Chapter 1

Introduction

1.1 Review of the Literature

In survey sampling, estimation of first order parameters such as population mean, total and distribution function is usually of primary concern. In the presence of auxiliary information there exists several general estimation procedures in recent literature to obtain more efficient estimators for these parameters. The auxiliary information may be in the form of known population totals, an example being the surface area of study region. In other cases, the auxiliary information is in the form of estimators, for example estimators from the first phase of a two-phase sample.

The relationship between the auxiliary and survey variable can be used in basically two ways. The traditional design strategy is to use randomized sampling and techniques such as stratification, systematic selection, probability proportional to size selection, and ratio or regression estimation. Measures of error are based on the distribution of estimators from all possible samples. The second approach is to state the relationship between the auxiliary and survey variables in terms of a model and to use selection and estimation schemes that minimizes a measure of error for the
distribution of the estimator over all possible replications of the population given the model. In some cases this approach results in only one optimal sample, whereas in others choice of possible samples is restricted to a set. If the model does not hold, the optimal design may be far optimal. Two examples of the second approach, which attempt to guard against departures from the model, were given by Royall and Herson (1973a; 1973b) and Brewer (1979).

The use of auxiliary information in estimating the finite distribution function has attracted increased attention in recent literature. Several estimators which incorporate knowledge of an auxiliary variable known for every unit in the population have been proposed and their performances examined and compared see, for examples, Chambers and Dunstan (1986), Rao, Kovar and Mantel (1990), Chambers, Dorfman and Hail (1992), Kuk (1993) and Wang and Dorfman (1996). Less attention has been given to the variance estimation problem. Wu and Sitter (2001a) develop jackknife and analytical variance estimators.

While a purely model-based prediction approach has been used by some researchers, the model-assisted approach has gained much popularity in recent literature. Several general procedures have been proposed, including generalized regression (GREG) estimators (Cassel et al., 1976; Särndal, 1980), calibration estimators (Deville and Särndal, 1992a), empirical likelihood methods (Chen and Qin, 1993; Chen and Sitter, 1999; Zhong and Rao, 2000), and, more recently the model calibration and model-calibrated pseudoempirical likelihood methods (Wu and Sitter, 2001a).

Estimation of quadratic or other higher-order finite population functions is also important. For example, efficient estimators for finite population variances, covariances between two response variables, or variances of linear estimators are highly
desirable. Shah and Patel (1996) presented several examples to illustrate why the estimation of population variances and covariances might be useful in their own right. However, because of the relative complexity of these functions, it is not obvious that one can obtain more efficient estimators for these higher-order population quantities when certain auxiliary information is available from survey data.

When survey results are presented, it is good practice to provide measure of precision/accuracy of an estimator used in the survey.

(A) To assess the accuracy or performance of an estimator, several important summary statistics may be considered: (i) the relative bias in percentage (RB), (ii) the relative standard error in percentage (RSE) and (iii) the confidence intervals. Unfortunately there may be a conflict between criteria (ii) and (iii): it is frequently the case that the estimator of variance with minimum mean square error (MSE) provides poorer interval estimates than some other variance estimators with larger MSE.

The most commonly used measure of precision for unbiased (biased) estimators is the variance (MSE) of survey estimator. Usually the exact value of the variance is unknown, because it depends on unknown population quantities. Royall and Cumberland's (1978a) in their introductory discussion of variance estimation gave two possible uses of a variance estimator viz., for planning a future survey and for inferential purposes. For inferential purposes, an important statistic is the 't' statistic which is used to construct confidence intervals.

(B) Although accuracy issues should dominate decision about variance estimators, administrative considerations such as cost and timing must also play an important role, particularly in the complex surveys. The cost of computing a highly
accurate estimate of variance for each survey statistics may be very formidable indeed, far exceeding the total survey budget. In such circumstances, methods of variance estimation that are cost effective may be highly desirable even though they may involve a certain loss of accuracy. Timing is another important practical consideration because modern complex survey often have rather strict closeout dates and publication deadlines. The methods of variance estimation must be evaluated in light of such deadlines and the efficiency of the computer environment to be used in preparing the survey estimators.

(C) A final issue, though perhaps subordinate to the accuracy, cost and timing consideration, is the simplicity of the variance estimating methodology. There are three aspects of simplicity. First, it will be necessary to use one or at most few methodology that involve a tolerance loss of accuracy for all at least the most important survey statistic. The second aspect of the simplicity involves the computer processing system used for the survey. The third aspect of simplicity is concerned with the survey sponsor and users of the survey data.

A major problem faced by statistician working on a large complex survey is the selection of a variance estimation procedure. Most of the basic theory developed in the standard sampling texts deals with variance estimation for linear estimators, and, therefore, is not applicable to complex survey involving adjustments for nonresponse and/or ratio and composite estimation procedures.

There are two main approaches to the variance estimation problem: (a) Design-based approach, (b) Model-based approach.

(a) Design-based Approach: In design inference (see, (Särndal et al., 1992)), it is the randomized selection of the sample, according to the given sampling design. In
this approach, the estimators are required to be asymptotic design unbiased (ADU) and the quality of the estimator is measured by the design variance. For confidence intervals, we need an approximately design-unbiased variance estimator.

