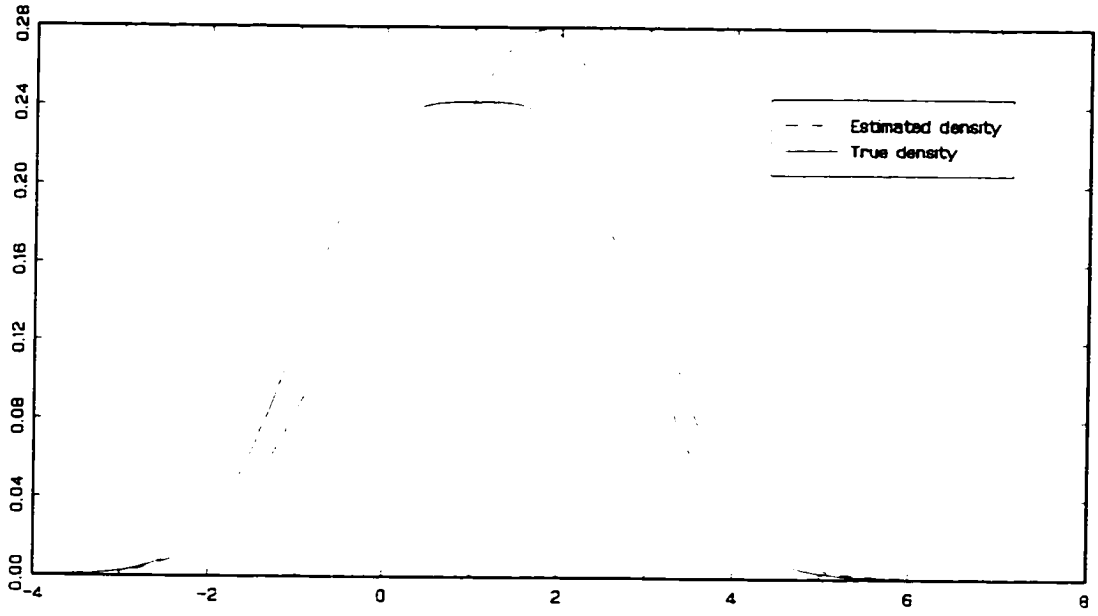


Figure 4.8b
Unweighted Estimate Using n_a
Non-Proportional Sampling n2/n1=2

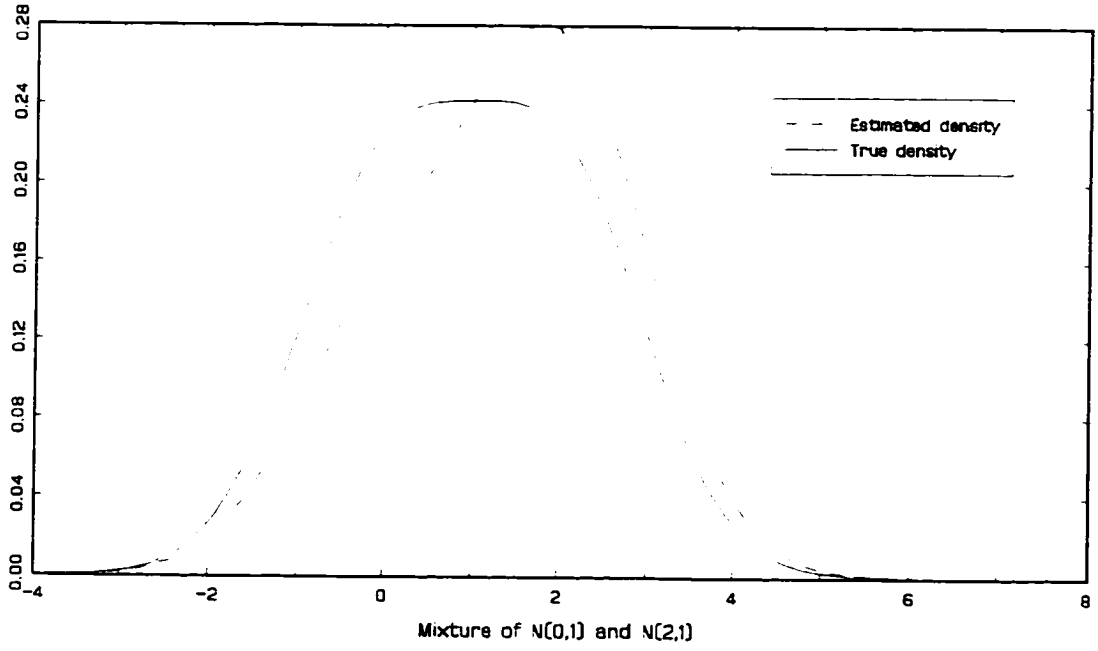


Figure 4.8c
Weighted Estimate Using h_{st}
Non-Proportional Sampling n2/n1=2

Figure 4.9a
Unweighted Estimate Using h^*
Non-Proportional Sampling n2/n1=2

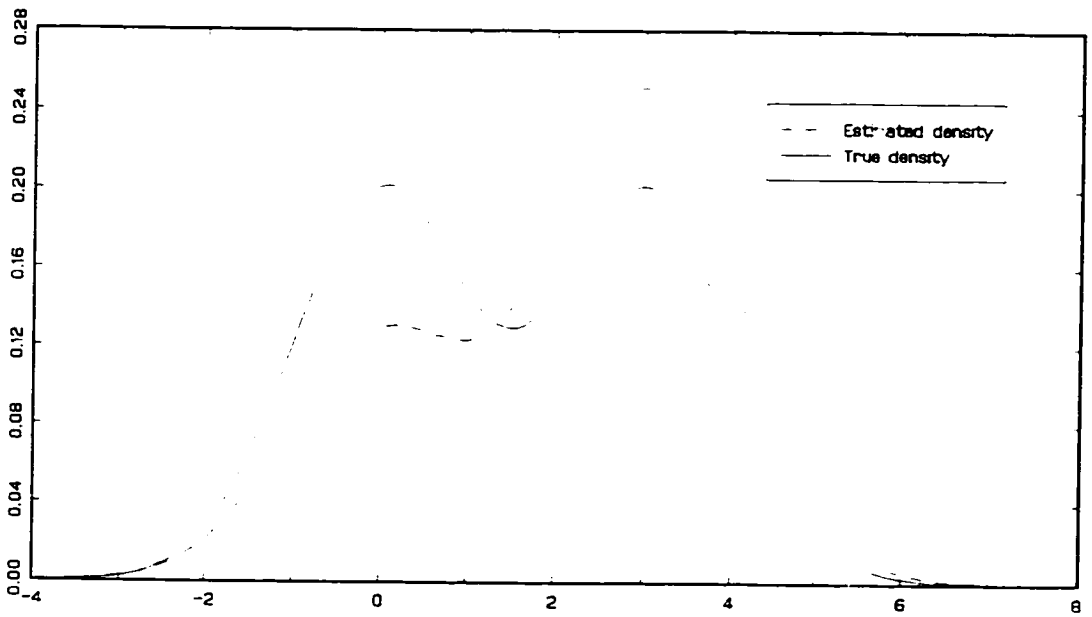
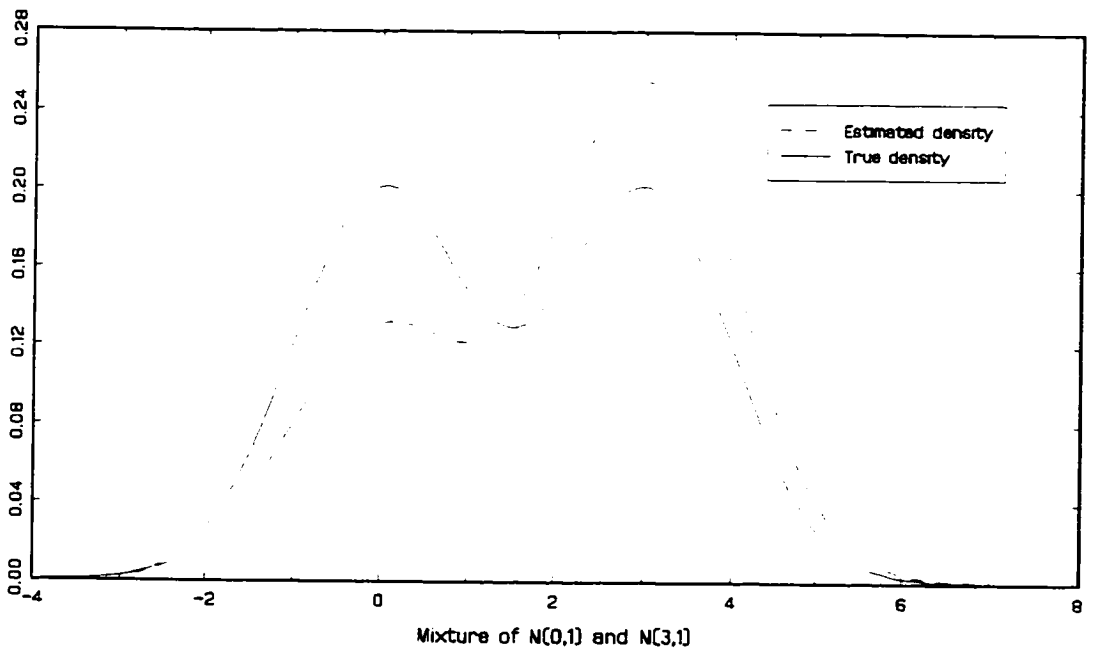


