University of Alberta

A Robust Estimator of Distribution Function and Quantile in the Presence of Auxiliary Information

By

Hong Li ©

A thesis
submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirement for the degree of a Master of Science
in
Statistics.

Department of Mathematical Sciences

Edmonton, Alberta, Canada

Spring 2000.

0-612-60140-4

University of Alberta

Library Release Form

Name of Author: Hong Li

Title of Thesis: A Robust Estimator of Distribution Function and Quantile in the presence of Auxiliary Information

Degree: Master of Science

Year this Degree Granted: 2000

Hong Li

c/o Department of Mathematical Sciences
University of Alberta
Edmonton, Alberta
Canada T6G 2E9

Date January 27, 2000

UNIVERSITY OF ALBERTA

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **A Robust Estimator of Distribution Function and Quantile in the Presence of Auxiliary Information** submitted by **Hong Li** in partial fulfillment of the requirements for the degree of Master of Science in Statistics.

Dr. R.J. Karunamuni (Chair)

Dr. N.G.N. Prasad (Supervisor)

Dr. Francis Yeh

Date: January 26, 2000

Abstract

Recently, estimation of distribution function and quantiles has received considerable attention in survey sampling. This thesis mainly considers estimation of distribution function and quantiles under single-stage and two-stage sampling designs. In the first part, for single-stage sampling process, estimation of distribution function and quantiles is considered through the use of auxiliary information. In the second part, along the lines of first part, the estimation problem under two-stage sampling plan is considered. The idea of estimating population total under model-assisted approach is also extended to estimate population distribution function and the resulting estimator is shown to perform better than the conventional estimator for moderate sample sizes. Through a Monte Carlo simulation study, proposed estimators are compared with conventional estimators.

Acknowledgement

Upon the completion of my thesis, I would like to take this opportunity to acknowledge all those people who have helped me.

First I would like to thank my supervisor, Dr. N. G. Narasimha Prasad for his guidance, patience and helps throughout the preparation of this thesis. Special thanks go to Dr. Rohana J. Karunamuni and Dr. Francis Yeh for serving as the committee members.

My gratitude extends to the Department of Mathematical Sciences, University of Alberta, for providing me the opportunity to study here, to Dr. Peter M. Hooper for academic and financial support in the first year of my study, and to the faculty members of the department who provided me with many interesting courses in Statistics.

Contents

List of Tables

List of Figures

Chapter 1
Introduction

In modern society, the need for statistical information has increased considerately for making policy decisions. In particular, data are regularly collected to satisfy the need for information about specified set of elements. The collection of all such elements is called a **finite population**. But collecting the data for each element in a population will be too expensive and/or time consuming; sometimes it is even impractical. So, one of the most important method of collecting information for policy decision is sample survey, that is, a partial investigation of the finite population.

A sample survey costs less than a complete enumeration and may even be more accurate than the complete enumeration due to the fact that personnel of higher quality can be employed for the data collection. So, designing an efficient sampling scheme at possible lowest cost and to develop some efficient methods to estimate population parameters from the sample data under given design have become major issues in sample surveys.

In this thesis, I will only concentrate on the estimation of the distribution function and associate quantiles of a finite population.

1.1 Background

Consider a finite survey population U which consists of N distinct elements. Each element is identified through label j $(j = 1, \dots, N)$. Let values of the population elements for variable Y be $Y_1, Y_2, ..., Y_N$, then for any given number t $(-\infty < t < \infty)$, the population distribution function $F_N(t)$ is defined as the proportion of elements in the population for which $Y_j \leq t$, for $j \in U$, that is

$$F_N(t) := N^{-1} \sum_{j \in U} \Delta(t - Y_j), \tag{1.1}$$

where

$$\Delta(t - Y_j) = \begin{cases} 1, & \text{when } t \geq Y_j; \\ 0, & \text{otherwise.} \end{cases} \tag{1.2}$$

Also, population α^{th} quantile $(0 < \alpha < 1)$ for Y is defined as

$$q_N(\alpha) := \inf\{t; F_N(t) \geq \alpha\}. \tag{1.3}$$

Without loss of generality, we omit N in notation of $F_N(t)$ or $q_N(\alpha)$.