A variance estimator should be:

a. Unbiased, or nearly so.

b. Stable, that is, the variance of the variance estimator should be small.

c. Nonnegative, that is, always take nonnegative values.

d. Produce confidence intervals that cover true value with a probability approximately equal to the stated confidence level.

So far the three general variance estimation methods have been dealt with.

1. Two unbiased estimators of the variance of the Horvitz-Thompson (HT) estimator or \( \pi \)-estimator were given, viz. the Horvitz-Thompson estimator and the Yates-Grundy-Sen estimator. Both estimators can also be modified and used when an estimator is a linear function of several \( \pi \) estimators.

2. When an estimator is a nonlinear function of \( \pi \) estimators, an exact expression for the variance cannot usually be found. In such a case the Taylor linearization was used to obtain an approximate variance.

3. The third method is a weighted residual variance estimator for the regression estimator. This variance estimator has advantages over the Taylor linearization estimator (Särndal et al., 1989).
For unequal probability sampling design, design-based variance estimation is cumbersome because (i) when both the \( \pi \)ps requirements i.e. inclusion probability proportional to a measure of size \( \pi_i \propto x_i \) for all \( i \) and fixed sample size \( n \) requirements are imposed on the scheme, it becomes tedious and often computationally difficult to calculate \( \pi_{ij} \), the joint inclusion probability of unit \( i \) and \( j \), especially if other desirable features such as \( \pi_{ij} - \pi_i \pi_j > 0 \) for all \( i \neq j \) are also required. (ii) The design-based variance estimation uses the \( \pi_{ij} \) in a cumbersome double-sum calculation with \( \frac{n(n-1)}{2} \) terms (using the Yates-Grundy formula, the sample size being fixed). Thus very large number of terms effectively rules out correct variance calculation in many \( \pi \)ps surveys. (It is sometimes assumed that the simpler \( \pi \)ps with replacement formula will be satisfactory, but then the correct \( \pi \)ps scenario has already been abandoned).

Another approach has been to obtain fairly single approximation to the unequal \( \pi_{ij} \). This still leaves the tedious double-sum calculations for the variance. Stehman and Overton (1994) proposed and reviewed possibilities of this kind. A weakness of the approach is that the variance computed by the double sum is then an approximation whose accuracy is not always known.

Due to the lack of or complexity of direct estimators of variance for complex surveys, simplified methods of variance estimation have been developed for use with complex surveys over the last few decades.

The majority of the simplified methods of variance estimation are based upon some form of repeated subsampling. The methods of random groups, jackknife, and balanced repeated replications differ primarily in the procedures for forming the subsamples. The selection of a method of variance estimation from among these choices for a particular survey design and estimator is not straightforward. Accuracy, time,
and cost all need to be considered.

These techniques are able to handle both complex estimators. Thus, they can be used in cases where methods (1) to (3) above are not easily applicable. Since they usually require extensive computation. They are sometimes called computer intensive. These variance estimators are theoretically complicated, because it is hard to derive general results on their statistical properties, for example, their bias. These techniques are primarily used with more complex estimators, however. In these cases, no theoretical results are available, and our knowledge about the statistical properties of the variance estimators is limited to conclusions drawn from simulation studies or other sources of empirical evidence.

(b) Model-based approach: By contrast, in model inference (see, (Valliant et al., 2000)), the argument is conditional on given sample and makes reference to an assumed model. If an estimator is model unbiased, its quality is measured by its model variance. Having found an estimated model variance one can construct model-based confidence intervals that are interpreted via repeated realizations of the term.

By imposing a super-population model on the actual finite population, several model-based variance estimators are proposed and studied in Royall and Eberhardt (1975) and Royall and Cumberland (1978a).

There is a considerable literature on variance estimation for survey estimators; a thorough account is given by Wolter (1985) and also by Särndal et al. (1992). Holt and Smith (1979), Royall and Cumberland (1978a; 1981a; 1981b) and other advocates of the model-based viewpoint have made important remarks on variance estimation. For widely used estimators, such as the ratio estimator and simple regression estimator, there have been many simulation studies done to compare different
variance estimators. Two such studies were those of Wu and Deng (1983) and Deng and Wu (1987). The findings of these and other simulation studies show that it is hard to single out a best variance estimator and the issue of variance estimation has not been finely resolved.

Fuller (1970) proposed a regression estimator of the variance of the Horvitz-Thompson (HT) estimator of the population total using as \( x \) variables the quantities \((\pi_i \pi_j - \pi_{ij})\) and \((\pi_i \pi_j - \pi_{ij})(i - j)^2\). Ogus and Clark (1971) proposed the use of ratio and difference estimator of the variance under a Poisson sampling design (see, (Hajek, 1964)) for the purpose of reducing the effect of random sample size on the variance estimator. Isaki (1983) has proposed and compared various regression type variance estimators of HT-estimator under the PPS one unit per stratum sampling design.

When mean \( \bar{X} \) of a single auxiliary variable \( x \) is known, a common strategy used by several authors is to restrict to a specific class of estimators, say, \( V_h = h(\bar{X})V \), where \( \bar{X} \) is the sample mean of the \( x \) variable, \( h(\cdot) \) is a smooth function satisfying \( h(1) = 1 \), and \( V \) is a conventional variance estimator not using any auxiliary information. The objective is then to find an optimal estimator within this class (see, (Das and Tripathi, 1978); (Deng and Wu, 1987); and references cited there in). When both \( \bar{X} \) and \( S_x^2 \) are known, a general class, say \( V_h = h(\bar{X}, S_x^2)V \), may be considered, where \( h(\cdot, \cdot) \) is a smooth function satisfying \( h(1, 1) = 1 \), \( S_x^2 \) and \( s_x^2 \) are the population and the sample variance of variable \( x \) respectively.