Figure 4.9b
Unweighted Estimate Using h_a
Non-Proportional Sampling n2/n1=2



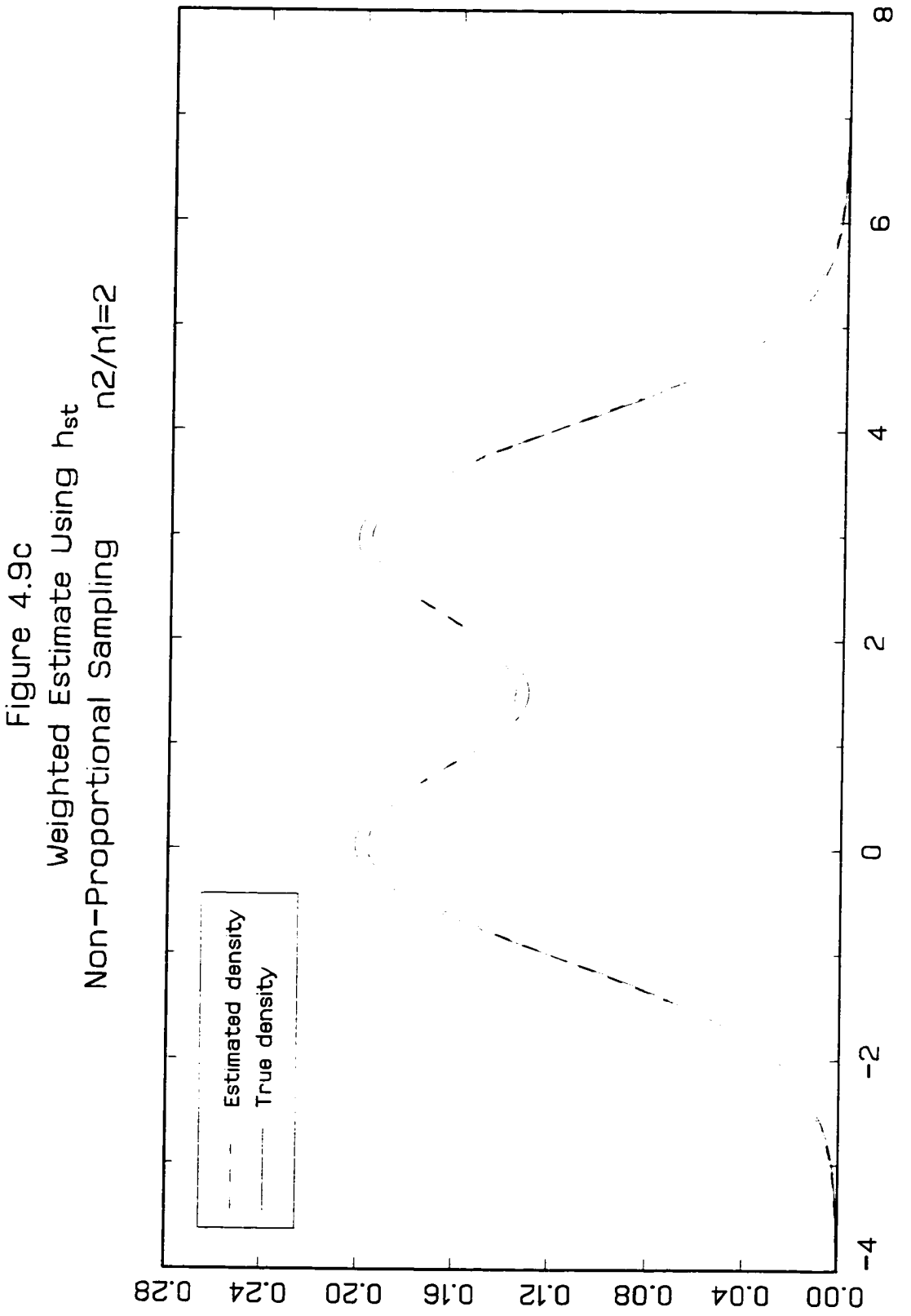


Figure 4.10a
Unweighted Estimate Using h*
Non-Proportional Sampling n2/n1=2

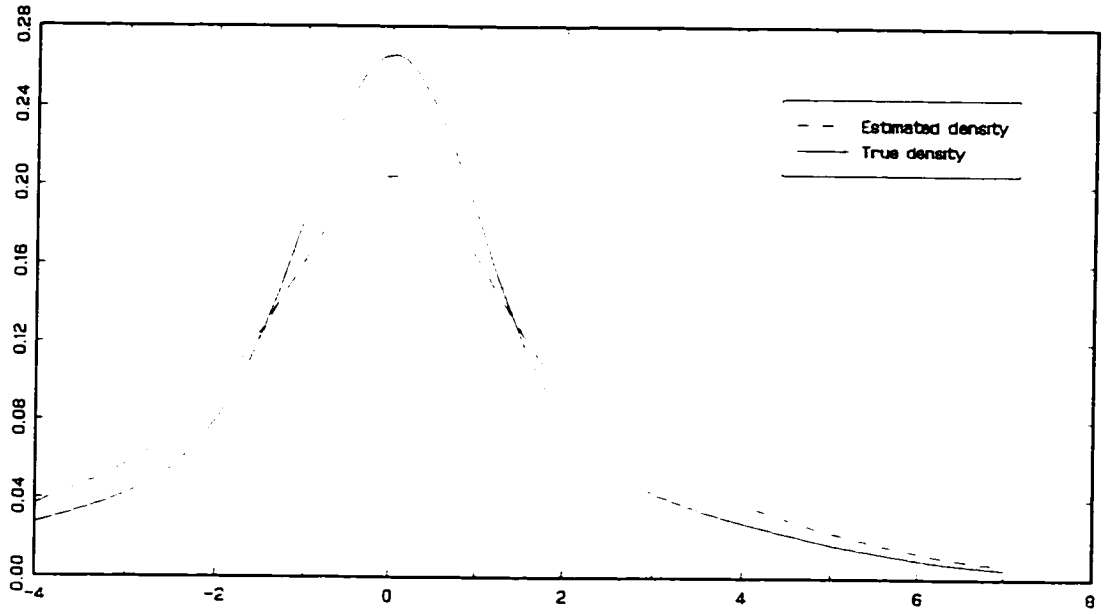


Figure 4.10b
Unweighted Estimate Using h_a
Non-Proportional Sampling n2/n1=2

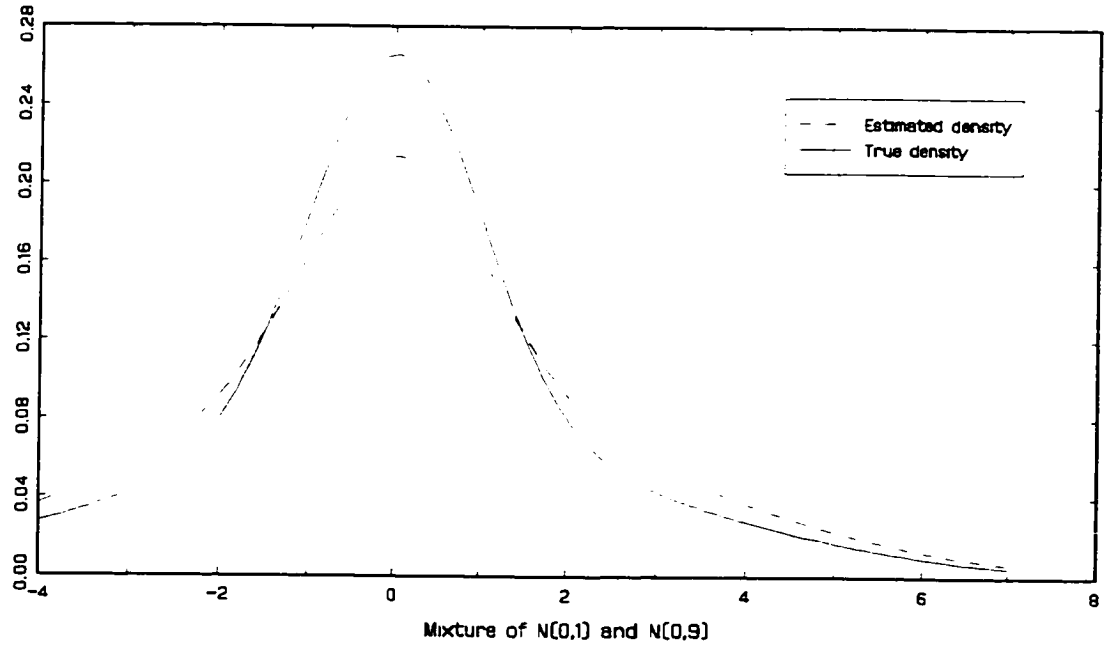


Figure 4.10c
Weighted Estimate Using h_{st}
Non-Proportional Sampling n2/n1=2

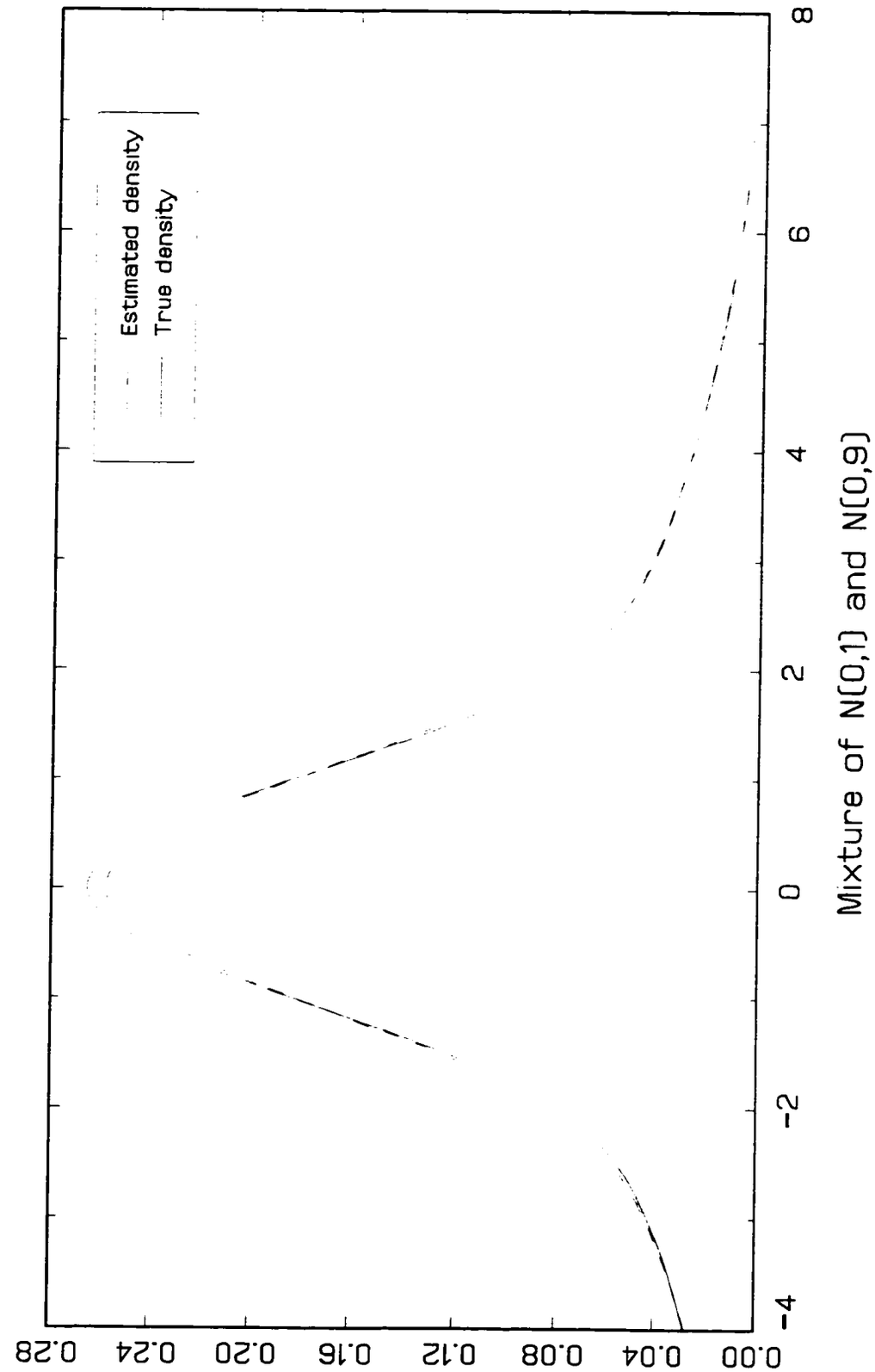


Figure 4.11a
Unweighted Estimate Using h^*
Non-Proportional Sampling n2/n1=2

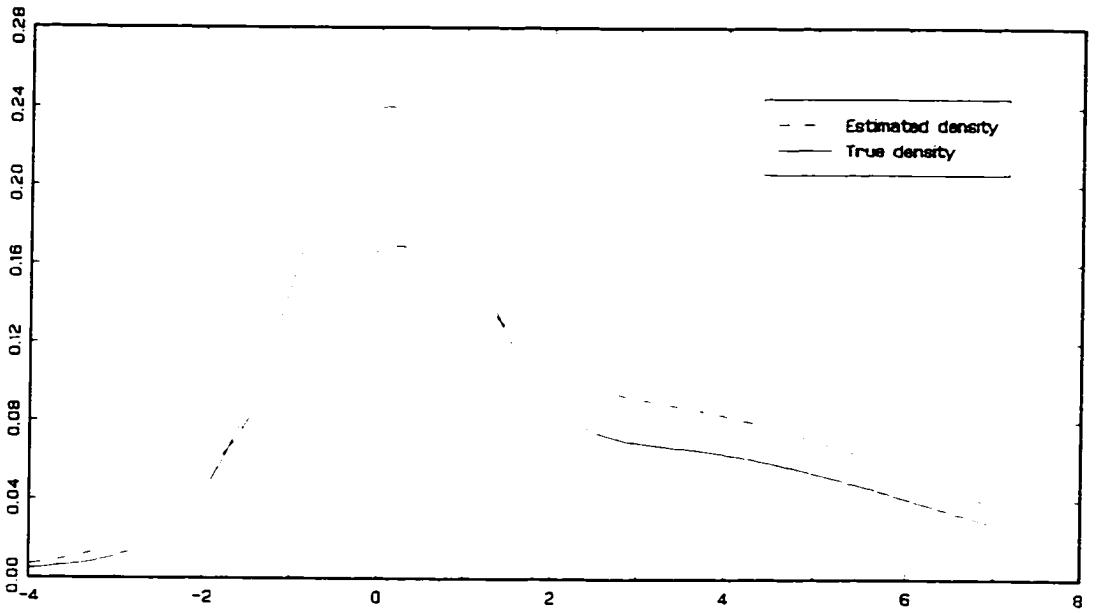


Figure 4.11b
Unweighted Estimate Using h_s
Non-Proportional Sampling n2/n1=2

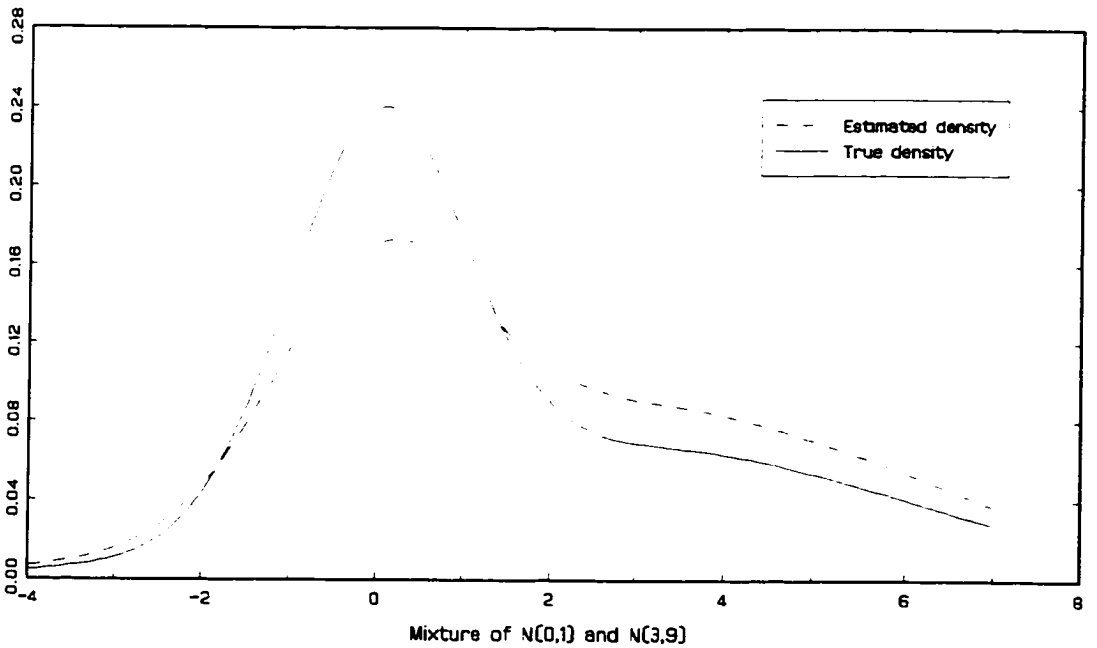


Figure 4.11c
Weighted Estimate Using h_{st}
Non-Proportional Sampling n2/n1=2

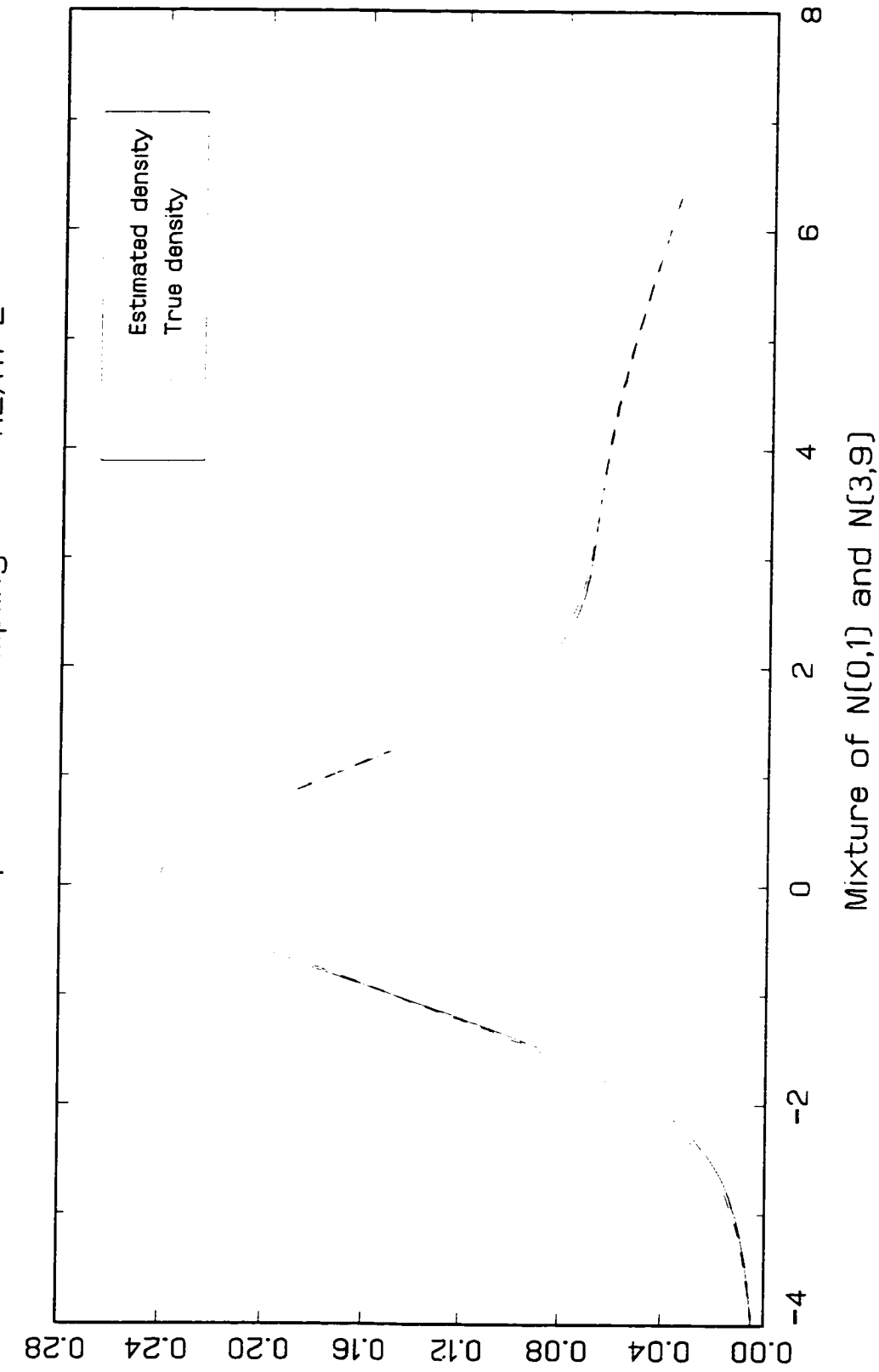


Table 4.2a
Comparison of IMSE from Weighted and Unweighted Estimation

Identical Standard deviations	$\mu_2-\mu_1$	h^*	h_{st}	IMSE(h^*)	IMSE(h_{st})	IMSE(h_{st})/ IMSE(h^*)
Proportional Sampling:	0.00	0.26626	0.26607	0.00133	0.00133	1.00000
	0.50	0.28290	0.27455	0.00129	0.00128	0.99815
	1.00	0.33283	0.30181	0.00119	0.00117	0.97922
	1.50	0.41603	0.34725	0.00110	0.00102	0.92610
	2.00	0.53252	0.36956	0.00135	0.00095	0.70549
	2.50	0.68229	0.34059	0.00375	0.00104	0.27622
	3.00	0.86535	0.31439	0.01320	0.00112	0.08496
	3.50	1.08168	0.30216	0.03859	0.00117	0.03024
	4.00	1.33130	0.29899	0.09292	0.00118	0.01269
	4.50	1.61420	0.30016	0.19669	0.00118	0.00597
	5.00	1.93039	0.30242	0.38729	0.00117	0.00301
Dis-proportional Sampling: $n_2/n_1=2$	0.00	0.26626	0.27243	0.00146	0.00146	0.99897
	0.50	0.28290	0.28112	0.00141	0.00141	0.99992
	1.00	0.33283	0.30903	0.00130	0.00128	0.98825
	1.50	0.41603	0.35556	0.00118	0.00112	0.94466
	2.00	0.53252	0.37840	0.00142	0.00105	0.73915
	2.50	0.68229	0.34874	0.00380	0.00114	0.29948
	3.00	0.86535	0.32191	0.01324	0.00123	0.09310
	3.50	1.08168	0.30939	0.03862	0.00128	0.03321
	4.00	1.33130	0.30615	0.09295	0.00130	0.01395
	4.50	1.61420	0.30735	0.19671	0.00129	0.00657
	5.00	1.93039	0.30966	0.38731	0.00128	0.00331

Table 4.2b
Comparison of IMSE from Weighted and Unweighted Estimation

Identical means	$\sigma_2-\sigma_1$	h^*	h_{st}	IMSE(h^*)	IMSE(h_{st})	IMSE(h_{st})/ IMSE(h^*)
Proportional Sampling:	1.00	0.26626	0.26607	0.00133	0.00133	1.00000
	1.50	0.43267	0.31478	0.00145	0.00112	0.77163
	2.00	0.66565	0.33664	0.00363	0.00105	0.28885
	2.50	0.96519	0.34506	0.01280	0.00102	0.07982
	3.00	1.33130	0.34834	0.04341	0.00101	0.02332
	3.50	1.76397	0.34971	0.13070	0.00101	0.00772
	4.00	2.26321	0.35034	0.35068	0.00101	0.00287
	4.50	2.82901	0.35066	0.85215	0.00101	0.00118
	5.00	3.46138	0.35082	1.90508	0.00101	0.00053
	5.50	4.16031	0.35092	3.97038	0.00101	0.00025
	6.00	4.92581	0.35097	7.79642	0.00101	0.00013
Dis-proportional Sampling: $n_2/n_1=2$	1.00	0.26626	0.27243	0.00146	0.00146	0.99897
	1.50	0.43267	0.32231	0.00153	0.00123	0.80293
	2.00	0.66565	0.34470	0.00368	0.00115	0.31292
	2.50	0.96519	0.35333	0.01284	0.00112	0.08749
	3.00	1.33130	0.35668	0.04343	0.00111	0.02562
	3.50	1.76397	0.35809	0.13072	0.00111	0.00848
	4.00	2.26321	0.35873	0.35070	0.00111	0.00316
	4.50	2.82901	0.35905	0.85216	0.00111	0.00130
	5.00	3.46138	0.35922	1.90509	0.00111	0.00058
	5.50	4.16031	0.35932	3.97039	0.00111	0.00028
	6.00	4.92581	0.35937	7.79643	0.00110	0.00014

Figure 4.12a
Ratio of Integrated Mean Squared Error
$\sigma_1=\sigma_2=1$ Proportional Sampling

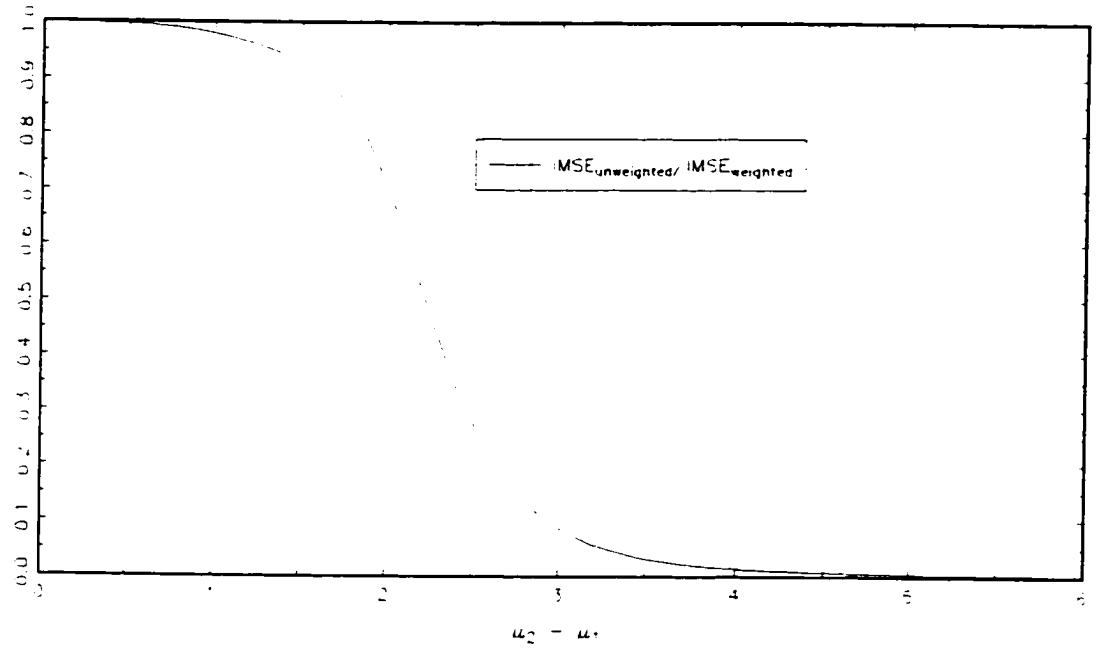


Figure 4.12b
Ratio of Integrated Mean Squared Error
$\mu_1=\mu_2=0$ Proportional Sampling