Stehman and Overton (1994) compared variance estimators of the HT estimator for randomized variable probability systematic sampling.

Chaudhari and Maiti (1994) developed two alternative modifications of Yates-Grundy type variance estimator of GREG-estimator so that the limiting values of the
design expectations of the model expectations of variance estimators match respectively the (a) model expectations of the Taylor approximation of the design variance of the GREG-estimator and the (b) limiting value of the design expectation between the GREG-estimator of the population total.

For unequal probability sampling, design-based variance estimation is cumbersome because it requires second-order inclusion probabilities. For most fixed size \( \pi ps \) scheme, these probabilities are difficult to compute, and variance estimation depends on them for a tedious double-sum calculation. Särndal (1996) introduced a simple weighted squared residual (WSR) form of the variance of GREG-estimator. This form leads directly to an extremely simple variance estimator free of the difficulty with the \( \pi_{ij} \) for specific sampling designs. He showed that stratified simple random sampling with a suitable Bernoulli sampling when used together with a GREG-estimator produce a variance of the WSR form. He also, showed that Poisson sampling combined with the GREG-estimator forms an efficient \( \pi ps \) strategy having WSR form.


Recently, Calibration has become a widely used procedure for estimation in sample survey. It uses auxiliary information to produce efficient estimators. Singh et al. (1999), following Deville and Särndal (1992b), have proposed a high level calibration approach for HT and GREG variance estimation. In the context of \( \pi ps \) sampling their estimators in contrast to the HT variance estimator require an additional auxiliary variable over and above the one used to define the inclusion probabilities. Kumar, Gupta and Agarwal (1985) proposed a variance estimator which is independent of \( \pi_{ij} \) and is model unbiased.
Royall (1976; 1986) and Valliant (1987) considered the model-based approach for the variance estimation for two-stage cluster sampling.

The jackknife technique is a standard tool in the sampler's ornament of variance estimator (see, e.g., (Krewski and Rao, 1981); (Rao and Wu, 1981); (Stukel et al., 1996); (Rust, 1985)). Duchesne (2000) derives the explicit jackknife variance estimators of the GREG-estimator using the random group technique and a corrected version is proposed that removes a large part of the positive model bias. He pointed out that jackknife estimators are perhaps applicable to exceptional situations like shortage of time, one-time use, etc. Their chief merit is that they require less programming efforts, e.g., there is no need to evaluate all the $\pi_{ij}$ as in the design-based variance estimation.

By viewing quadratic and other second-order finite population functions as total or means over a derived synthetic finite population, Sitter and Wu (2002) have shown that the recently proposed model-calibration (Wu and Sitter, 2001a) and pseudoempirical likelihood methods (Chen and Sitter, 1999) for effective use of auxiliary information from survey data can be readily extended to obtain efficient estimators of quadratic and other second-order finite population functions. In particular, estimation of a finite population variance, covariance, or variance of a linear estimator can be greatly improved when auxiliary information is available.

1.2 Notation and Terminology

Let $U = \{1, \cdots, i, \cdots, N\}$ be the set of labels of the finite population. Associated with unit $i$ are values of response variable ($y_i$) and covariate ($x_i$). We assume that the
values \( x_1, \ldots, x_N \) are known for the entire finite population (referred to as complete auxiliary information) but \( y_i \) is known only if the \( i^{th} \) unit is selected in the sample \( s \) drawn using sampling design \( p(s) \). A sampling design without replacement is a probability distribution \( p(\cdot) \) on all the non-empty subsets \( s \subseteq U \) such that

\[
\sum_{s \subseteq U} p(s) = 1 \quad \text{and} \quad p(s) \geq 0.
\]

A sampling design of fixed size \( n \) is such that \( p(s) = 0 \), if \( \text{card}(s) \neq n \). The first and second order inclusion probabilities of population units can be computed as follows:

\[
\pi_i = \sum_{s \ni} p(s)
\]

\[
\pi_{ij} = \sum_{s \ni,j} p(s)
\]

The use of unequal inclusion probability sampling can often be interesting in survey sampling. The expectation and the variance of an estimator \( \hat{\theta} \) of the parameter \( \theta \) are defined respectively by

\[
E(\hat{\theta}) = \sum_{s \ni \mathcal{S}} p(s) \hat{\theta}(s)
\]

and

\[
V(\hat{\theta}) = \sum_{s \ni \mathcal{S}} p(s)(\hat{\theta}(s) - \theta)^2
\]

where \( \mathcal{S} = \{s : s \subseteq U\} \). Two important measures of the quality of an estimator \( \hat{\theta} \) are the bias and the mean squared error (MSE). The bias of \( \hat{\theta} \) is defined as

\[
B(\hat{\theta}) = E(\hat{\theta}) - \theta.
\]

An estimator \( \hat{\theta} \) is said to be unbiased for \( \theta \) if \( B(\hat{\theta}) = 0 \) for all \( (y_1, \ldots, y_N) \in R_N \). The MSE of \( \hat{\theta} \) is defined as \( \text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + [B(\hat{\theta})]^2 \).

If \( \hat{\theta} \) is unbiased for \( \theta \), \( \text{MSE}(\hat{\theta}) = V(\hat{\theta}) \).
Because the bias, variance, and other statistical quantities are compared by averaging over all the samples that might be drawn under a particular sampling design, these are sometimes referred to as repeated sampling or design-based properties.

If $\pi_i > 0, i \in U$, the total of the variable $y$

$$Y = \sum_{i \in U} y_i$$

can be estimated unbiasedly by the $\pi$-estimator (HT-estimator) given by

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

where $s$ is the random sample drawn by means of the sampling design $p(\cdot)$ in such a way that $P(S = s) = p(s)$.

The Horvitz-Thompson (1952) variance of $\hat{Y}_{HT}$ is

$$V_{HT}(\hat{Y}_{HT}) = \sum_{i \in U} \left(\frac{1}{\pi_i} - 1\right) y_i^2 + \sum_{i \neq j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) y_i y_j$$

and the unbiased estimator of the variance is

$$V_{HT}(\hat{Y}_{HT}) = \sum_{i \in s} \left(\frac{1}{\pi_i} - 1\right) \frac{y_i^2}{\pi_i} + \sum_{i \neq j \in s} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) \frac{y_i y_j}{\pi_{ij}}. \quad (1.2.1)$$

If the sample size is fixed, then the variance can be written

$$V_{YG}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{i \neq j \in U} \sum_{\pi_{ij} - \pi_i \pi_j} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

This variance expression is due to Yates and Grundy (1953) and can be estimated by

$$V_{YG}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{i \neq j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \quad (1.2.2)$$

In order that these variance estimators be unbiased, a necessary and sufficient condition is that the joint inclusion probabilities be strictly positive : $\pi_{ij} > 0$ for all
Moreover a sufficient condition in order that $V_{yg} > 0$ is that $\pi_i \pi_j > \pi_{ij}$, for all $i, j \in U, i \neq j$. The interest of the use of the unequal probabilities is that $y_i \propto \pi_i$, then $V_{HT}$ equals zero. Thus, if an auxiliary variable $x$ close to $y$ and such that $x_i > 0, i \in U$, is known, it is often more interesting to select the units with unequal probabilities proportional to the $x_i$.

The abundance of works about sampling designs without replacement with fixed sample size is quite disconcerting. No less than fifty sampling procedures are presented in the famous paper of Hanif and Brewer (1980). Several other recent papers deal with this matter as well. Among them, one can cite the Sunter method (Sunter, 1977; Sunter, 1986), the Chao updating procedure (1982) and the Pomix-sampling (Kröger et al., 1999).

The search for a sampling design with unequal probabilities is a relatively open problem. A "good" solution should however respect the following properties (see, (Tille, 1996)):

1. The procedure should be exact in the sense that the units should be selected exactly with probabilities equalling $\pi_i, i \in U$.

2. The procedure should be general i.e. it should be possible to apply it to any set of first-order inclusion probabilities fixed a priori, which satisfy the relation

$$\sum_{i \in U} \pi_i = n \quad \text{where} \quad 0 < \pi_i < 1, \ i \in U.$$

3. The algorithm should be fast, the selection of the sample should be made without computing $p(s)$ for the $N! / \{n!(N - n)!\}$ possible samples of size $n$.

4. The algorithm should be sequential i.e. it should be possible to apply it to a
data file in only one reading by examining the units in accordance with their order number on the data file.

5. The \( p(s) \) obtained by means of the sampling method should not depend on the order of the units on the data file.

6. The joint inclusion probabilities should be easy to compute without examining all the probabilities \( p(s) \).

7. The joint inclusion probabilities should be strictly positive.

8. The joint inclusion probabilities should verify the Yates-Grundy condition:
   \[
   \pi_{ij} < \pi_i \pi_j \text{ for all } i \neq j.
   \]

9. The method should give an estimator with a smaller variance than sampling with replacement.

A model assisted approach has been proposed by Särndal et al. (1989; 1992). In this approach, design consistent estimator that are also model-unbiased under an assumed "working" model are constructed and used in conjunction with design consistent variance estimators. They are also model unbiased for the conditional (model) variance.

In case of single auxiliary variable \( x \), the model assisted estimator of the population total \( Y \) is given by

\[
\hat{Y}_{\text{REG}} = \hat{Y}_{HT} + \hat{B}(X - \hat{X}_{HT})
\]

where \( \hat{Y}_{HT} \) and \( \hat{X}_{HT} \) are the basic HT estimators of \( Y \) and \( X \), \( \hat{Y}_{HT} = \sum_s d_i y_i \) and \( \hat{X}_{HT} = \sum_s d_i x_i \) and \( \hat{B} = \hat{T}^{-1} \sum_s d_i q_i^{-1} x_i y_i \) with \( \hat{T} = \sum_s d_i q_i^{-1} x_i^2 \). Here \( d_i \) which
denotes the basic design weight attached to \( i \in s \), is the inverse of the inclusion probability \( \pi_i \). The estimator \( \hat{Y}_{GREG} \), named the generalized regression (GREG) estimator, may also be written as \( \sum_s w_i y_i \) with \( w_i = d_i g_i \) and

\[
g_i = 1 + (X - \bar{X})T^{-1}q_i
\]

(Särndal et al., 1992). Also \( \hat{Y}_{GREG} \) is a calibration estimator in the sense \( \sum_s w_i x_i = X \), that is, it ensures consistency with the known total \( X \). Under the ratio model, \( \hat{Y}_{GREG} \) reduces to the customary ratio estimator \( \hat{Y}_R = \frac{\hat{Y}_{HT}}{\bar{X}_{HT}}X \). Note that \( \hat{Y}_{HT} \) uses the \( x \) variable only at sampling stage, whereas \( \hat{Y}_{GREG} \) uses the \( x \) variable both at the sampling as well as estimation stage.