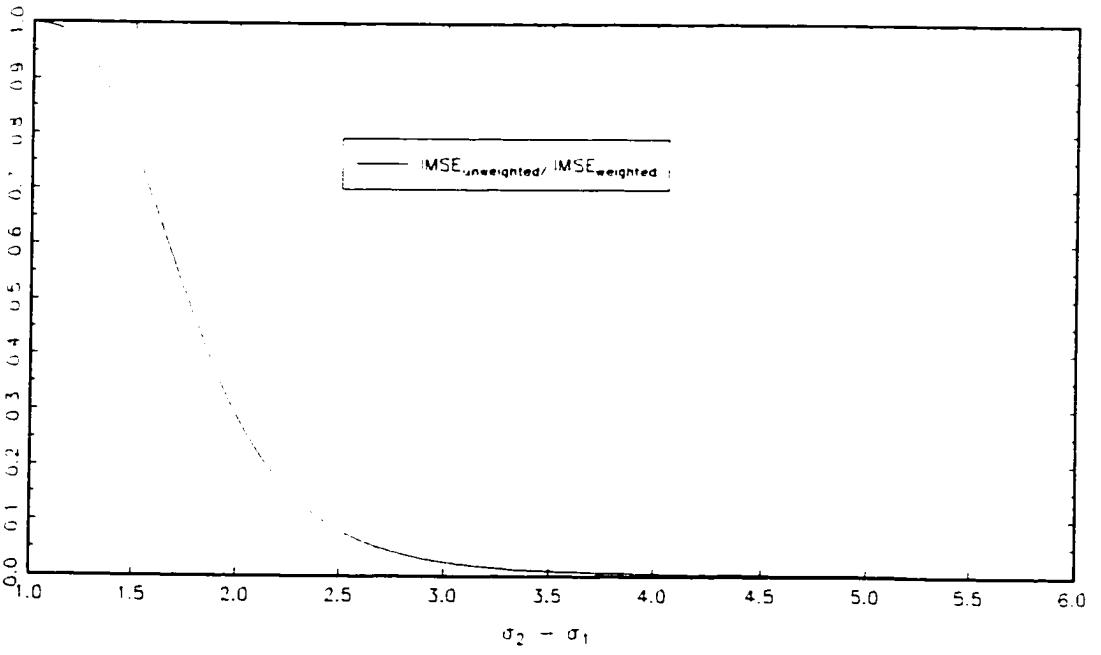


Figure 4.13a
Ratio of Integrated Mean Squared Error
$\sigma_1=\sigma_2=1$ Non-Proportional Sampling $n2/n1=2$

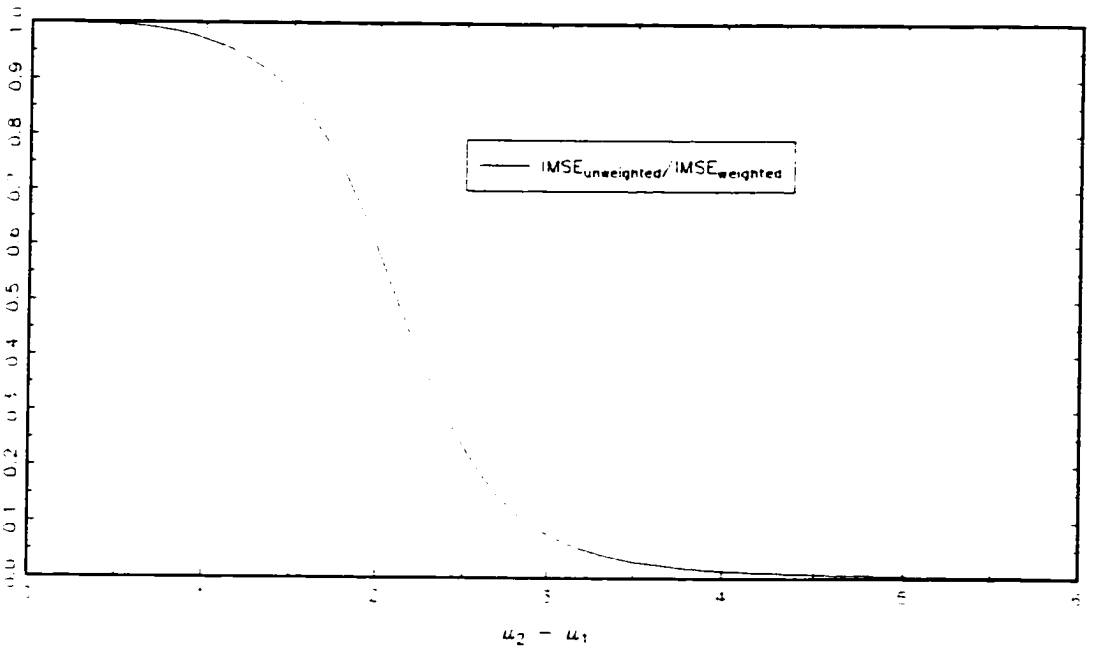
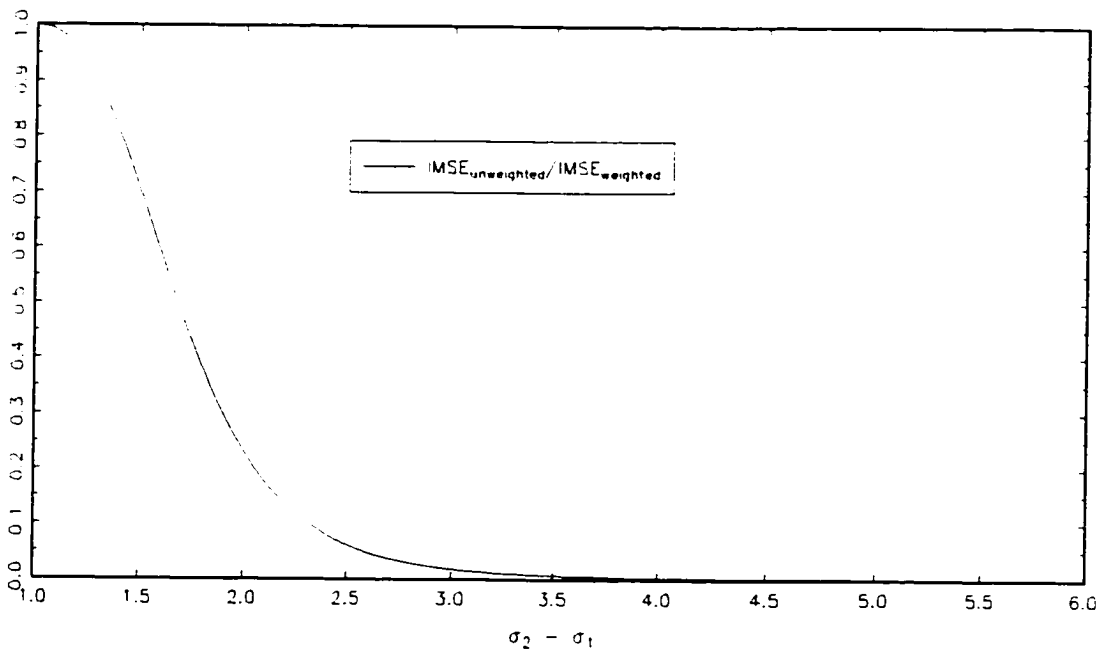


Figure 4.13b
Ratio of Integrated Mean Squared Error
$\mu_1=\mu_2=0$ Non-Proportional Sampling $n2/n1=2$



4.3 Clustered Sampling

Much of the data used for economic analysis is gathered using survey methods leading to sample data which may violate the i.i.d. assumption. Serial correlation of data is a well-known problem in time-series data, but it is also present in much cross-sectional data, where it is usually ignored by analysts. Most cross-sectional data for economic analysis is gathered through some type of complex survey (see Ullah and Breunig (1998)). Data is usually selected from populations which are stratified and clustered using well-known survey sampling techniques. Clustering, frequently employed to reduce the cost of data collection, generally leads to positive correlation between data points in the same cluster. Much applied cross-sectional, econometric analysis ignores the correlation which is present in such data. In particular, for the case of non-parametric density estimation, such effects from the sampling structure have not been considered in the literature.

Below, we relax the assumption of independently distributed data and consider the problem of kernel density estimation for clustered data. As mentioned above, the choice of window width for kernel density estimation with i.i.d. data has been considered by many authors. Silverman (1986) provides an excellent review.

In this section, we obtain the approximate integrated mean squared error (IMSE) for the kernel density estimation under cluster sampling. An optimal window width

is proposed which minimizes the approximate IMSE. This result suggests that the usual optimal window width for i.i.d. data does not hold in the case of clustered data. The combination of a fourth-order kernel and a window width which depends on the degree of correlation in the data turns out to perform well in application and the suggested optimal window width performs better in an integrated mean-squared error sense.

The population model which we have in mind is the same as in (59) above

$$Y_{hci} = \mu_{hc} + U_{hci} \tag{214}$$

where $\sum_c \sum_i U_{hci} = 0$ by definition of μ_{hc}. We will confine our analysis to a single stratum for simplicity, so we can write the population model for one stratum as

$$\begin{aligned} Y_{ci} &= \mu + U_{ci} \\ &= \mu + \alpha_c + \epsilon_{ci}. \end{aligned} \tag{215}$$

Consider the case of clustered data, where a sample of n units has been drawn from this population using cluster sampling. It is assumed that the data is drawn in two stages; a sample of k "clusters" is randomly chosen at the first stage; and in the second stage a sample of n_c elements is chosen from each cluster, $c = 1, ..., k$.[16] The

[16]Clustering is frequently found in economic data gathered from surveys. One common example is the income and expenditure survey, where first a sample of villages is chosen and then, within each village, households are randomly selected. Households within the same village (or cluster) can be assumed to face similar conditions–for example we expect heating fuel costs to be correlated for households in the same area. In this paper, I assume that the data has already been gathered and that the analyst has information about the structure of the data. Kish (1965) and Thomspon (1992) provide details of how clustered surveys are conducted.

total sample size is thus $n = \sum_c n_c$. The sample model is

$$y_{ci} = \mu + u_{ci}, \qquad c = 1, ..., C, \; i = 1, ..., n_c. \tag{216}$$

For simplicity, I assume that in the second stage, $n_c = \frac{n}{k} = m$ elements are chosen from each cluster. Also known as "balanced" clusters, this assumption will be relaxed below. We will assume that the sample is drawn as a random sample with replacement such that

$$\begin{aligned} E\, u_{ci} &= 0, \; E\, u_{ci}^2 = \sigma^2, \; E\, u_{ci} u_{cj} = \rho\sigma^2, \; i \neq j \\ E\, u_{ci}\, u_{c'j} &= 0, \; c \neq c'; \end{aligned} \tag{217}$$

$\rho > 0$ is called the intra-cluster correlation coefficient. (217) implies that the elements within clusters are correlated, but are uncorrelated across clusters[17]

The problem of non-parametrically estimating the density for the case of i.i.d. data is well-studied. (See Silverman (1986), Härdle (1990), Pagan and Ullah (1997)). Choosing the optimal window width, h^*, by minimizing the approximate integrated mean squared error (AMISE) of the density estimator provides $h^* \; \alpha \; n^{-\frac{1}{5}}$ when a second-order kernel is used. (Generally, $h^* = cn^{-\frac{1}{(2P+1)}}$ for a Pth order kernel.) Furthermore, if the underlying true density of the data is normal with variance σ^2 and if the kernel is Gaussian, then the optimal (in the sense of minimizing the AIMSE)

[17]The intra-cluster correlation coefficient, ρ, can be surprisingly large in cross-sectional data. Deaton (1996) provides examples using World Bank data where intra-cluster correlation coefficients range from .2 to .5.

window width is

$$h^* = 1.06\sigma n^{-\frac{1}{5}}. \tag{218}$$

(See Silverman (1986).)

I will examine how the result changes if the data follows (217) above. Specifically, I will use the same method of minimizing the approximate integrated mean squared error with respect to h, and solving for the optimal window width.

The non-parametric, kernel estimate of the density at any point y is

$$\widehat{f(y)} = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{y_j - y}{h}\right) \tag{219}$$

where h is the window width which is assumed to satisfy (A1) from section 4.2, n is the sample size, and the kernel $K(\cdot)$ is a symmetric function which satisfies:

(i) $\int \psi^3 K(\psi)d\psi = \mu_3 < \infty$

(ii) $\int \psi^4 K(\psi)d\psi = \mu_4 < \infty$

in addition to satisfying (A2) from section 4.2.

For the case of clustered data, we can re-write the Kernel density estimate as

$$\widehat{f(y)} = \frac{1}{nh}\sum_{c=1}^{k}\sum_{i=1}^{n_c} K\left(\frac{y_{ci} - y}{h}\right) \tag{220}$$

where n_c is the number of observations in the cth cluster. For the derivation which follows, we have assumed $n_c = \frac{n}{k} = m$. It is straightforward to show that the bias of $\widehat{f(y)}$ upto $O(h^2)$ is

$$\text{bias } \widehat{f(y)} = \frac{h^2}{2} f''(y)\mu_2 \tag{221}$$

(see Silverman (1986), p. 39).

To find the variance of $\widehat{f(y)}$ under the assumptions about the data (217), we first re-write $\widehat{f(y)}$ as

$$\widehat{f(y)} = \frac{1}{n}\sum_{c=1}^{k}\sum_{i=1}^{n_c} W_{ci} \tag{222}$$

where

$$W_{ci} = \frac{1}{h}K\left(\frac{y_{ci}-y}{h}\right). \tag{223}$$

Then,

$$Var(\widehat{f(y)}) = \frac{1}{n^2}\sum_{c=1}^{k}\sum_{i=1}^{n_c} Var(W_{ci}) + \frac{1}{n^2}\sum_{c=1}^{k}\sum_{c'=1}^{k}\sum_{j=1}^{n_{c'}}\sum_{\substack{i=1 \\ i\neq j \text{ for } c=c'}}^{n_c} Cov(W_{ci}, W_{c'j}). \tag{224}$$

$Var(W_{ci}) = EW_{ci}^2 - (EW_{ci})^2$, and since the data are identically distributed, $Var(W_{ci}) = Var(W_{11})$. Using this information and the assumption that elements across clusters are uncorrelated

$$Var(\widehat{f}(y)) = \frac{1}{n}EW_{11}^2 - \frac{1}{k}(EW_{11})^2 + \left(\frac{1}{k}-\frac{1}{n}\right)E\left(W_{11}\cdot W_{12}\right). \tag{225}$$

First, consider term 1:

$$\tfrac{1}{n}E(W_{11})^2 = \tfrac{1}{n}\int_{y_{11}}\left(\tfrac{1}{h}K\left(\tfrac{y_{11}-y}{h}\right)\right)^2 f(y_{11})dy_{11} = \tfrac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})\, f(h\psi_{11}+y)\, d\psi_{11}.$$

Expanding $f(y_{11})$ around the point $y_{11} = y$ by the method of Taylor's series,

$$\frac{1}{n}E(W_{11})^2 = \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})\left[f(y) + f'(y)h\psi_{11} + f''(y)(h\psi_{11})^2 + \cdots\right] d\psi_{11}.$$

Keeping terms up to $O\left(\frac{1}{nh}\right)$ gives an approximation for the first term of $Var(\widehat{f(y)})$,

$$\frac{1}{n}E(W_{11})^2 = \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})\, f(y) d\psi_{11}. \tag{226}$$

Now, consider term two:

$$\frac{1}{k}(EW_{11})^2 = \frac{1}{k}\left\{\frac{1}{h}\int_{y_{11}} K^{\cdot}\left(\frac{y_{11}-y}{h}\right) f(y_{11})dy_{11}\right\}^2 = \frac{1}{k}\left\{\int_{\psi_{11}} K^{\cdot}(\psi_{11})\, f(h\psi_{11}+y)d\psi_{11}\right\}^2.$$

Here, $f(y_{11})$ is replaced with a Taylor's series expansion around the point $y_{11} = y$.

$$\frac{1}{k}(EW_{11})^2 = \frac{1}{k}\left\{\int_{\psi_{11}} K^{\cdot}(\psi_{11})\left[f(y) + f'(y)h\psi_{11} + \frac{f''(y)}{2}(h\psi_{11})^2 + \frac{f'''(y)}{6}(h\psi_{11})^3 + \frac{f''''(y)}{24}(h\psi_{11})^4 + \cdots\right] d\psi_{11}\right\}^2$$

$$= \frac{1}{k}\left\{f(y)\int_{\psi_{11}} K^{\cdot}(\psi_{11})\,d\psi_{11} + hf'(y)\int_{\psi_{11}} \psi_{11}K^{\cdot}(\psi_{11})\,d\psi_{11} + \frac{h^2}{2}f''(y)\int_{\psi_{11}} \psi_{11}K^{\cdot}(\psi_{11})\,d\psi_{11} + \frac{h^3}{6}f'''(y)\int_{\psi_{11}} \psi_{11}K^{\cdot}(\psi_{11})\,d\psi_{11} + \frac{h^4}{24}f''''(y)\int_{\psi_{11}} \psi_{11}K^{\cdot}(\psi_{11})\,d\psi_{11}\right\}^2$$

$$= \frac{1}{k}\left\{f(y) + \frac{h^2 f''(y)\mu_2}{2} + \frac{h^3 f'''(y)\mu_3}{6} + \frac{h^4 f''''(y)\mu_4}{24}\right\}^2. \tag{227}$$