A model-assisted approach is certainly useful, but its limitations should be noted. (Chambers, 1996). First if the working model is indeed true, then it leads to inferences inferior to those obtained under the model dependent approach, because the latter uses just linear unbiased predictors. Secondly, under the model misspecification, it appeals to unconditional design based inference in the case of a purely design based approach so that it is essentially design based. Thus, the desirable conditional features of the model dependent approach are forsaken.

### 1.3 Description of Problem

The use of an important general estimator viz, the Horvitz-Thompson estimator \( \hat{Y}_{HT} \) in probability sampling for estimating a population total \( Y \) is intuitively appealing when \( \pi_i \propto x_i \). This estimator used in many practical surveys, e.g. two large scale environmental surveys, the National Stream Survey (Kaufmann et al. 1988; Sale et al. 1988) and Environmental Monitoring and Assessment Program (Messer, Linthurst,
and Overton 1991) use variable probability, systematic sampling, and the Horvitz-Thompson estimator to estimate population parameters of ecological interest.

The inference procedure is completed by calculating variance estimator and a confidence interval of $Y$. An unbiased variance estimator is given by the HT formula, $\nu_{HT}$ given in (1.2.1), or (for fixed sample size design) by the YG alternative, $\nu_{YG}$ given in (1.2.2). The $\nu_{YG}$ is generally considered superior to the $\nu_{HT}$ because of fewer negative estimates and small sampling variance (Royall and Cumberland, 1981b; Rao and Singh, 1973). However, $\nu_{YG}$ requires fixed sample size, whereas $\nu_{HT}$ does not. This restriction of $\nu_{YG}$ to fixed sample size design eliminated this variance estimator from consideration of many applications in survey (Stehman and Overton, 1994). Random-size designs sometimes have advantages from a survey operational viewpoint, e.g., the use of Bernoulli sampling reduces delays in data capture and provides relatively uniform workload to operations staff (Armstrong et al., 1993).

The estimation of the variance of the Horvitz-Thompson estimator / GREG estimator is indeed a difficult, and in a sense, open problem. As in common practice, survey objectives required special design and analysis features that strongly influenced selection of a variance estimator. This thesis deals with the estimation of variances for Horvitz-Thompson / GREG estimators if auxiliary information is available. This problem is well known and described in the scientific literature. From this point of view the thesis incorporates no purely new issues. On the other hand various estimators, derived by various approaches, are proposed and are optimal under some conditions but by different criteria and investigate properties of these variance estimators via an empirical study. The most significant feature of these approaches is that it effectively uses auxiliary information at the estimation stage under a general sampling design.
1.4 Chapterwise Summary

Chapterwise summary of the thesis is as under:

CHAPTER 2. This chapter deals with a model-based variance estimation of the HT estimator, although attention is concentrated on sampling design unbiased estimators for the finite population quadratic functions. We shall present the results of a small Monte Carlo study that sheds light on the repeated-sampling properties of the estimators.

CHAPTER 3. In this chapter a model-based variance estimation of the Horvitz-Thompson estimator of the finite population total is derived and it is shown that this estimator has smaller expected mean square error among the class of all model-unbiased estimators. A small simulation study is presented to compare the performance of the suggested estimator with the Yates-Grundy variance estimator for a fixed size sampling and with the Horvitz-Thompson variance estimator for a random size sampling.

CHAPTER 4. In this chapter, motivated by the ratio method of estimation, a \( \pi \)-weighted ratio type estimator for HT-variance is suggested and is shown to be asymptotically design unbiased and consistent. An empirical study is conducted to compare its performance. The data used in the simulation come from a variety of populations, many of them real data, and many where the target variable has been modelled on an auxiliary variable. To assess the performances, several important summary statistics such as the percentage relative biases, the percentage relative standard errors, and the empirical coverage rate of the resultant confidence intervals are computed and presented.

CHAPTER 5. The Horvitz-Thompson and Yates-Grundy designed-based variance
estimators of the HT estimator of the population total do not make use of auxiliary population information at the estimation stage. To remedy this, ratio-type and difference-type estimators of the HT variance under a general sampling design are obtained. Also, following Singh et al. (1999), a calibration estimator is derived. These estimators are compared with customary estimator through a simulation study.

CHAPTER 6. In this chapter, various procedures for improving variance estimation with the aid of auxiliary information are discussed. Two estimators of the variance of the regression estimator of a finite population total that have that objective are discussed, and their properties are summarized via an empirical study.

CHAPTER 7. This chapter deals with a design-based variance estimation for the GREG estimator. A \( \pi \)-weighted ratio-type estimator is suggested which is asymptotically design unbiased and consistent. To study the repeated sampling properties of the suggested estimator a Monte Carlo simulation is conducted. The estimators proposed by Särndal et al. (1989; 1992), Kott (1990) and Chaudhuri and Maiti (1994) are included in the comparison.

The following simulation will be used repeatedly in the sequel.

1.5 Monte Carlo Simulation

The performance of the different variance estimators will be examine using a Monte Carlo simulation study. We now provide a description of the simulated population and sample creation.

A finite population of size \( N = 400 \) was created. The characteristics \( x \) and \( y \) for
the $i^{th}$ unit were generated using the model

\[ y_i = \beta x_i + \sigma_x^2 \varepsilon_i \]  \hspace{1cm} (1.5.1)

for specified values of $\beta, \gamma, g, h, \text{and} \sigma_x^2$, where $\varepsilon_i \sim N(0, \sigma_x^2)$ independent of $x_i \sim \text{Gamma}(g, h)$. Thus the mean, variance, and coefficient of variation of $x_i$ are given by $\mu_x = gh; \sigma_x^2 = gh^2; \text{and} \ C_x = \frac{\sigma_x}{\mu_x} = g^{-1/2}$.

Further the mean of $y_i$ is $\mu_y = \beta \mu_x$, variance of $y_i$ is $\sigma_y^2 = \beta^2 \sigma_x^2 + \sigma_x^2 E(x_i^\gamma)$ and the correlation coefficient $\text{Corr}(x_i, y_i) = \rho = \frac{\sigma_x}{\mu_x}$ vary depending on the choice of $\gamma$. Here $\gamma = 0, 1 \text{ and } 2$ were considered so that $E(x_i^\gamma) = 1, \mu_x$ and $\mu_x^2 + \sigma_x^2$; for each of these cases $\sigma_x^2$ and $\sigma_x^2$ were then chosen to match various values of $\rho$ and $C_x$. The three $(\beta, \mu_x, \gamma)$ combinations were: (a) $\beta = 1, \mu_x = 100, \gamma = 0$; (b) $\beta = 1, \mu_x = 100, \gamma = 1$; and (c) $\beta = 1, \mu_x = 100, \gamma = 2$. Figure 1 gives a scatter plot of the resulting finite population values of $x$ and $y$ for each of (a) - (c) with $\rho = 0.8$ and $C_x = 0.75$. 
For each of (a) - (c) and each \( p \) and \( C_x \) combination, the finite population was created and a sample of size \( n = 30 \) was drawn using a specific sampling design. The variance estimators were computed from each sample. This process was repeated \( M = 10,000 \) times. The performance of the different variance estimators was measured and compared in terms of relative bias in percentage (RB), relative standard error in percentage (RSE), relative efficiency (RE) and empirical coverage rate (ECR). The simulated values of RB, RSE, RE and ECR for a particular variance estimator \( \mathcal{V} \) were computed as

\[
RB(\mathcal{V}) = 100 \times \frac{\bar{\mathcal{V}} - V}{V}
\]

where \( \bar{\mathcal{V}} \) is obtained computationally as

\[
\bar{\mathcal{V}} = \frac{1}{M} \sum_{j=1}^{M} \mathcal{V}_j
\]

The relative percentage standard error and relative efficiency of \( \mathcal{V} \) are given by

\[
RSE(\mathcal{V}) = \frac{\sqrt{MSE(\mathcal{V})}}{V} \times 100
\]

and

\[
RE(\mathcal{V}) = \frac{MSE(\mathcal{V}_{YG})}{MSE(\mathcal{V})} \quad \text{or} \quad RE(\mathcal{V}) = \frac{MSE(\mathcal{V}_{HT})}{MSE(\mathcal{V})}
\]

where

\[
MSE(\mathcal{V}) = \frac{1}{M-1} \sum_{j=1}^{M} (\mathcal{V}_j - V)^2
\]

And the empirical coverage rate is

\[
ECR(\mathcal{V}) = \frac{1}{M} \times \sum_{j=1}^{M} I_{\{\mathcal{V} \in \bar{\mathcal{Y}} \pm z_{\alpha/2} \sqrt{\mathcal{V}}\}}
\]
where \( I[\cdot] \) is an indicator function, \( \hat{Y} \) denotes the value of the HT-estimator given a specific sample and \( Z_{a/2} \) denotes a tabular value from the standard normal distribution or \( t \)-distribution.

1.6 Sampling Designs

The fact that a \( \pi ps \) design (inclusion probabilities proportional to a size measure) combined with the Horvitz-Thompson estimator \( \hat{Y}_{HT} \) is an 'optimal' strategy (here a strategy is a combination of a sampling design and an estimator) for the estimation of the population total \( Y \) has stimulate the enormous research work in that direction. A more general unequal probability sampling design would be a computationally nightmare. This naturally raise the question as to which strategy is preferable.

In the simulation studies we shall use the following sampling schemes for sample selection. Their subroutines (in 'C++' Programming Language) are given in the Appendix G.

1.6.1 Simple Random Sampling

This scheme, was suggested by Fan, Muller and Rezucha (Fan et al., 1962). This is a least sequential scheme which gives SRSWOR of size \( n \).