Which gives, up to order $O(h^4)$,

$$= \frac{1}{k}\left\{(f(y))^2 + \frac{h^4(f''(y))^2\mu_2^2}{4} + f(y)h^2f''(y)\mu_2 + \frac{f(y)h^3f'''(y)\mu_3}{3} + \frac{f(y)h^4f''''(y)\mu_4}{12}\right\}. \tag{228}$$

Now, consider term three:

$$\left(\frac{1}{k}-\frac{1}{n}\right)E(W_{11}W_{12}) = \left(\frac{1}{k}-\frac{1}{n}\right)\int_{y_{11}}\int_{y_{12}} \frac{1}{h^2}K^{\cdot}\left(\frac{y_{11}-y}{h}\right)K^{\cdot}\left(\frac{y_{12}-y}{h}\right) f(y_{11},y_{12})dy_{11}dy_{12}. \tag{229}$$

We first transform the density $f(y_{11}, y_{12})$ following Rao (1973)

$$f(\psi_{11},\psi_{12}) = f(y_{11},y_{12})\begin{vmatrix} \frac{\partial y_{11}}{\partial(\psi_{11})} & \frac{\partial y_{12}}{\partial(\psi_{11})} \\ \frac{\partial y_{11}}{\partial(\psi_{12})} & \frac{\partial y_{12}}{\partial(\psi_{12})} \end{vmatrix}. \tag{230}$$

Evaluation of the determinant in equation (230) under assumptions (217) gives $h^2(1-\rho^2)$, thus after replacing y_{11} with $h\psi_{11}+y$ and y_{12} with $h\psi_{12}+y$, term three becomes

$$\left(\frac{h^2(1-\rho^2)}{h^2}\right)\left(\frac{1}{k}-\frac{1}{n}\right)\int_{\psi_{11}}\int_{\psi_{12}} K(\psi_{11})K(\psi_{12})f(h\psi_{11}+y,h\psi_{12}+y)d\psi_{11}d\psi_{12} \quad (231)$$

Using a bivariate Taylor series expansion of $f(y_{11},y_{12})$ around the point $y_{11}=y$, $y_{12}=y$, this term becomes

$$\left(1-\rho^2\right)\left(\frac{1}{k}-\frac{1}{n}\right)\int_{\psi_{11}}\int_{\psi_{12}} K(\psi_{11})K(\psi_{12})\left[f(y,y)+f_1(y,y)h\psi_{11}+f_2(y,y)h\psi_{12}\right.$$
$$\left.+\frac{1}{2}f_{11}(y,y)(h\psi_{11})^2+\frac{1}{2}f_{22}(y,y)(h\psi_{12})^2+f_{12}(y,y)(h\psi_{11})(h\psi_{12})+\cdots\right]d\psi_{11}d\psi_{12} \quad (232)$$

which after simplification, and keeping only terms upto $O(\max\left\{\frac{1}{nh},h^4\right\})$

$$= \left(\frac{1-\rho^2}{k}\right)\left[f(y,y)+h^2f_{11}(y,y)\mu_2+\frac{h^3}{3}f_{111}(y,y)\mu_3\right.$$
$$\left.+\frac{h^4}{12}f_{1111}(y,y)\mu_4+\frac{h^4}{4}f_{1122}(y,y)\mu_2^2\right] \quad (233)$$

where $f_1(y_1,y_2)=\frac{\partial f(y_1,y_2)}{\partial y_1}$, $f_{11}(y_1,y_2)=\frac{\partial^2 f(y_1,y_2)}{(\partial y_1)^2}$, etc. and $f_{12}(y_1,y_2)=\frac{\partial^2 f(y_1,y_2)}{\partial y_1\partial y_2}$, etc.

Proposition 4.2: If the data is characterized by (217), and $\widehat{f(y)}$ is estimated using a kernel which satisfies (A2), then the $Var(\widehat{f(y)})$ upto $O(\max\left\{\frac{1}{nh},h^4\right\})$ is

$$Var(\widehat{f(y)})=\frac{1}{nh}\int_{\psi_{11}}K^2(\psi_{11})f(y)d\psi_{11}$$
$$-\frac{1}{k}\left\{(f(y))^2+\frac{h^4(f''(y))^2\mu_2^2}{4}+f(y)h^2f''(y)\mu_2+\frac{f(y)h^3f'''(y)\mu_3}{3}+\frac{f(y)h^4f''''(y)\mu_4}{12}\right\}$$

$$+\left(\frac{1-\rho^2}{k}\right)\left[f(y,y)+\frac{h^4}{4}f_{1122}(y,y)\mu_2^2+h^2f_{11}(y,y)\mu_2+\frac{h^3}{3}f_{111}(y,y)\mu_3+\frac{h^4}{12}f_{1111}(y,y)\mu_4\right]. \tag{234}$$

Proof: combine equations (226), (228), and (233).

The kernel density estimator will be consistent if $k \to \infty$ as $n \to \infty$. This is a reasonable assumption, as it characterizes the way that sampling is done for most economic surveys. Average cluster sizes tend to be fairly small (10-12 elements per cluster) while increases in sample size are normally achieved by increasing k, the number of clusters sampled.

Corollary 4.2: If the data is characterized by (217) where $\rho = 0$, then the $Var(\widehat{f(y)})$ upto $O(\frac{1}{nh})$ is

$$Var(\widehat{f(y)}) = \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})\, f(y) d\psi_{11}. \tag{235}$$

Proof: if the data is independent, then $f(y,y) = f(y)f(y)$, $f_{1122}(y,y) = f''(y)f''(y)$, $f_{11}(y,y) = f''(y)f(y)$, $f_{111}(y,y) = f'''(y)f(y)$, and $f_{1111}(y,y) = f''''(y)f(y)$ and the second and third terms in equation (234) will cancel out.

Note that (235) is simply the variance of the kernel estimate of $f(y)$ for the i.i.d. case (see Silverman, p.40).

The window width which minimizes the AIMSE of $\widehat{f(y)}$ in the i.i.d. case (218) will no longer be optimal in the case of correlated data. If we use the method of minimizing the approximate integrated mean squared error using equations (234) and (221), the resulting solution will be a complex polynomial in h, which will include

terms containing μ_2, μ_3, and μ_4. (Hall et. al. (1991) consider a similar problem where the AIMSE is also a 7th-degree polynomial in h. They provide an optimal h which is asymptotically equivalent to the implicit minimizer of the 7th-degree polynomial.)

The solution pursued here is to choose a higher-order kernel. Higher-order kernels have been used to reduce bias in kernel density estimation (see Pagan and Ullah (1997), chapter 2, section 4.3). By choosing a fourth-order kernel, terms involving μ_2 and μ_3 will be zero. Minimizing the approximate integrated mean squared error will then yield a simple solution for the optimal h upto the order of approximation considered. The proposed optimal h below is exactly that window width which minimizes the integrated mean squared error (to the order considered) and not simply an asymptotic equivalent.

Replace assumption (A2) with the following:

(A2)' (i) $\int K(\psi)d\psi = 1$

(ii) $\int \psi K(\psi)d\psi = 0$

(iii) $\int \psi^2 K(\psi)d\psi = 0$

(iv) $\int \psi^3 K(\psi)d\psi = 0$

(v) $\int \psi^4 K(\psi)d\psi = \mu_4 > 0$

Here, the fourth moment of the kernel is required to be positive instead of the usual, less restrictive assumption that μ_4 be finite. Since μ_4 appears below in the expression for the optimal window width (239) and is raised to a fractional power, it must be positive in order to have a reasonable (i.e. a positive real number) window

width.

We can construct a fourth-order kernel which meets assumption (A2)' and is based upon the standard normal kernel ($K^{*}(\psi)$) by assigning a value for μ_4. Here we simply set $\mu_4 = 1$. The resulting kernel is

$$\left(2 - \frac{3}{2}\psi^2 + \frac{1}{6}\psi^4\right) K^{*}(\psi)\,. \tag{236}$$

Figure 4.13 presents a graph of this kernel and the standard normal kernel.

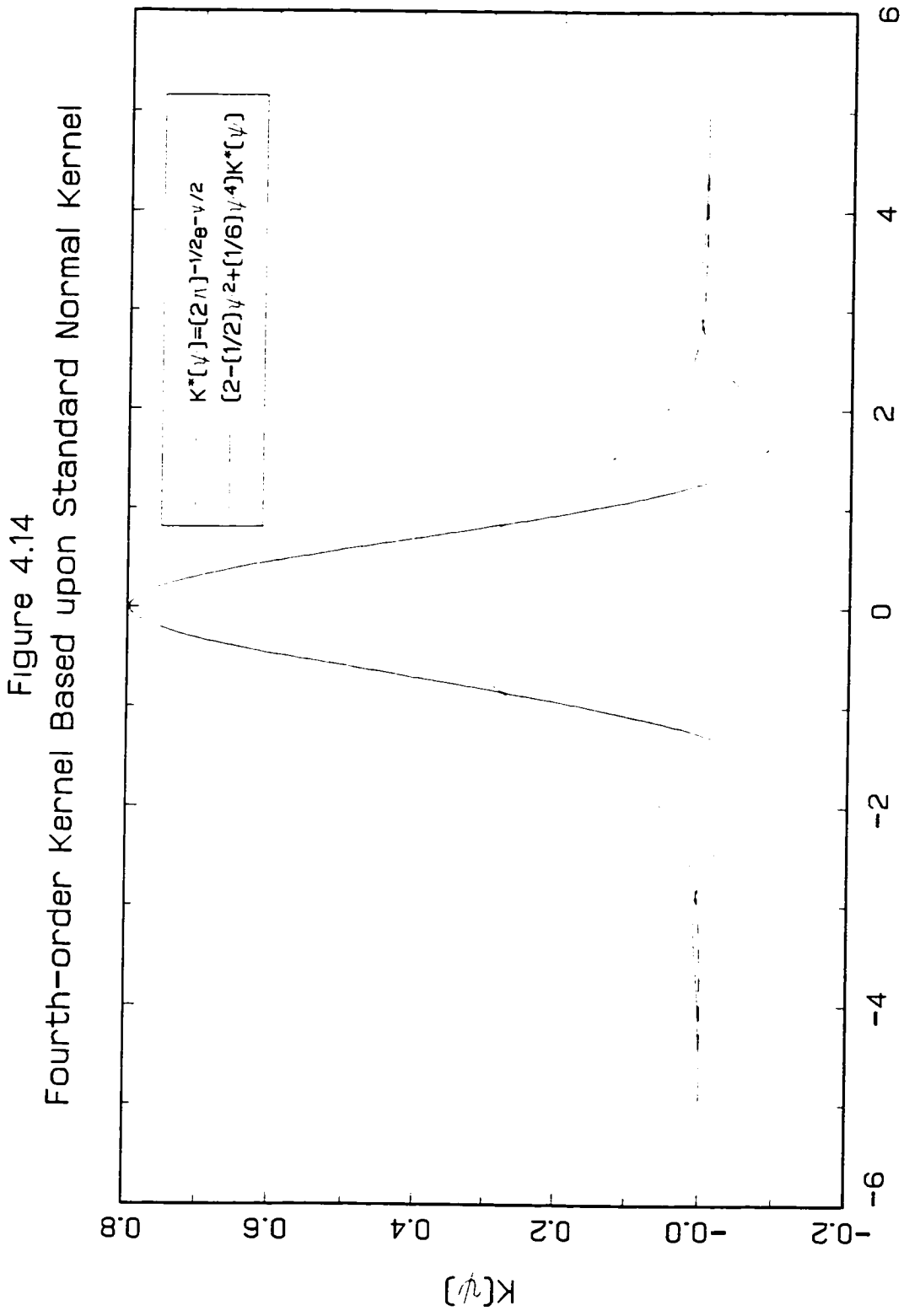


Figure 4.14
Fourth-order Kernel Based upon Standard Normal Kernel

Now, the

$$Var(\widehat{f(y)}) = \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})\, f(y) d\psi_{11} - \frac{1}{k}\left\{(f(y))^2 + \frac{h^4}{12} f''''(y) f(y) \mu_4\right\} + \left(\frac{1-\rho^2}{k}\right)\left[f(y,y) + \frac{h^4}{12} f_{1111}(y,y)\mu_4\right]. \quad (237)$$

The integrated mean-squared error (IMSE) is

$$\int_y \left\{(bias(\widehat{f(y)}))^2 + Var(\widehat{f(y)})\right\} dy. \quad (238)$$

Since we are using a higher-order kernel, this bias will be of $O(h^4)$ instead of having the form in (221). $(bias(\widehat{f(y)}))^2$ will thus be $O(h^8)$, and the approximate integrated mean squared error upto $O(\frac{1}{nh})$ will be equivalent to the integrated variance.

Corollary 4.3: Given (217), and (A2)', the optimal window width, h_{opt} will be

$$h_{opt} = \left[\int_\psi K^2(\psi) d\psi\right]^{\frac{1}{5}} \left[\frac{n\mu_4}{3k}\right]^{\frac{-1}{5}} \left[\int_y \left(\left(1-\rho^2\right) f_{1111}(y,y) - f''''(y) f(y)\right) dy\right]^{\frac{-1}{5}}. \quad (239)$$

Proof: Use (237) and (238) and minimize the expression for the AIMSE with respect to h and rearrange to solve for h_{opt}.

As in the i.i.d. case, the optimal window width is proportional to $n^{\frac{-1}{5}}$ and will depend upon both the kernel and the true, underlying density of the data. Analogous to the i.i.d. case, we consider the case where (y_1, y_2) is distributed as a bivariate normal and we choose the fourth-order kernel of (236) allowing us to give an exact

value to h_{opt}. For this special case, we have

$$h = \left[\frac{467}{384\sqrt{\pi}}\right]^{\frac{1}{5}} \left[\frac{n}{3k}\right]^{\frac{-1}{5}} \left(\sigma^{-5} \int_z \left\{\left(1-\rho^2\right)\phi^*_{1111}(z_1, z_2; \rho) - \phi''''(z) \cdot \phi(z)\right\} dz\right)^{\frac{-1}{5}} \tag{240}$$

where ϕ is the standard normal distribution and ϕ^* is the standard normal bivariate with correlation ρ (see Morrison, p. 86). Rearranging, we can write the optimal window width as

$$h = \kappa \sigma n^{-\frac{1}{5}} \tag{241}$$

where

$$\kappa = \left[\frac{467}{128\sqrt{\pi}}\right]^{\frac{1}{5}} [k]^{\frac{1}{5}} \left(\Phi(\rho) - \frac{3}{8\sqrt{\pi}}\right)^{\frac{-1}{5}} \tag{242}$$

and $\Phi(\rho) = \int_z (1-\rho^2)\phi^*_{1111}(z_1, z_2; \rho) dz$. Unlike the case of i.i.d. data where y is normally distributed, κ will no longer be constant, but will depend upon both the number of clusters, k, and the intra-cluster correlation coefficient, ρ.

As the cluster size increases, the window width will increase for a given size sample, n. If we re-write (22) as

$$h = \tilde{\kappa} \sigma \left(\frac{n}{k}\right)^{-1/5} \tag{243}$$

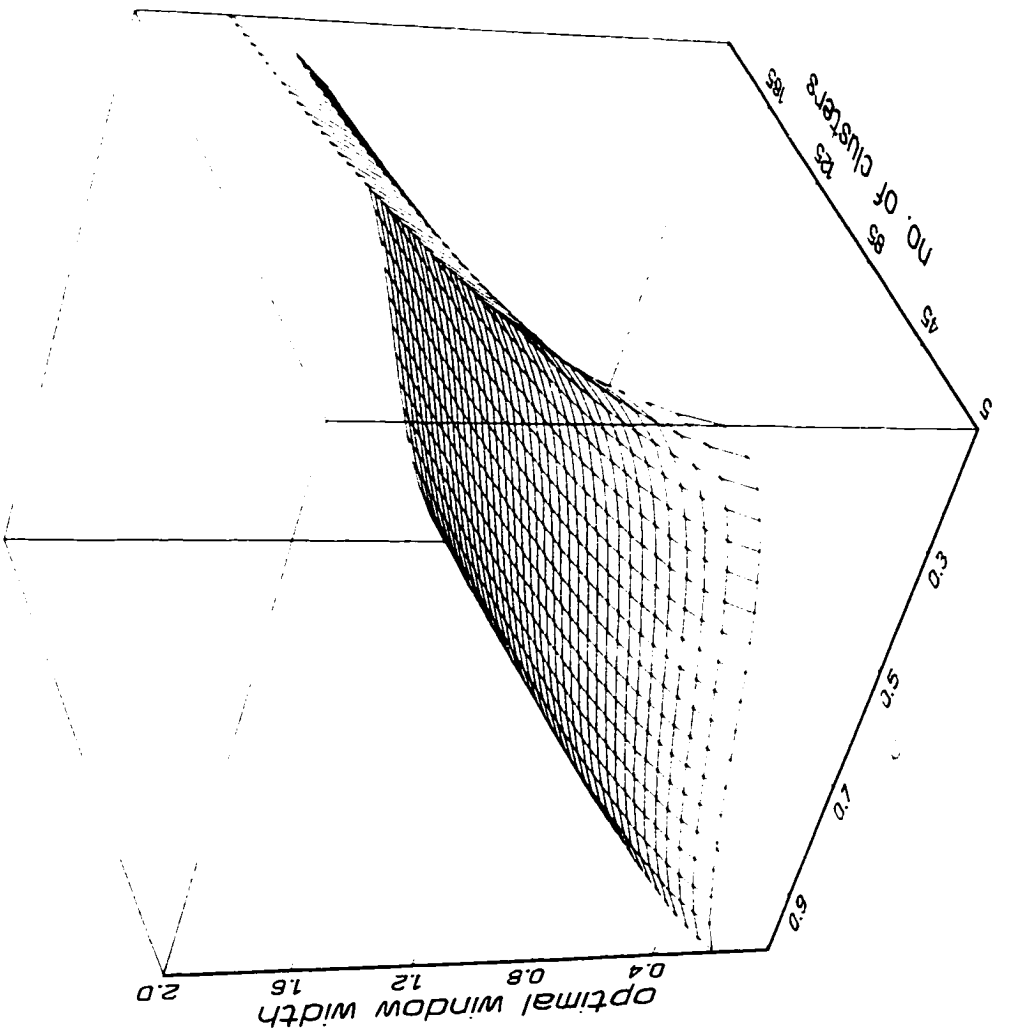
where

$$\tilde{\kappa} = \left[\frac{467}{128\sqrt{\pi}}\right]^{\frac{1}{5}} \left(\Phi(\rho) - \frac{3}{8\sqrt{\pi}}\right)^{\frac{-1}{5}} \tag{244}$$

we can see that an increase in the number of clusters (or a decrease in the average cluster size) acts as a decrease in the "effective" sample size, $\left(\frac{n}{k}\right)$.