Let \( \varepsilon_1, \varepsilon_2, \ldots \), be independent random numbers drawn from the \( Uniform(0,1) \) distribution. If \( \varepsilon_1 < n/N \), the element \( i = 1 \) is selected, otherwise not. For subsequent elements, \( i = 2, 3, \ldots \), let \( n_i \) be the number of elements selected among the first \( i - 1 \) elements in the population list. If

\[
\varepsilon_i < \frac{n - n_i}{N - i - 1}
\]
the element $i$ is selected, otherwise not. The procedure terminates when $n_i = n$.

1.6.2 Poisson Sampling (Random-size strictly $\pi$ps)

Poisson sampling (PO) is an unequal probability sampling. The PO design leads to random sample size and is the generalization of Bernoulli sampling. This design is appealing because it is simple to execute and it is strictly $\pi$ps. Each unit $i$ is examined and independently given some probability $\pi_i$ of being selected. The joint inclusion probability of two units is simply $\pi_{ij} = \pi_i \pi_j$. Let $x_1, x_2, \ldots, x_N$ be the auxiliary variable. The procedure of Poisson sampling is as follows:

(i) Compute inclusion probabilities $\pi_i = \frac{n x_i}{\sum_i x_i}, i = 1, 2, \ldots, N$.

(ii) Generate $N$ independent $Uniform(0,1)$ random numbers, say $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$.

(iii) If $\varepsilon_i < \pi_i$, the element $i$ is selected, otherwise not; $i = 1, 2, \ldots, N$.

It is true that HT estimator $\hat{Y}_{HT}$ incurs a variance penalty for random sample size. But one should not discard random sample size design as inferior; they can be used to advantage together with the GREG estimator.

1.6.3 Midzuno's Sampling Scheme (Fixed-size Non-$\pi$ps)

The simple procedure of selecting a sample was suggested by Midzuno (1952), which consists in selecting the first unit with pps and the remaining $(n - 1)$ units from $(N - 1)$ units of the population by SRSWOR.

For this selection procedure, the inclusion probabilities for individual and pairwise
units are given by

$$\pi_i = p_i + \left( \frac{n - 1}{N - 1} \right)(1 - p_i)$$

$$= \frac{n - 1}{N - 1} + \frac{N - n}{N - 1} p_i$$

and

$$\pi_{ij} = \frac{n - 1}{N - 1} \left[ \frac{N - 1}{N - 2} (\pi_i + \pi_j) - \frac{n}{N - 2} \right]$$

1.6.4 Sunter's Sampling Scheme (Fixed-size approximate \(\pi ps\))

In \(\pi ps\) sampling scheme, the inclusion probabilities satisfy \(\pi_i \propto x_i\), where \(x_1, \cdots, x_N\) are known, positive numbers.

For most important part of the population, Sunter's scheme gives the inclusion probability \(\pi_i\) strictly proportional to \(x_i\). The scheme is as follows:

(i) Given data \(x_1, \cdots, x_N\), arrange the data in descending order of magnitude, say, \(u_1, \cdots, u_N\).

(ii) Generate Uniform(0,1) random number say \(\varepsilon_1\). Compute \(\pi_1 = \frac{n u_1}{T_N}, T_N = \sum u_i x_i\).

(iii) If \(\varepsilon_1 < \pi_1\), then select \(u_1\) in the sample, otherwise not.

(iv) For each \(i = 2, 3, \cdots\), generate an independent Uniform(0,1) random number \(\varepsilon_i\).

Compute

$$\pi'_i = \frac{(n - n_i)x_i}{t_i}$$
where \( t_i = u_i + u_{i+1} + \cdots + u_N = \sum_{k=i}^{N} u_k \)

\( n_i \): number of elements selected among first \((i - 1)\) elements in the population list obtained in (ii).

(v) If \( \varepsilon_i < \pi_i' \) then the element \( i \) is selected, otherwise not.

(vi) The process described in (iii) to (iv) ends when \( n_i = n \) or \( i = i^* \) whichever occur first, where \( i^* = \min \{i_0, N - n + 1\} \) with \( i_0 \) equal to smallest \( i \) for which \( \frac{n x_i}{t_i} \geq 1 \).

(vii) If \( n_{i^*} < n \), the process (iii) to (vi) has not produced the full size \( n \).

For \( i = i^*, i^* + 1, \cdots \) generate an independent Uniform \((0,1)\) random number \( \varepsilon_i \) and compute

\[
\pi_i^o = \frac{n - n_i}{N - i + 1}
\]

if \( \varepsilon_i < \pi_i^o \) then element \( k \) is selected in the sample, otherwise not. The process ends when \( n_i = n \).
1.7 List of Formulae

\[ \hat{Y}_{HT} = \sum_{i} \frac{y_i}{\pi_i} \]

\[ V_{HT}(\hat{Y}_{HT}) = \sum_{U} \left( \frac{1}{\pi_i} - 1 \right) y_i^2 + \sum_{U} \sum_{U} \left( \frac{\pi_{ij}}{\pi_i\pi_j} - 1 \right) y_i y_j \]

\[ V_{YG}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{U} \sum_{U} \left( \frac{\pi_{ij} - \pi_{i}\pi_j}{\pi_i} \right) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \]

\[ V_{HT}(\hat{Y}_{HT}) = \sum_{s} \left( \frac{1}{\pi_i} - 1 \right) y_i^2 + \sum_{s} \sum_{s} \left( \frac{\pi_{ij}}{\pi_i\pi_j} - 1 \right) y_i y_j \]