It can be shown that $\Phi(\rho)$ is increasing in ρ and therefore κ will be decreasing in ρ. Intuitively, as the intra-cluster correlation coefficient increases, data will be clustered more tightly together, and thus a finer window width will be optimal. For the case where $f(y, y)$ is bivariate normal with $\sigma = 1$ and $n = 1000$ Figure 4.15 shows the effect on the optimal window width of changing ρ and k simultaneously. Figure 4.16 provides a cross-sectional view of Figure 4.15 for the case where $n = 1000$ and $k = 100$. Here we can see quite clearly the effect of the optimal bandwidth constant decreasing as the data gets packed more tightly together (i.e. as the cross-correlation coefficient increases.) Table 4.3 gives values of $\tilde{\kappa}$ for a range of ρ values. Table 4.4 shows how $\Phi(\rho)$ is increasing in ρ.

Figure 4.15
Optimal Window Width for Different Cluster Sizes
and Intra-cluster Correlation Coefficients



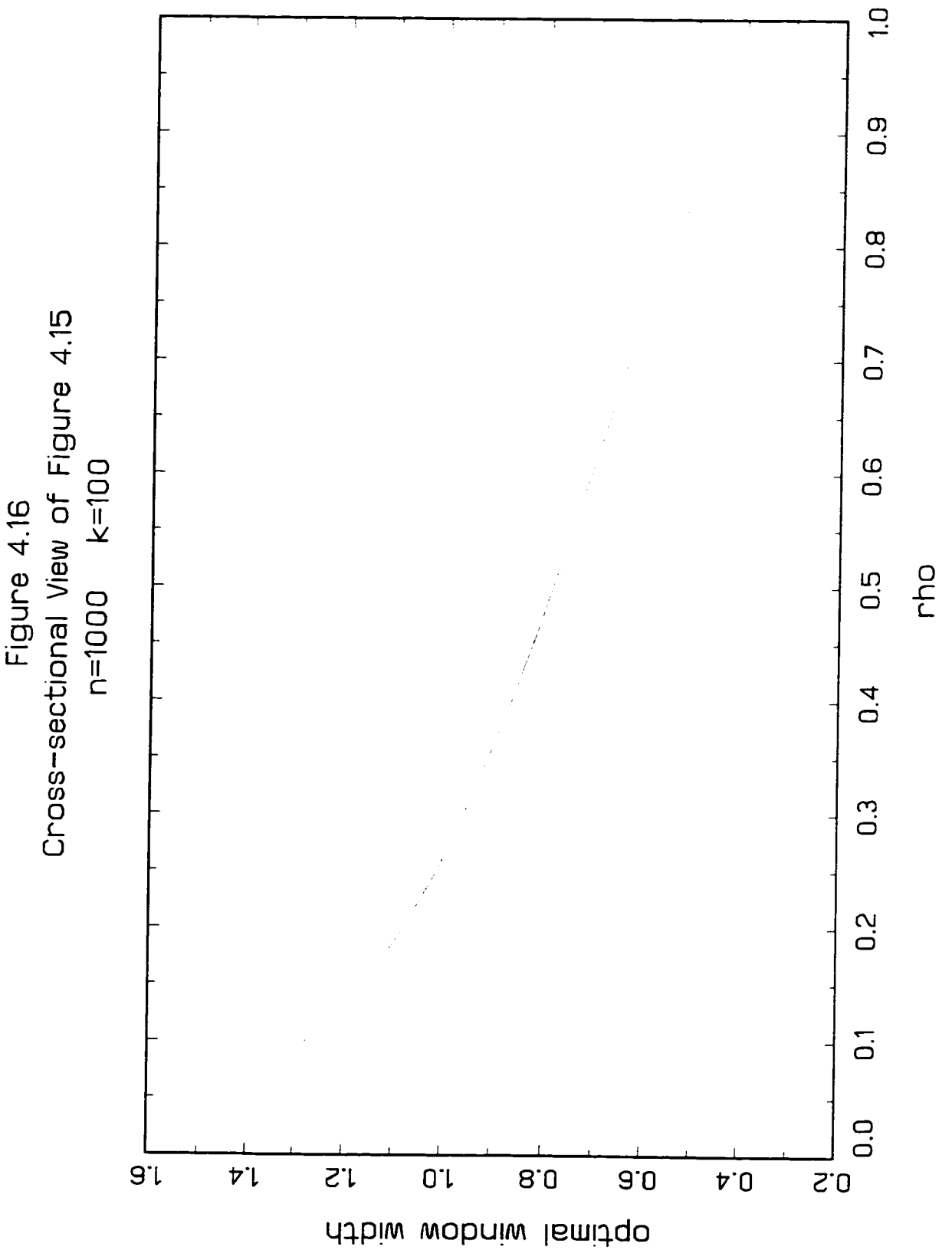


Figure 4.16
Cross-sectional View of Figure 4.15
n=1000 k=100

Table 4.3
Optimal Constant for Window Width
under Clustered Sampling

ρ	$\tilde{\kappa}$
0.05	2.3561866
0.10	2.0213222
0.15	1.835119
0.20	1.7040834
0.25	1.6011561
0.30	1.5148296
0.35	1.4391234
0.40	1.3704895
0.45	1.3065959
0.50	1.2457628
0.55	1.1866615
0.60	1.1281282
0.65	1.0690206
0.70	1.0080711
0.75	0.94368371
0.80	0.87356777
0.85	0.79392504
0.90	0.69714355
0.95	0.56208067

Entries in table represent $\tilde{\kappa}$ in equation (141) for different values of ρ.

Table 4.4

Values of $\Phi(\rho)$ and $\left(\Phi(\rho)-\frac{3}{8\sqrt{\pi}}\right)^{-1/5}$ for Various Values of ρ

ρ	$\phi_{1111}(z,z)$	$(1-\rho^2)$	$\Phi(\rho)$	$\Phi(\rho)-\frac{3}{8\sqrt{\pi}}$	$\left(\Phi(\rho)-\frac{3}{8\sqrt{\pi}}\right)^{-1/5}$
0.00	0.21157109	1.0000	0.21157109	0	0
0.05	0.24051785	0.9975	0.23991656	0.028345464	2.0394041
0.10	0.2753278	0.9900	0.27257452	0.061003424	1.7495613
0.15	0.31762086	0.9775	0.31047439	0.098903297	1.5883926
0.20	0.36959949	0.9600	0.35481551	0.14324442	1.4749743
0.25	0.43431334	0.9375	0.40716876	0.19559767	1.3858852
0.30	0.51607311	0.9100	0.46962653	0.25805544	1.3111651
0.35	0.62111632	0.8775	0.54502957	0.33345848	1.2456374
0.40	0.75871419	0.8400	0.63731992	0.42574882	1.1862311
0.45	0.94308233	0.7975	0.75210816	0.54053706	1.1309279
0.50	1.1968268	0.7500	0.89762013	0.68604904	1.0782736
0.55	1.5574892	0.6975	1.0863487	0.87477763	1.0271183
0.60	2.0907704	0.6400	1.3380931	1.126522	0.97645462
0.65	2.9193504	0.5775	1.6859248	1.4743537	0.92529382
0.70	4.2919356	0.5100	2.1888871	1.9773161	0.87253887
0.75	6.770275	0.4375	2.9619953	2.7504242	0.81680815
0.80	11.827184	0.3600	4.2577861	4.046215	0.7561191
0.85	24.278854	0.2775	6.737382	6.5258109	0.6871841
0.90	66.904654	0.1900	12.711884	12.500313	0.60341461
0.95	378.46988	0.0975	36.900813	36.689242	0.48651054

In practice, ρ can be replaced by a consistent estimator, $\widehat{\rho}$:

$$\widehat{\rho} = \frac{\sum_{c=1}^{k} \sum_{i=1}^{n_c} \sum_{j \neq i}^{n_c} (y_{ci} - \overline{y})(y_{cj} - \overline{y})}{\widehat{\sigma}^2 \sum_{c=1}^{k} n_c(n_c - 1)} \tag{245}$$

and the optimal window width can then be calculated from (241)and (242). A computer program is available from the author for exact calculation of $\Phi(\rho)$. Alternately, the appropriate value from Table 4.3 could be used to determine $\widetilde{\kappa}$ and plugged into (243).

For the case of unbalanced clusters, we can replace $[k]^{\frac{1}{5}}$ in (241) with $\left[\widetilde{k}\right]^{-\frac{1}{5}}$ where

$$\widetilde{k} = \sum \left(\frac{n_c}{n}\right)^2. \tag{246}$$

It is easy to ascertain that $\widetilde{k} = \frac{1}{k}$ when clusters are balanced.

Fan and Marron (1992) posit that higher-order kernels have not seen much use in application because of the unclear meaning of negative weights which higher-order kernels give to some data points and because the gains from using higher-order kernels are negligible for most sample sizes. In the case of clustering, however, higher-order kernels provide a simple way to solve for the optimal window width and allow for kernel density estimation which is easy to implement and is similar to the i.i.d. case.

4.3.1 Numerical Properties of h_{opt}: Clustered Sampling

What are the gains in efficiency from using the optimal window width, h_{opt}, of (241)? We can compare the mean squared error of $\hat{f}(y)$ using h_{opt} with two possible alternatives, $h^* = 1.06\sigma n^{-\frac{1}{5}}$ and $h^{**} = cn^{-\frac{1}{9}}$. h^* is the optimal window width for the univariate density estimation problem when the underlying density is normal and the kernel is Gaussian. h^{**} is the optimal window width in the i.i.d. case when a 4th-order kernel is used. We will set

$$c = 1.44$$

which is the optimal proportionality constant given the kernel of (236) and a true, underlying distribution that is normal (but ignoring, of course, the dependence in the data.)[18]

The AIMSE is calculated upto $O(\frac{1}{nh})$ using the kernel in (236), the standardized bivariate normal distribution, and the three window widths, h_{opt}, h^*, and h^{**}. The AIMSE upto $O\left(\frac{1}{nh}\right)$ will be

$$AIMSE = \frac{1}{nh}\int_{\psi} K^2(\psi)d\psi + \frac{1}{k}\sigma^{-1}\int_{z}\left\{\left(1-\rho^2\right)\phi^*(z,z) - (\phi(z))^2\right\}dz$$

$$+\frac{h^4}{12k}\sigma^{-5}\int_{z}\left\{\left(1-\rho^2\right)\phi^*_{1111}(z,z) - \phi''''(z)\phi(z)\right\}dz. \qquad (247)$$

[18] In general, for an r-th order kernel, $h^{**} = \left(\frac{\lambda_2(r!)^2}{2r\lambda_{1r}}\right)^{\frac{1}{2r+1}} n^{-\frac{1}{2r+1}}$ where $\lambda_{1r} = \mu_r^2\int_x (f^r(x))^2\,dx$ and $\lambda_2 = \int_{\psi} K^2(\psi)d\psi$. $\lambda_{1r} = \frac{105}{32\sqrt{\pi}}$ for this density.

Specifying a reference distribution allows exact calculation of the approximate integrated mean squared error. Results are given in Table 4.5.

h_{opt} outperforms both h^* and h^{**} in the integrated mean squared error sense–not surprising given that it is chosen to minimize the AIMSE. The last two columns of Table 4.5 show that the gains in approximate IMSE calculated from (247) are quite substantial when compared to h^*–generally on the order of 50%. The gains from using h_{opt} compared to h^{**} are somewhat smaller, but using h_{opt} provides a lower mean squared error. The gains in integrated mean squared error in this case may be even greater depending upon the values of n and k chosen.

Table 4.5
Difference in IMSE between h_{opt} and h* and h**

ρ	h_{opt}	$h^* = 1.06\sigma n^{-1/5}$	$h^{**} = 1.44\sigma n^{-1/5}$	IMSE (h_{opt})	IMSE (h*)	IMSE (h**)	$\frac{IMSE(h_{opt})}{IMSE(h^*)}$	$\frac{IMSE(h_{opt})}{IMSE(h^{**})}$
.05	1.486	0.2663	0.6686	0.000643	0.002643	0.001097	0.243260	0.586081
.10	1.275	0.2663	0.6686	0.000795	0.002700	0.001159	0.294566	0.686057
.15	1.157	0.2663	0.6686	0.000911	0.002747	0.001213	0.331480	0.750947
.20	1.075	0.2663	0.6686	0.001004	0.002784	0.001257	0.360763	0.799154
.25	1.010	0.2663	0.6686	0.001082	0.002811	0.001292	0.384895	0.837503
.30	0.956	0.2663	0.6686	0.001145	0.002825	0.001317	0.405131	0.869422
.35	0.908	0.2663	0.6686	0.001194	0.002828	0.001331	0.422223	0.896890
.40	0.865	0.2663	0.6686	0.001230	0.002817	0.001335	0.436661	0.921129
.45	0.824	0.2663	0.6686	0.001253	0.002792	0.001329	0.448786	0.942859
.50	0.786	0.2663	0.6686	0.001262	0.002751	0.001312	0.458852	0.962372
.55	0.749	0.2663	0.6686	0.001258	0.002693	0.001284	0.467057	0.979417
.60	0.712	0.2663	0.6686	0.001239	0.002615	0.001247	0.473589	0.992858
.65	0.675	0.2663	0.6686	0.001204	0.002516	0.001204	0.478679	0.999836
.70	0.636	0.2663	0.6686	0.001154	0.002391	0.001161	0.482718	0.993924
.75	0.595	0.2663	0.6686	0.001088	0.002236	0.001132	0.486536	0.961323
.80	0.551	0.2663	0.6686	0.001006	0.002044	0.001150	0.492192	0.874860
.85	0.501	0.2663	0.6686	0.000912	0.001805	0.001313	0.505621	0.694891
.90	0.440	0.2663	0.6686	0.000824	0.001503	0.001981	0.548011	0.415765
.95	0.355	0.2663	0.6686	0.000827	0.001140	0.005544	0.726027	0.149258

n=1000, k=100, average cluster size=10, s=1

Figures 4.17 through 4.19 provide an illustration of the difference in density estimates from using h_{opt} as opposed to h^*. A detailed simulation study was conducted by the author and the results are given here for three values ρ from .2 to .6.

The see how the simulation was conducted, first re-write the model to capture assumptions (217) as follows:

$$y_{ci} = \mu + u_c + \varepsilon_{ci} \tag{248}$$

where $u_c \tilde{} D(0, \sigma_c^2)$ is an effect common to all elements in cluster c and $\varepsilon_{ci} \tilde{} D(0, \sigma_\varepsilon^2)$ is an idiosyncratic error term with $cov(\mu_c, \varepsilon_{ci}) = 0$ for all i=1, . . ., n and c=1, . . ., k. The element variance will then be $\sigma_c^2 + \sigma_\varepsilon^2$, and the intra-cluster correlation coefficient ρ will equal $\frac{\sigma_c^2}{\sigma_c^2+\sigma_\varepsilon^2}$. Data was chosen in repeated simulation from a normal distribution for both the cluster-specific and the idiosyncratic errors. σ_c^2 and σ_ε^2 were fixed so that the total element variance equals one. This allows data with different degrees of correlation to be generated using any simple random number generating method. For the simulation, the cluster size was set at 500 and the total sample size at 10,000. (Thus for each simulation, 10,000 numbers were drawn from a $N(0, \sigma_\varepsilon^2)$ and 500 from a $N(0, \sigma_c^2)$. All elements in the same cluster share the same draw of the cluster-specific error term.) Average cluster size is 20 and since clusters were chosen to be balanced, the clusters are all the same size.

Figures 4.17 through 4.19 are typical realizations of this simulation exercise. The

density estimates using h^* tend to under-smooth the data as can be seen by the spurious variation in the density estimates. The standard normal distribution (the marginal distribution of the true density) is shown for reference.

For cluster data with values of ρ greater than .15, the suggested kernel (236) and h_{opt} (241) perform very well in simulation. For small values of ρ this combination tends to over-smooth the data, and it is probably best to use the second-order Gaussian kernel and the usual window width. This needs further investigation.

The above kernel and optimal window width give a method for applying kernel density estimation for clustered data in a way that takes into account the dependence in the data. As in the i.i.d. case with a second-order kernel, the optimal window width is proportional to $n^{-\frac{1}{5}}$, however a different proportionality constant is now required to minimize the AIMSE. This optimal window width and fourth-order kernel perform well for the levels of correlation commonly found in cross-sectional, survey data.

Figure 4.17
Density Estimation for Clustered Data (ρ=.2)
h^* vs. h_{opt}

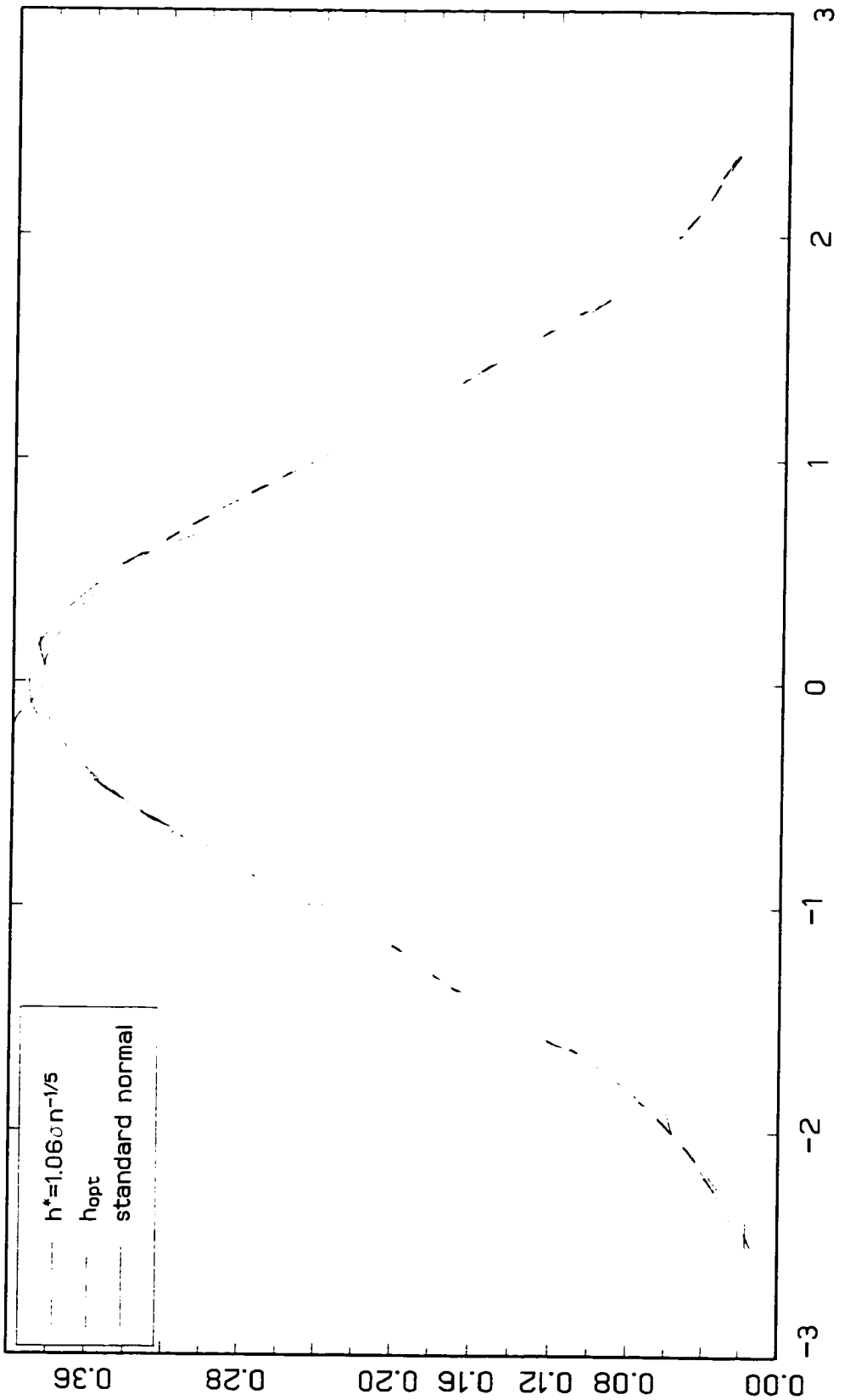


Figure 4.18
Density Estimation for Clustered Data (ρ=.4)
h^* vs. h_{opt}

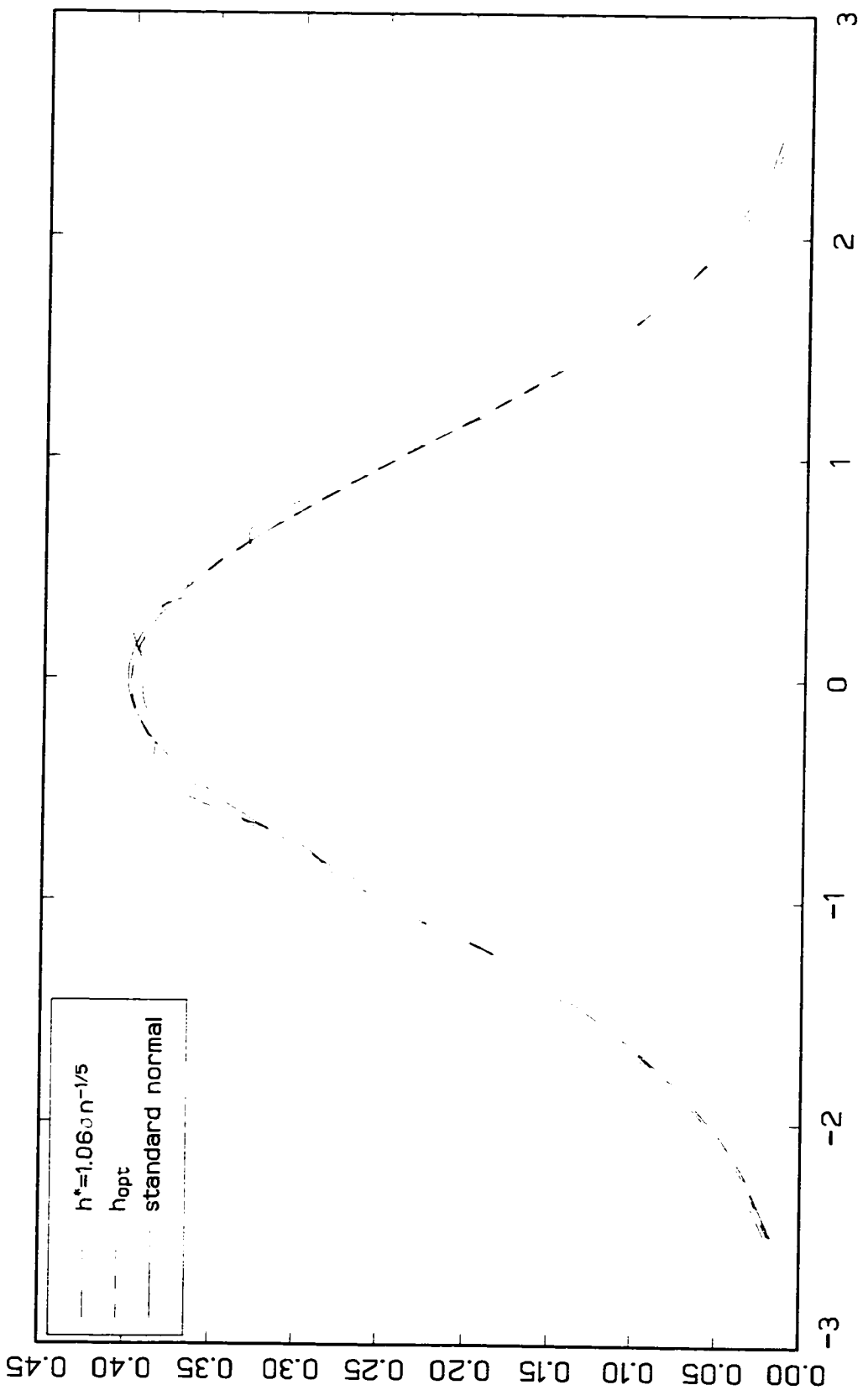
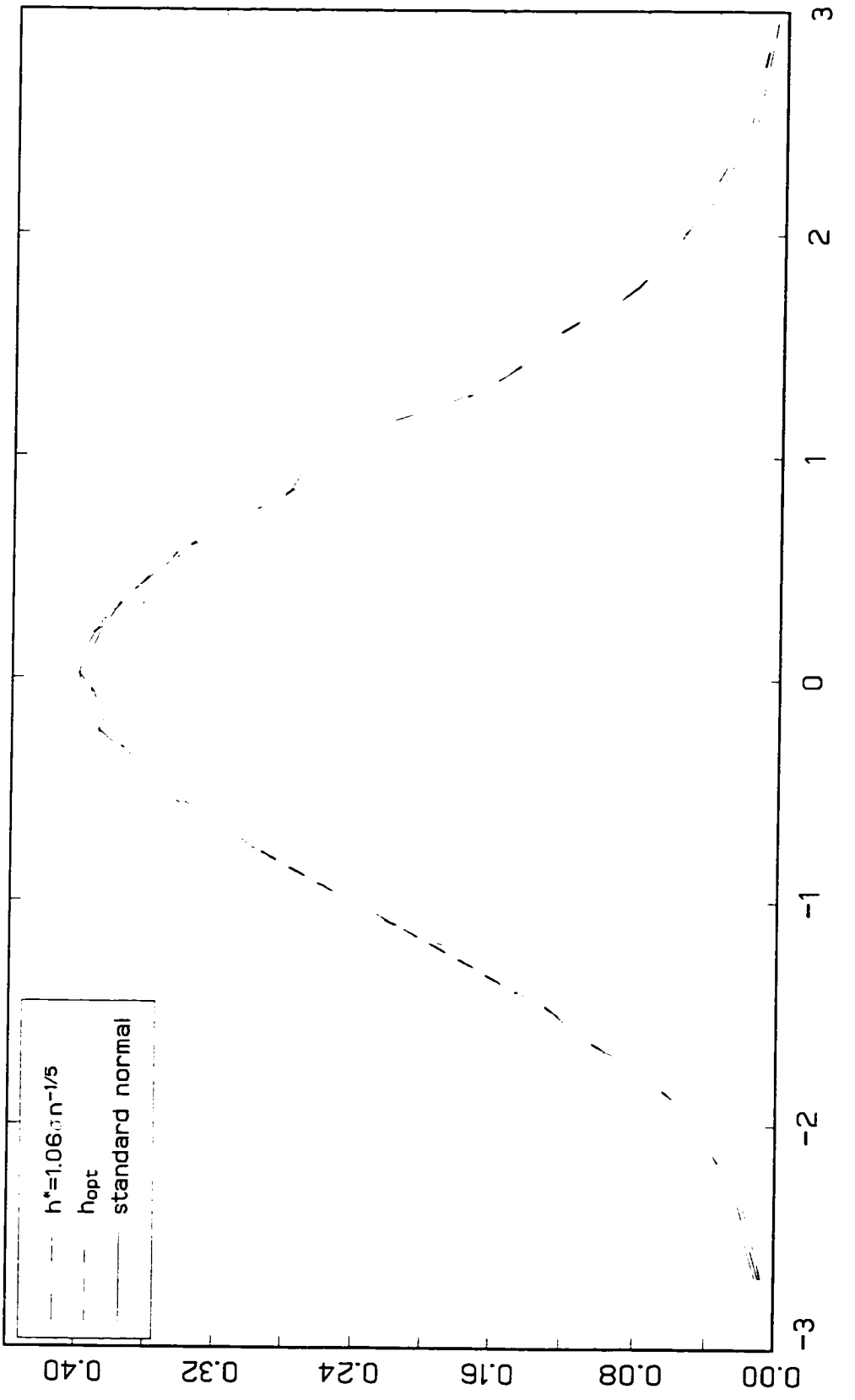


Figure 4.19
Density Estimation for Clustered Data (ρ=.6)
h^* vs. h_{opt}



4.4 Conclusion

Non-parametric density estimation is a useful tool to provide a visual image of the economic data under consideration. The use of non-parametric techniques has spread widely in econometric analysis in the past decade. Much of the analysis has been conducted using data gathered in surveys. As we have seen in this section, ignorning the stratification and clustering in the data can lead to biased and inefficient density estimation. These same results will carry over to the regression case. Exploring the exact nature of the problem for non-parametric regression is a promising area of future work.

Here, for the density estimation case, we have shown why the usual optimal window width will no longer be optimal and we have suggested data-based methods for choosing a new optimal window width for the case of stratification and for the case of clustering. In addition, we have proposed a weighted, non-parametric density estimator for the case of unequal probability, stratified sampling. Through simulation we have shown the huge bias which may result from ignoring the stratified nature of the data.

For non-parametric density estimation, when data is both stratified and clustered, the results of sections 4.2 and 4.3 need to be combined. For approximately normal data in the case of stratified and clustered sampling, we can estimate each stratum separately, using a fourth order kernel with appropriate modification to the window

width (213) for the *ith* stratum

$$h_i = \tilde{\kappa}\,(\rho)\left(\frac{n_i}{k_i}\right)^{-1/5} \tag{249}$$

where

$$\tilde{\kappa} = \left[\frac{467}{128\sqrt{\pi}}\right]^{\frac{1}{5}}\left(\Phi(\rho) - \frac{3}{8\sqrt{\pi}}\right)^{\frac{-1}{5}}. \tag{250}$$

Values of $\tilde{\kappa}$ for different ρ are given in section 4.3. ρ may be estimated using (70) and is generally assumed to be constant for all clusters in the sample, following the same methodology presented for dealing with clustered data throughout this paper.

References

[1] Atkinson, A. (1970) On the Measurement of Inequality, *Journal of Economic Theory*, 2: 244-263.

[2] Beach, C. M., R. Davidson, and G. A. Slotsve (1994), Distribution Free Statistical Inference for Inequality Dominance with Crossing Lorenz Curves, manuscript, Vanderbilt University, Nashville.

[3] Bellhouse, D.R. (1988a) A Brief History of Random Sampling Methods, *Handbook of Statistics, Volume 6*, edited by P. R. Krishnaiah and C.R. Rao., Elsevier Science Publishers, Amsterdam.

[4] Bellhouse, D.R. (1988b) Systematic Sampling, *Handbook of Statistics, Volume 6*, edited by P. R. Krishnaiah and C.R. Rao., Elsevier Science Publishers, Amsterdam.

[5] Blackorby, C. and D. Donaldson (1978) Measures of Relative Equality and Their Meaning in Terms of Social Welfare, *Journal of Economic Theory*, 18: 59-80.

[6] Bourguignon, F. (1979) Decomposable Income Inequality Measures, *Econometrica*, 47:901-920.

[7] Bowley, A. L. (1907) *Elements of Statistics, 3rd edition.* King and Son, London.

[8] Bowley, A. L. (1913) Working-class households in Reading, *Journal of the Royal Statistical Society*, 76: 672-701.

[9] Bowley, A. L. (1926) Measurement of the precision attained in sampling, *Bulletin of the International Statistical Institute*, 22: Supplement to Liv. 1, 6-62.

[10] Bowman, K.O. and L. R. Shenton (1981) Moment Series for the Coefficient of Variation in Weibull Sampling, *American Statistical Association–Proceedings of the Statistical Computing Section*: 148-153.

[11] Breunig, R. (1997a) Almost Unbiased Estimation of an Inequality Measure, manuscript, University of California, Riverside.

[12] Breunig, R. (1997b) Density Estimation for Clustered Data, manuscript, University of California, Riverside.

[13] Cochran, W. G. (1939) The Use of the Analysis of Variance in Enumeration by Sampling, *Journal of the American Statistical Association 34*: 492–510.

[14] Cochran, W. G. (1953) *Sampling Techniques*, John Wiley & Sons, New York.

[15] Cochran, W. G. (1978) Laplace's Ratio Estimator, *Contributions fo Survey Sampling and Applied Statistics*, edited by H.A. David, Academic Press, New York, 3-10.

[16] Coulter, P. B. (1989) *Measuring Inequality: A Methodological Handbook*. Westview Press, Boulder, CO.

[17] Cowell, F.A. (1989) Sampling Variance and Decomposable Inequality Measures, *Journal of Econometrics* 42: 27-41.

[18] Cowell, F. A. (1995) *Measuring Inequality*, 2nd edition, Prentice Hall Harvester Wheatsheaf, London.

[19] Cramer, H. (1946) *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

[20] Dalenius, T. (1988) A First Course in Survey Sampling, *Handbook of Statistics, Volume 6*, edited by P. R. Krishnaiah and C.R. Rao., Elsevier Science Publishers, Amsterdam.

[21] Dalton, H. (1920) The Measurement of the Inequality of Incomes, *Economic Journal*, 30: 348-361.

[22] Davidson, R. and J. MacKinnon (1993) *Estimation and Inference in Econometrics*, Oxford University Press, New York.

[23] Davies, J. B., D. A. Green, and H. J. Paarsch (1998) Economic Statistics and Social Welfare Comparisons: A Review. *Handbook of Applied Economic Statistics*, edited by D. Giles and A. Ullah. Marcel Dekker, Inc., New York.

[24] Deaton, A. (1997) *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Johns Hopkins University Press, Baltimore.

[25] Duncan, J. W. and W. C. Shelton (1978) *Revolution in U.S. Government Statistics, 1926-1976*. U.S. Department of Commerce, Washington.

[26] Efron, B. (1982) The Jackknife, the Bootstrap, and Other Resampling Plans, Philadelphia: Society for Industrial and Applied Mathematics.

[27] Engel, E. (1857) Die Productions-Und Consummations-Erhaltnisse des Konigreichs Sachsen, reproduced in *Die Lebenkosten belgischer Arbielter Fa* (E. Engel, ed.), Dresden, 1895.

[28] Evans, M., N. Hastings, and B. Peacock (1993) *Statistical Distributions*, 2nd edition. John Wiley and Sons, New York.

[29] Fan, J. and J.S. Marron (1992) "Best Possible Constant for Bandwidth Selection," *Annals of Statistics*, 20:4, 2057-2070.

[30] Federal Trade Commission (1997) Comments of the Staff of Federal Trade Commission, Food Labeling: Net Quantity of Contents; Compliance 21 C.F.R. Parts 101, 161 and 50, Federal Trade Commission, Washington.

[31] Fisher, R. A. (1925) *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.

[32] Foster, J., J. Greer and E. Thorbecke (1984) A Class of Decomposable Poverty Measures, *Econometrica*, 52: 761-766.

[33] Gastwirth, J. L. (1974) Large Sample Theory of Some Measures of Income Inequality, *Econometrica*, 42:191-196.

[34] Gastwirth, J. L. and M. H. Gail (1985) Simple Asymptotically Distribution-Free Methods for Comparing Lorenz Curves and Gini Indices Obtained from Complete Data. *Advances in Econometrics*, Volume 4: 229-243.

[35] Gastwirth, J. L., T. K. Nayak, and A. M. Krieger (1986) Large Sample Theory for the Bounds on the Gini and Related Indices of Inequality Estimated from Grouped Data, *Journal of Business and Economic Statistics*, 4:269-273.

[36] Gini, C. (1928) Une application de la méthode representative aux matériaux du dernier recensement de la population italienne (1er décembre 1921). *Bulletin of the International Statistical Institute*, 23: 198-215.

[37] Githinji, M. (1996), "Dualism and Income Distribution in Kenya," Ph.D. Manuscript, University of California Riverside.

[38] Godambe, V.P. (1952) A Unified Theory of Sampling from Finite Populations, *Journal of the Royal Statistical Society*, Series B, 17: 269-278.