\[ V_{YG}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{s} \sum_{s} \left( \frac{\pi_{ij} - \pi_{i}\pi_j}{\pi_i} \right) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \]

\[ V_{OPT}(\hat{Y}_{HT}) = \left( \sum_{U} \left( \frac{1}{\pi_i} - 1 \right) x_i^2 \right) \sum_{s} \frac{y_i^2}{n \pi_i^2} \]

\[ + \left( \sum_{U} \sum_{U} \left( \frac{\pi_{ij}}{\pi_i\pi_j} - 1 \right) x_i x_j \right) \sum_{s} \sum_{s} \frac{y_i y_j}{n(n-1)x_i x_j} \]

\[ V_{PR}(\hat{Y}_{HT}) = \sum_{s} \left( \frac{1}{\pi_i} - 1 \right) y_i^2 + \sum_{s} \sum_{s} \left( \frac{\pi_{ij}}{\pi_i\pi_j} - 1 \right) y_i y_j \]

\[ + \left\{ \sum_{U} \left( \frac{1}{\pi_k} - 1 \right) x_k^2 - \sum_{s} \left( \frac{1}{\pi_k} - 1 \right) x_k^2 \right\} \sum_{s} \frac{y_i^2}{n \pi_i^2} \]

\[ + \left\{ \sum_{U} \sum_{U} \left( \frac{\pi_{kl}}{\pi_k\pi_l} - 1 \right) x_k x_l \right\} \sum_{s} \sum_{s} \frac{y_i y_j}{n(n-1)x_i x_j} \]

\[ V_{WR}(\hat{Y}_{HT}) = \begin{cases} 
\left( \sum_{s} \left( \frac{1}{\pi_i} - 1 \right) x_i^2 \right) \sum_{s} \left( \frac{1}{\pi_i} - 1 \right) x_i^2 & \text{if } \pi_{ij} = \pi_i \pi_j \forall i \neq j \in U \\
\left( \sum_{s} \left( \frac{1}{\pi_i} - 1 \right) \frac{x_i}{\pi_i} \right) \sum_{s} \left( \frac{1}{\pi_i} - 1 \right) \frac{x_i}{\pi_i} + \left( \sum_{s} \sum_{s} \left( \frac{\pi_{ij}}{\pi_i\pi_j} - 1 \right) y_i y_j \right) \sum_{s} \sum_{s} \frac{y_i y_j}{\pi_i \pi_j} & \text{otherwise.} 
\end{cases} \]
\[
V_{RHT}(\hat{Y}_{HT}) = \frac{V_{HT}(\hat{X}_{HT})}{V_{HT}(\hat{X}_{HT})} V_{HT}(\hat{X}_{HT})
\]

\[
V_{RYG}(\hat{Y}_{HT}) = \frac{V_{YG}(\hat{Y}_{HT})}{V_{YG}(\hat{X}_{HT})} V_{YG}(\hat{X}_{HT})
\]

\[
V_{CYG}(\hat{Y}_{HT}) = V_{YG}(\hat{Y}_{HT}) + \gamma_{CYG}[V_{YG}(\hat{X}_{HT}) - V_{YG}(\hat{X}_{HT})]
\]

\[
\gamma_{CYG} = \sum_{s} \left( \frac{\pi_{ij} - \pi_{i} \pi_{j}}{\pi_{ij}} \right) \left( \frac{y_{i} - y_{j}}{\pi_{i} - \pi_{j}} \right)^{2} \left( \frac{x_{i} - x_{j}}{\pi_{i} - \pi_{j}} \right)^{4}
\]

\[
V_{CHT}(\hat{Y}_{HT}) = V_{HT}(\hat{Y}_{HT}) + \gamma_{CHT}[V_{HT}(\hat{X}_{HT}) - V_{HT}(\hat{X}_{HT})]
\]

\[
\gamma_{CHT} = \sum_{s} \left( \frac{1}{\pi_{i} - 1} \right) \frac{x_{i} y_{i}^{2}}{\pi_{i}} + \sum_{s} \left( \frac{\pi_{ij} - \pi_{i} \pi_{j}}{\pi_{ij}} \right) \frac{x_{i} x_{j} y_{i} y_{j}}{\pi_{i} \pi_{j}}
\]

\[
\frac{RB(V)}{V} = 100 \times \frac{\hat{V} - V}{V}
\]

\[
\frac{RSE(V)}{V} = 100 \times \frac{\sqrt{MSE}}{V}
\]

\[
\frac{RE(V)}{V} = \frac{MSE(V_YG)}{MSE(V_Y)}
\]

\[
ECR(V) = \frac{1}{M} \times \sum_{j=1}^{M} I_{[Y \in \hat{Y} \pm Z_{\alpha/2} \sqrt{V}]}
\]

**Estimator Acronyms**

- HT - Horvitz-Thompson
- OPT - Optimum
- πWR - π-Weighted Ratio-Type
- CHT - Calibration using HT-Variance estimator
- CYG - Calibration using YG-Variance estimator
- RHT - Ratio-type using HT-Variance estimator
- RYG - Ratio-type using YG-Variance estimator

[YG - Yates-Grundy](#)

[PR - Predictor](#)

[GREG - Generalized Regression](#)

★ ★ ★