[39] Godambe, V.P. (1976) A Historical Perspective of the Recent Developments in the Theory of Sampling from Actual Populations, *Dr. Panse Memorial Lecture*. Indian Society of Agricultural Statistics, New Dehli.

[40] Godambe, V.P. (1991), editor. *Estimating Functions*. Oxford University Press: Clarendon Press, New York.

[41] Godambe, V.P. (1995) Estimation of Parameters in Survey Sampling: Optimality, *Canadian Journal of Statistics*, 23: 227-243.

[42] Godambe, V.P. (1997) Estimation of Parameters in Survey Sampling, *Journal of the American Statistical Association*, forthcoming.

[43] Gourieroux, C. *Theorie des Sondages*, Economica, Paris.

[44] Greene, W.M. (1993) *Econometric Analysis*, 2nd edition, Macmillan , New York.

[45] Gupta, R. and S. Ma (1996) Testing the Equality of Coefficients of Variation in K Normal Populations, *Communications in Statistics–Theory and Methods*, 25: 115-132.

[46] Hall, P., S. N. Lahiri, and J. Polzehl (1995), "On Bandwidth Choice in Nonparametric Regression with Both Short- and Long-Range Dependent Errors," *Annals of Statistics*, 23:6, 1921-1936.

[47] Hall, P., S. J. Sheather, M.C. Jones, and J. S. Marron (1991), "On Optimal Data-Based Bandwidth Selection in Density Estimation," *Biometrika*, 78, 263-271.

[48] Hansen, M. H. (1987) History of Survey Sampling, *Statistical Science, Vol. 2:* 180–190.

[49] Hansen, M. H., W. G. Madow, and B. J. Tepping (1983) An Evaluation of Model-dependent and Probability-sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, 78:776-807.

[50] Hardle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press, New York.

[51] Herrmann, E., T. Gasser, and A. Kneip (1992), "Choice of Bandwidth for Kernel Regression when Residuals are Correlated," *Biometrika*, 79, 783-795.

[52] Hinkley, D. V. (1978), "Improving the Jackknife with Special Reference to Correlation Estimation," Biometrika, Volume 65, number 1, pp. 13-21.

[53] Howes, S. and J. O. Lanjouw (1994) Making Poverty Comparisons Taking into Account Survey Design: How and Why, manuscript, World Bank.

[54] Iachan, R. (1982) Systematic Sampling–A Critical Review, *International Statistical Review*, 50: 293-303.

[55] Isserlis, L. (1918) On the Value of a Mean as Calculated from a Sample, *Journal of the Royal Statistical Society*, 81: 75-81.

[56] Johnston, J. (1991) *Econometric Methods*, 3rd edition, McGraw-Hill, London.

[57] Jones, C.I. (1997) On the Evolution of the World Income Distribution, *Journal of Economic Perpsectives*, 11,3: 19-36.

[58] Kalton, G. (1983) Introduction to Survey Sampling, Sage University Paper #35.

[59] Kakwani, N. (1980) *Income Inequality and Poverty: Methods of Estimation and Policy Applications*, World Bank, Oxford University Press, Oxford.

[60] Kakwani, N. (1990) Large Sample Distributions of Several Inequality Measures: With Application to Cote d'Ivoire in *Contributions to Econometric Theory and Application: Essays in honour of A. L. Nagar*, edited by R.A.L. Carter, J. Dutta, and A. Ullah, Springer-Verlag, New York.

[61] Kendall, M. and L. R. Stuart (1977), The Advanced Theory of Statistics. Volume 1. Fourth edition. London: Charles Griffin & Company Ltd.

[62] Khan, A. R., K. Griffin, C. Riskin, and Z. Renwei (1991), "Household Income and Its Distribution in China." University of California Riverside, Working paper Series.

[63] Kiaer, A. N. (1897) The Representative Method of Statistical Surveys (translation 1976, original in Norwegian), *Kristiania Videnskabsselskabets Skrifter, Historisk-filosofiske Klasse*, 4: (34–56), Statistisk Sentralbyra, Oslo.

[64] Kish, L. (1965) *Survey Sampling*, John-Wiley, New York.

[65] Kuznets, S. (1963) Quantitative Aspects of the Economic Growth of Nations: Part VIII, Distribution of Income by Size, *Economic Development and Cultural Change*, 11:2.

[66] Levy, P.S. and S. Lemeshow (1991) *Sampling of Populations: Methods and Applications*, 2nd edition, John Wiley & Sons, New York.

[67] Maasoumi, E. (1991), guest editor. Measurement of Welfare and Inequality. Annals of Econometrics, Journal of Econometrics. Volume 50, numbers 1 and 2.

[68] Maasoumi, E. and H. Theil (1979), "The Effect of the Shape of the Income Distribution on Two Inequality Measures." Economics Letters, 4, 289-291.

[69] Mahalanobis, P. C. (1940) A Sample Survey of the Acreage Under Jute in Bengal, *Sankhyā*, 4: 511-530.

[70] McKay, A. T. (1932) Distributions of the Coefficient of Variation and the Extended t-distribution, *Journal of the Royal Statistical Society*, 95: 695-698.

[71] Morrison, D.F. (1976) *Multivariate Statistical Methods*. New York: McGraw-Hill.

[72] Neuts, M. (1982) On the Coefficient of Variation of Mixtures of Probability Distributions, *Communications in Statistics–Simulation and Computation 11:* 649–657.

[73] Neyman, J. (1938) Contribution to the Theory of Sampling Human Populations, *Journal of the American Statistical Association 33*: 101–116.

[74] 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 Royal Statistical Society* 97: 558–606.

[75] Nygård, F. (1981) Mätning av inkomstjämnhet-en studie av ett statistiskt operationaliseringsproblem. Meddelanden från ekonomisk-statsvetenskapliga fakulteten vid Åbo Akademi, Statistiska Institutionen, Series A: 163, Turku.

[76] Nygård, F. and A. Sandström (1981) *Measuring Income Inequality*, Stockholm Studies in Statistics 1, Almqvist and Wiksell International, Stockholm.

[77] Nygård, F. and A. Sandström (1985) The Estimation of the Gini and Entropy Inequality Parameters in Finite Populations, *Journal of Official Statistics*. 1:399-412.

[78] Pagan, A. and A. Ullah (1997) Nonparametric Econometrics, Manuscript, Australian National University, Australia.

[79] Parzen, E. (1962) On Estimation of a Probability Density Function and Mode, *Annals of Mathematical Statistics* 33: 1065-1076.

[80] Pudney, S. (1989) *Modelling Individual Choice: The Econometrics of Corners, Kinks, and Holes*, Basil Blackwell, Oxford.

[81] Rao, C.R. (1973) *Linear Statistical Inference and its Applications*. 2nd edition. New York: John Wiley & Sons.

[82] Ravallion, M. (1994) *Poverty Comparisons*, Harwood Academic Publishers, Langhorne, PA.

[83] Rawls, J. (1972) *A Theory of Justice*. Oxford University Press, London.

[84] Rosenblatt, M. (1956) Remarks on Some Nonparametric Estimates of Density Function, *Annals of Mathematical Statistics 27:* 832–837.

[85] Sandström, A.(1983) Estimating Income Inequality, Large Sample Inference in Finite Populations. Research Report 1983:5, Department of Statistics, University of Stockholm.

[86] Sen, A. (1992) *Inequality Reexamined*, Russell Sage Foundation, Oxford University Press, New York.

[87] Shorrocks, A.F. (1980) The Class of Additively Decomposable Inequality Measures, *Econometrica* 48:613-625.

[88] Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[89] Singh, M. (1973), "Behaviour of Sample Coefficient of Variation Drawn from Several Distributions" Sankhya: The Indian Journal of Statistics, 55, 65-76.

[90] Stephan, F.F. (1948) History of the Uses of Modern Sampling Procedures, *Journal of American Statistical Association 43* .

[91] Stigler, G.J. (1954) The Early History of Empirical Studies of Consumer Behavior, *Journal of Political Economy*, 42: 95-113.

[92] Stigler, S.M. (1986) *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press, Cambridge, MA.

[93] Sukhatme, P. V. and Sukhatme (1984) *Sampling Theory of Surveys with Applications*, Iowa State University Press, Ames, IA.

[94] Survey of Income and Program Participation, Users Guide (1991) U.S. Department of Commerce, Economics and Statistics Administration, Bureau of the Census, Washington, DC.

[95] Theil, H. (1967) *Economics and Information Theory*, North-Holland, Amsterdam.

[96] Thompson, S. (1992) *Sampling*, John Wiley & Sons, New York.

[97] Tschuprow, A. (1923) On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observation, *Metron* 2: 461–493, 646–680.

[98] Ullah, A. and R. Breunig (1996) On the Bias of the Standard Errors of the LS Residual and the Regression Coefficients Under Normal and Non-Normal Errors, *Econometric Theory, Problems and Solutions*, 12:868.

[99] Ullah, A. and R. Breunig (1998) Econometric Analysis in Complex Surveys, *Handbook of Applied Economic Statistics*, edited by D. Giles and A. Ullah, Marcel Dekker, Inc., New York.

[100] Ullah, A., R.A.L. Carter and V. K. Srivastava, (1984) "The Sampling Distribution of Shrinkage Estimators and their F-Ratios in the Regression Model." Journal of Econometrics, 25, 109–122.

[101] Warren, W. G. (1982) "On the Adequacy of the Chi-squared Approximation for the Coefficient of Variation." Communications in Statistics–Simulation and Computation, 11, 659–666.

[102] Yates, F. (1946) A Review of Recent Statistical Developments in Sampling and Sampling Surveys, *Journal of the Royal Statistical Society*, Series A, 109: 12–42.

[103] Yates, F. and I. Zacapony (1935) The Estimation of the Efficiency of Sampling with Special Reference to Sampling for Yield in Cereal Experiments, *Journal of Agricultural Science 25*: 543–577.

Appendix A

In order to derive the results given in Proposition 3.1 we first introduce, in matrix notation,

$$w = s^2 - \sigma^2 = \frac{1}{n-1}\hat{u}'\hat{u} - \sigma^2 = O_p(\frac{1}{\sqrt{n}}) \tag{i}$$

$$\bar{u} = \frac{e'u}{n} = O_p(\frac{1}{\sqrt{n}}) \tag{ii}$$

where u is an n x 1 vector of u_i satisfying (2.2), e is an n x 1 vector of unit elements and

$$M = I - \frac{ee'}{n} \tag{iii}$$

is an n x n idempotent matrix with tr (M) = n - 1. Then from (2.1) and (2.4)

$$\hat{\theta} = \frac{s^2}{\bar{y}^2} = \frac{(w+\sigma^2)}{\mu^2}[1 + \frac{\bar{u}}{\mu}]^{-2} \tag{iv}$$

Expanding the right hand side we get

$$\hat{\theta} - \theta = f_{-1/2} + f_{-1} + f_{-3/2} \tag{v}$$

where, denoting f_{-r} as a term of $O_p(n^{-r})$

$$\begin{aligned} f_{-1/2} &= \frac{w}{\mu^2} - \frac{2\theta}{\mu}\bar{u} \\ f_{-1} &= -\frac{2}{\mu^3}w\bar{u} + \frac{3\theta}{\mu^2}\bar{u}^2 \\ f_{-3/2} &= \frac{3w\bar{u}^2}{\mu^4} - \frac{4\theta\bar{u}^3}{\mu^3} \end{aligned} \tag{vi}$$

Thus, the bias of $\hat{\theta}$ to O (n^{-1}) is given,

$$Bias\ (\hat{\theta}) = E\ (f_{-1/2}) + E\ (f_{-1}) \tag{vii}$$

Now we state the following results:

$$\sigma^{-2}\ E[u'Bu] = trB \tag{viii}$$

$$\sigma^{-3}\ E[u'Buu] = \gamma_1(I * B)e \tag{ix}$$

$$\begin{aligned}
\sigma^{-4}\ E[u'Buuu'] &= \gamma_2(I * B) + (trB)I + 2B \\
\sigma^{-5}\ E\{u'Auu'Bu.u] &= \gamma_3(I * A * B)e + \gamma_1[(trA + 2A) \\
&\quad (I * B)e + (trB + 2B)(I * A)e + 4(I * AB)e] \\
\sigma^{-6}\ E[u'Au.u'Bu.uu'] &= \gamma_4(I * A * B) + \gamma_2\ [tr(A * B)I + tr\ A(I * B) \\
&\quad + trB(I * A) + 4(I * AB) + 2A(I * B) + \\
&\quad + 2B(I * A) + 2(I * B)A + 2(I * A)B] + \gamma_1^2[4(A * B) \\
&\quad + (I * A)ee'(I * B) + (I * B)ee'(I * A) \\
&\quad + 2I * \{A(I * B)\}ee' + 2I * \{B(I * A)\}ee'] \\
&\quad + 2(tr\ A)B + 4AB + 2(tr\ B)A + 4BA \\
&\quad + [2(tr\ AB) + tr(A)\ tr(B)]I.
\end{aligned}$$

where A and B are n x n symmetric matrixes with non-stochastic elements—see, e.g. Ullah et. al. (1984, p. 398).

Using (i), (ii), (viii), and (ix) it is easy to see that

$$\begin{aligned} Ef_{-1/2} &= 0 \\ Ef_{-1} &= \frac{1}{n}\theta^{3/2}(3\theta^{1/2} - 2\gamma_1) \end{aligned} \tag{xi}$$

which, when substituted in (vii) gives the result in Proposition 3.1.

To obtaining the result in Proposition 3.2, we write, from (v), upto $O(n^{-2})$,

$$E(\hat{\theta} - \theta)^2 = E\, f_{-1/2}^2 + Ef_{-1}^2 + 2Ef_{-1/2}f_{-3/2} + 2Ef_{-1/2}f_{-1}. \tag{xii}$$

It is easy to verify that

$$\begin{aligned} Ef_{-1/2}^2 &= \theta^2\left(\frac{\gamma_2}{n} + \frac{2}{n-1}\right) - 4\theta^{5/2}\frac{\gamma_1}{n} + \frac{4\theta^3}{n} \\ Ef_{-1}^2 + 2Ef_{-1/2}\, f_{-3/2} &= \frac{\theta^3}{n^2}\left(10\gamma_2\frac{n+3}{n-1} + \frac{20n}{n-1} - 96\theta^{1/2}\gamma_1 + 75\theta + 20\gamma_1^2\frac{(n+1)}{(n-1)}\right) \\ 2Ef_{-1/2}f_{-1} &= \frac{\theta^{5/2}}{n^2}\left(14\theta^{1/2}\gamma_2\left(\frac{n}{n-1}\right) - 4\gamma_3 - 16\gamma_1\frac{n}{n-1} - 12\theta\gamma_1\right) \end{aligned} \tag{xiii}$$

To prove the result in Proposition 3.3, it is necessary to expand the sample skewness coefficient as well as the coefficient of variation squared. Letting $v = \hat{\mu} - \mu = \frac{1}{n}\sum \hat{u}_i^3 - \mu$ and

$$\hat{\gamma}_1 - \gamma_1 = g_{-1/2} + g_{-1} + g_{-3/2} \tag{xiv}$$

the mean squared error of $\tilde{\theta}$ is

$$\begin{aligned} MSE\left(\tilde{\theta}\right) &= MSE\left(\hat{\theta}\right) + \frac{1}{n^2}\left\{9\theta^4 + 4\theta^3\gamma_1^2 - 12\gamma_1\theta^{7/2}\right\} \\ &\quad -\frac{6}{n}\theta E\left(\theta f_{-1/2} + \theta f_{-1} + 2\theta f_{-1/2}^2\right) - \frac{4}{n}\theta^{3/2}\gamma_1 E\left(f_{-1/2} + f_{-1}\right) \\ &\quad -\frac{6}{n}\theta^{1/2}\gamma_1 Ef_{-1/2} - \frac{4}{n}\theta^{3/2}Ef_{-1/2}g_{-1/2} + o\left(\frac{1}{n^2}\right) \end{aligned} \tag{xv}$$

and

$$Ef_{-1/2}g_{-1/2} = E\left[\frac{w}{\mu^2} - \frac{2\theta\overline{u}}{\mu}\right]\left[\frac{v}{\sigma^3} - \frac{3w\mu_3}{2\sigma^5}\right] \tag{xvi}$$

which upto $O(n^{-1})$, gives

$$Ef_{-1/2}g_{-1/2} = \frac{\theta}{n}\gamma_3 + \frac{6\theta}{n}\gamma_1 - \frac{2\theta^{\frac{3}{2}}}{n}\gamma_2 - \frac{3\theta}{n}\gamma_1\gamma_2 - \frac{3\theta}{n-1}\gamma_1 + \frac{3\theta^{\frac{3}{2}}}{n}\gamma_1^2. \tag{xvii}$$

Proposition 3.3 follows from straightforward algebra.

Appendix B

To calculate the variance of $\hat{\beta}_2 = \frac{1}{n}\sum y_i^2$, we first write

$$Var(\hat{\beta}_2) = E(\hat{\beta}_2 - E\hat{\beta}_2)^2. \tag{xviii}$$

Using $y_i = \mu + u_i$,

$$\begin{aligned} E\hat{\beta}_2 &= \frac{1}{n}E\sum y_i^2 \\ &= \frac{1}{n}E\sum_{i=1}^{n}(\mu + u_i)^2 \\ &= \frac{1}{n}E\sum_{i=1}^{n}\mu^2 + \frac{1}{n}\mu\sum_{i=1}^{n}Eu_i + \frac{1}{n}E\sum_{i=1}^{n}u_i^2 \\ &= \mu^2 + \sigma^2. \end{aligned} \tag{xix}$$

We note that this result holds for the case of random sampling with replacement, random sampling without replacement, and cluster sampling, since in all of those cases we assume that the data share a common mean and that the expected value of the error term is zero.

We thus have

$$\begin{aligned} Var(\hat{\beta}_2) &= E(\hat{\beta}_2 - \mu^2 - \sigma^2)^2 \\ &= E(\hat{\beta}_2)^2 + (\mu^2 + \sigma^2)^2 - 2E(\hat{\beta}_2)\left(\mu^2 + \sigma^2\right) \end{aligned} \tag{xx}$$

Since $E(\widehat{\beta}_2) = \mu^2 + \sigma^2$, this yields

$$Var(\widehat{\beta}_2) = E(\widehat{\beta}_2)^2 - (\mu^2 + \sigma^2)^2. \tag{xxi}$$

To solve for the expected value of $\widehat{\beta}_2^2$, write

$$\begin{aligned} E(\widehat{\beta}_2)^2 &= E\left(\frac{1}{n}\sum_{i=1}^{n}(\mu + u_i)^2\right)^2 \\ &= \frac{1}{n^2}E\left(\sum_{i=1}^{n}(\mu + u_i)^2\right)\left(\sum_{i=1}^{n}(\mu + u_i)^2\right) \\ &= \frac{1}{n^2}E\sum_{i=1}^{n}(\mu + u_i)^4 + \frac{1}{n^2}E\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}(\mu + u_i)^2(\mu + u_i)^2 \end{aligned} \tag{xxii}$$

which can be expanded to give

$$\begin{aligned} E(\widehat{\beta}_2)^2 = {} & \frac{1}{n^2}E\sum_{i=1}^{n}\left(\mu^4 + u_i^4 + 4\mu^3 u_i + 4\mu u_i^3 + 6\mu^2 u_i^2\right) \\ & + \frac{1}{n^2}E\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\left(\mu^4 + u_i^2 u_j^2 + 2\mu^3 u_i + 2\mu^3 u_j\right. \\ & \left. + 4\mu^2 u_i u_j + 2\mu u_i^2 u_j + 2\mu u_i u_j^2 + \mu^2 u_j^2 + \mu^2 u_i^2\right). \end{aligned} \tag{xxiii}$$

In the most general form, we can evaluate these expressions and write the following

$$\begin{aligned} E(\widehat{\beta}_2)^2 = {} & \frac{\mu^4}{n} + \frac{(\gamma_2 + 3)\sigma^4}{n} + \frac{4\mu\gamma_1\sigma^3}{n} + \frac{6\mu^2\sigma^2}{n} \\ & + \frac{\mu^4}{n}(n-1) + \frac{2\mu^2\sigma^2}{n}(n-1) \\ & + \frac{4(n-1)}{n}\left[\mu^2\sigma_{12} + \mu\sigma_{112} + \frac{1}{4}\sigma_{1122}\right]. \end{aligned} \tag{xxiv}$$

For the case of RSWR, the last line will be zero, and we have

$$\begin{aligned} Var_{RSWR}(\widehat{\beta}_2) &= \frac{\mu^4}{n} + \frac{(\gamma_2+3)\,\sigma^4}{n} + \frac{4\mu\gamma_1\sigma^3}{n} + \frac{6\mu^2\sigma^2}{n} \\ &\quad + \frac{\mu^4}{n}(n-1) + \frac{2\mu^2\sigma^2}{n}(n-1) - (\mu^2+\sigma^2)^2 \end{aligned} \tag{xxv}$$

which reduces to the expression in equation (102)

$$Var_{RSWR}(\widehat{\beta}_2) = \frac{1}{n}\left[(\gamma_2+2)\,\sigma^4 + 4\mu\gamma_1\sigma^3 + 4\mu^2\sigma^2\right]. \tag{xxvi}$$

When the sampling is without replacement,

$$\begin{aligned} \sigma_{12} &= \frac{-\sigma^2}{(N-1)} \\ \sigma_{112} &= \frac{-\gamma_1\sigma^3}{(N-1)} \\ \sigma_{1122} &= \frac{N-(\gamma_2+3)}{(N-1)}\sigma^4 \end{aligned} \tag{xxvii}$$

and then we can show that

$$\begin{aligned} Var_{RSWOR}(\widehat{\beta}_2) &= \frac{(\gamma_2+3)\,\sigma^4}{n} + \frac{Nn-N+(\gamma_2+3)-n\,(\gamma_2+3)-Nn+n}{n(N-1)}\sigma^4 \\ &\quad + \frac{4\mu^2\sigma^2}{n} + \frac{4\mu^2\sigma^2}{n}\frac{(1-n)}{(N-1)} \\ &\quad + \frac{4\mu\gamma_1\sigma^3}{n} + \frac{4\mu\gamma_1\sigma^3}{n}\frac{(1-n)}{(N-1)} \end{aligned} \tag{xxviii}$$

which after simplification gives

$$\begin{aligned} Var_{RSWOR}(\widehat{\beta}_2) &= \frac{(\gamma_2+3)\,\sigma^4}{n}\frac{(N-n)}{(N-1)} \\ &\quad + \frac{4\mu^2\sigma^2}{n}\frac{(N-n)}{(N-1)} + \frac{4\mu\gamma_1\sigma^3}{n}\frac{(N-n)}{(N-1)}. \end{aligned} \tag{xxix}$$

The finite population correction is this case is the same as the fpc for the mean case, and we can write the variance of $\hat{\beta}_2$ for the RSWOR case in terms of the variance in the sampling with replacement case as

$$Var_{RSWOR}(\hat{\beta}_2) = Var_{RSWR}(\hat{\beta}_2)\frac{(N-n)}{(N-1)}. \tag{xxx}$$

For the case of clustered data, we need to re-write equation (xxiii) as

$$\begin{aligned} E(\hat{\beta}_2)^2 &= \frac{1}{n^2}E\sum_{c=1}^{C}\sum_{i=1}^{M_c}\left(\mu^4 + u_{ci}^4 + 4\mu^3 u_{ci} + 4\mu u_{ci}^3 + 6\mu^2 u_{ci}^2\right) \\ &+\frac{1}{n^2}E\sum_{\substack{c=1\\ i\neq j \text{ for } c=s}}^{C}\sum_{s=1}^{C}\sum_{i=1}^{M_c}\sum_{j=1}^{M_s}\left(\mu^4 + u_{ci}^2u_{sj}^2 + 2\mu^3 u_{ci} + 2\mu^3 u_{sj}\right. \\ &\left.+4\mu^2 u_{ci}u_{sj} + 2\mu u_{ci}^2 u_{sj} + 2\mu u_{ci}u_{sj}^2 + \mu^2 u_{sj}^2 + \mu^2 u_{ci}^2\right). \end{aligned} \tag{xxxi}$$

But we note that we can further divide this second summation into terms within the same cluster (of which there are $n\left(\overline{M}-1\right)$)

$$\begin{aligned} &\frac{1}{n^2}E\sum_{c=1}^{C}\sum_{\substack{i=1\\ i\neq j}}^{M_c}\sum_{j=1}^{M_c}\left(\mu^4 + u_{ci}^2u_{cj}^2 + 2\mu^3 u_{ci} + 2\mu^3 u_{cj}\right. \\ &\left.+4\mu^2 u_{ci}u_{cj} + 2\mu u_{ci}^2 u_{cj} + 2\mu u_{ci}u_{cj}^2 + \mu^2 u_{cj}^2 + \mu^2 u_{ci}^2\right) \end{aligned} \tag{xxxii}$$

and terms in different clusters (of which there are $n^2 - n\overline{M}$)

$$\begin{aligned} &+\frac{1}{n^2}E\sum_{\substack{c=1\\ c\neq s}}^{C}\sum_{s=1}^{C}\sum_{i=1}^{M_c}\sum_{j=1}^{M_s}\left(\mu^4 + u_{ci}^2u_{sj}^2 + 2\mu^3 u_{ci} + 2\mu^3 u_{sj}\right. \\ &\left.+4\mu^2 u_{ci}u_{sj} + 2\mu u_{ci}^2 u_{sj} + 2\mu u_{ci}u_{sj}^2 + \mu^2 u_{sj}^2 + \mu^2 u_{ci}^2\right). \end{aligned} \tag{xxxiii}$$

We can then evaluate these expressions using the assumptions of (167) and (168).

Thus for the case of clustered data we will have

$$\begin{aligned} Var_{clust}(\hat{\beta}_2) &= \frac{\mu^4}{n} + \frac{(\gamma_2+3)\,\sigma^4}{n} + \frac{4\mu\gamma_1\sigma^3}{n} + \frac{6\mu^2\sigma^2}{n} \\ &+\frac{\mu^4}{n}(\overline{M}-1) + \frac{2\mu^2\sigma^2}{n}(\overline{M}-1) + \frac{(\overline{M}-1)}{n}\sigma_{1122,w} \\ &+\frac{4\mu(\overline{M}-1)}{n}\sigma_{112,w} + \frac{4\mu^2(\overline{M}-1)}{n}\sigma_{12,w} \\ &+\frac{\mu^4}{n}(n-\overline{M}) + \frac{2\mu^2\sigma^2}{n}(n-\overline{M}) + \frac{\sigma^4}{n}(n-\overline{M}) \\ &-(\mu^2+\sigma^2)^2. \end{aligned} \tag{xxxiv}$$

One additional algebraic manipulation gives

$$\begin{aligned} Var_{clust}(\hat{\beta}_2) &= \frac{1}{n}\left[\left(\gamma_2+3-\overline{M}\right)\sigma^4 + 4\mu\gamma_1\sigma^3 + 4\mu^2\sigma^2\right] \\ &+\frac{4(\overline{M}-1)}{n}\left[\frac{1}{4}\sigma_{1122,w} + \mu\sigma_{112,w} + \mu^2\sigma_{12,w}\right]. \end{aligned} \tag{xxxv}$$

We also promised to provide the $Cov(\hat{\beta}_2, \overline{y})$. We first write

$$\begin{aligned} Cov(\hat{\beta}_2, \overline{y}) &= E\left(\hat{\beta}_2 - E\hat{\beta}_2\right)(\overline{y} - E\overline{y}) \\ &= E\left(\frac{1}{n}\sum x_i^2 - \mu^2 - \sigma^2\right)\left(\frac{1}{n}\sum x_i - \mu\right). \end{aligned} \tag{xxxvi}$$

Expanding this term

$$\begin{aligned} Cov(\hat{\beta}_2, \overline{y}) &= \frac{1}{n^2}E\sum_{i=1}^{n}\left(\mu^3 + u_i^3 + 3\mu^2 u_i + 3\mu u_i^2\right) \\ &+\frac{1}{n^2}E\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\left[\mu^3 + \mu u_i^2 + u_i^2 u_j + 2\mu^2 u_i + \mu^2 u_j + 2\mu u_i u_j\right] \\ &-\mu(\mu^2+\sigma^2) \end{aligned} \tag{xxxvii}$$

and evaluating the expectations

$$Cov(\widehat{\beta}_2, \overline{y}) = \frac{1}{n}\left\{\mu^3 + \gamma_1\sigma^3 + 3\mu\sigma^2\right\} + \frac{n-1}{n}\left\{\mu^3 + \mu\sigma^2\right\} + \frac{n-1}{n}\left\{2\mu\sigma_{12} + \sigma_{112}\right\} - \mu^3 - \mu\sigma^2. \quad \text{(xxxviii)}$$

When the sampling is with replacement $\sigma_{12} = \sigma_{112} = 0$, so this reduces to the term in equation (103)

$$Cov_{RSWR}(\widehat{\beta}_2, \overline{y}) = \frac{1}{n}\left[\sigma^3\gamma_1 + 2\mu\sigma^2\right] \quad \text{(ixl)}$$

and when the sampling is without replacement it is simple to show that

$$Cov_{RSWOR}(\widehat{\beta}_2, \overline{y}) = \frac{\sigma^3\gamma_1 + 2\mu\sigma^2}{n}\frac{(N-n)}{(N-1)} = Cov_{RSWR}(\widehat{\beta}_2, \overline{y})\frac{(N-n)}{(N-1)}. \quad \text{(xl)}$$

For the case of clustered sampling (with replacement), we first note that the following equalities will hold under the assumptions (167) and (168)

$$\begin{aligned} E\sum_{\substack{k=1\\k\neq l}}^{n}\sum_{l=1}^{n} u_k^2 u_l &= n(n-1)\sigma_{112} \\ &= E\sum_{\substack{c=1\\i\neq j}}^{C}\sum_{\substack{s=1\\for\ c=s}}^{C}\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ci}^2 u_{sj} \quad \text{(xli)} \\ &= E\sum_{\substack{c=1\\i\neq j}}^{C}\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ci}^2 u_{cj} = n(\overline{M}-1)\sigma_{112,w} \end{aligned}$$

and

$$E\sum_{\substack{k=1\\k\neq l}}^{n}\sum_{l=1}^{n} u_k u_l = n(n-1)\sigma_{12}$$

$$= E\sum_{\substack{c=1\\i\neq j}}^{C}\sum_{\substack{s=1\\for\ c=s}}^{C}\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ci}u_{sj} \qquad \text{(xlii)}$$

$$= E\sum_{c=1}^{C}\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n} u_{ci}u_{cj} = n(\overline{M}-1)\sigma_{12,w}.$$

Substituting these into (xxxvii) gives

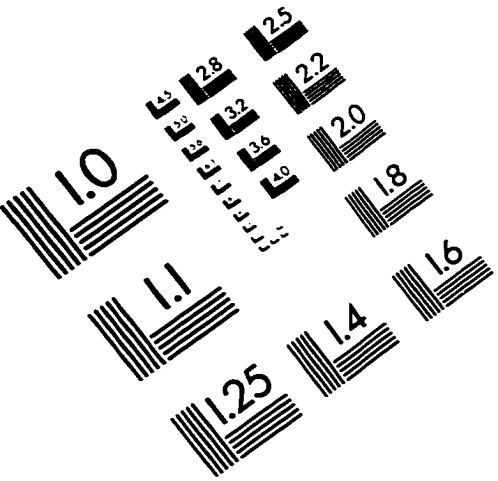
$$Cov_{clust}(\hat{\beta}_2,\bar{y}) = \frac{1}{n}\left\{\mu^3+\gamma_1\sigma^3+3\mu\sigma^2\right\}+\frac{n-1}{n}\left\{\mu^3+\mu\sigma^2\right\}$$
$$+\frac{\overline{M}-1}{n}\left\{2\mu\sigma_{12,w}+\sigma_{112,w}\right\}-\mu^3-\mu\sigma^2. \qquad \text{(xliii)}$$

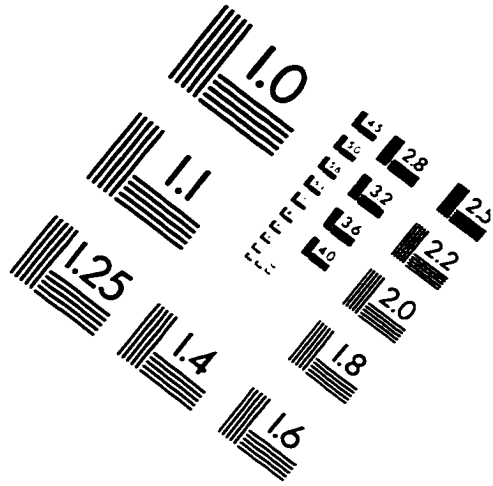
We can simplify this expression to give

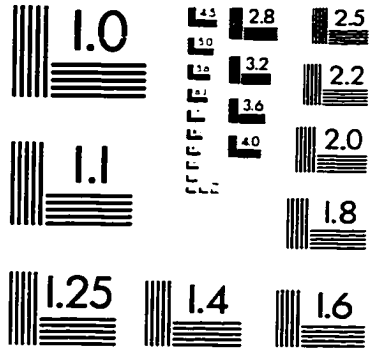
$$Cov_{clust}(\hat{\beta}_2,\bar{y}) = \frac{\gamma_1\sigma^3}{n}+\frac{2\mu\sigma^2}{n}+\frac{\overline{M}-1}{n}\left[2\mu\sigma_{12,w}+\sigma_{112,w}\right]. \qquad \text{(xliv)}$$

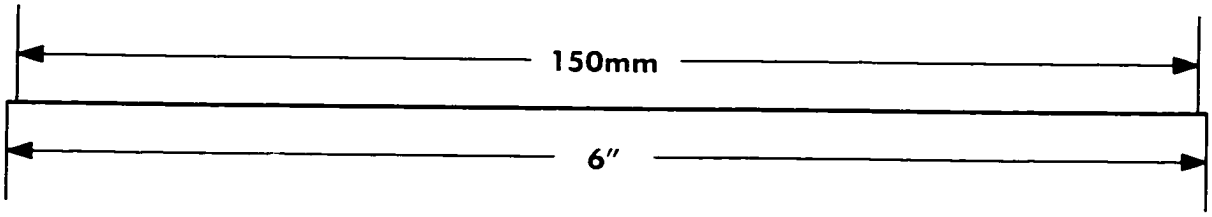
Combining these results gives the variance of the coefficient of variation under clustered sampling.

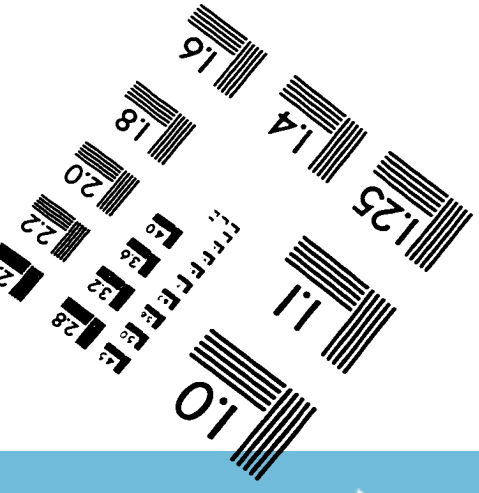
IMAGE EVALUATION TEST TARGET (QA-3)

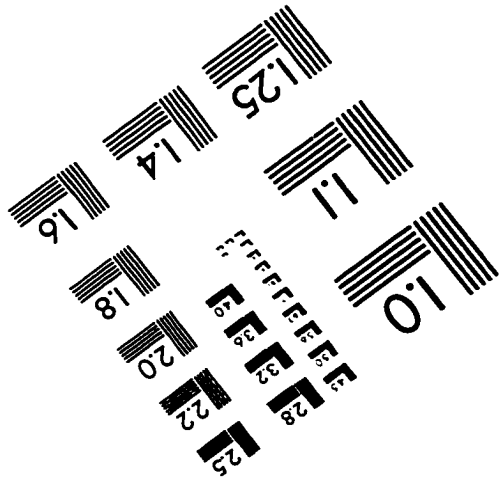












APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989