Consider an auxiliary variable X, associated with Y, with the known values X_1, X_2, ..., X_N.

Now, if a single-stage sampling scheme is undertaken to select a sample $s = \{j_1, j_2$, ... $, j_n\}$ of size n under some sampling design $\{S,\ p(\cdot)\}$, where S denotes the set of all possible samples of size n, $p(s)$ is the probability of selecting the sample s, $s \in S$, such that $\sum_{s \in S} p(s) = 1$. The values for Y in set s are observed as $y_{j_1}, y_{j_2}, ..., y_{j_n}$. Then, only with this information, a general class of design-based estimator of $F(t)$ is given by Rao (1994) as:

$$\hat{F}(t) = \begin{cases} \dfrac{\sum_{i \in s} d_i(s) \Delta(t - y_i)}{N}, & \text{when } N \text{ is known;} \\ \dfrac{\sum_{i \in s} d_i(s) \Delta(t - y_i)}{\sum_{i \in s} d_i(s)}, & \text{when } N \text{ is unknown.} \end{cases} \tag{1.4}$$

And naturally, the α^{th} quantile $q(\alpha)$ may be estimated as

$$\hat{q}(\alpha) = \inf\{t;\ \hat{F}(t) \geq \alpha\}. \tag{1.5}$$

So, in the following I will focus on the estimation of distribution function. The associate quantile estimator will be defined as (1.5) accordingly. In (1.4), $d_i(s)$ is the basic design weight for single-stage sampling which can depend on both s and i ($i \in s$) (see, Godambe, 1955), which also satisfies the design-unbiasedness condition: $\sum_{\{s:\ i\in s\}} p(s) d_i(s) = 1$ for $i = 1, \ldots, N$. To get an unbiased estimator of F, it is also essential to assume that the inclusion probabilities $\pi_i = \sum_{\{s:\ i\in s\}} p(s)$, for $i = 1$, ... , N, are all positive (Rao, 1975). Under simple random sampling scheme, $\hat{F}(t)$ reduces to the naive empirical distribution function estimator

$$\hat{F}_{edf}\ (t) = n^{-1} \sum_{i\in s} \Delta(t - y_i). \tag{1.6}$$

Plugging $\hat{F}_{edf}$ for $\hat{F}$ (1.5) gives the naive quantile estimator $\hat{q}_{edf}(\alpha)$.

The above approach to estimate distribution function is called **probability sampling approach** which is free from model assumption. However, the above estimators do not incorporate any supplementary information directly into the estimation process, although the use of auxiliary information on population is known to increase the precision of an estimator. In theory, information on one auxiliary variable may be used at the survey design stage by defining appropriate inclusion probabilities π_i. But this is not possible when information on several auxiliary variables is available. So we may wish to introduce the auxiliary information at the survey estimation stage also (Chambers and Dunstan, 1986). In recent years, several estimators of distribution function of a finite population have been proposed with the use of auxiliary population information (Rao, 1994).

1.2 Objective

For single-stage sampling scheme, the most notable distribution function estimators are the model assisted estimator (Rao, 1994) and the pseudo-empirical maximum likelihood estimator (Chen and Sitter, 1996). The former uses the complete information of auxiliary variable but with the assumption that the relationship between Y and X follow a linear model. Eventhough, the latter does not assume any model, the resulting estimator is optimal under a linear model. Also, the latter method is computational intensive. In this thesis, I will develop an estimator of the finite population distribution function and associate quantiles such that it uses the complete information of X and is exempt of model assumption. More precisely, this estimator uses a nonparametric approach.

For a two-stage sampling scheme, Royall (1976) proposed the linear least-squares prediction approach in the estimation of population total, with the cluster size as supplementary information. In this thesis, The same approach will be used to get an estimator of the finite population distribution function and associate quantiles.

All of the estimators proposed in this thesis will be compared with conventional estimators through a Monto Carlo simulation study.

1.3 Thesis Overview

In chapter 2, some of the literature on the estimation of distribution function and quantiles for a finite population under single-stage and two-stage sampling scheme are reviewed. In chapter 3, by exploring the advantages of different methods under single-stage sampling plan, a robust distribution function estimator is proposed. This estimator can be used under single-stage sampling plan without model assumption. The proposed method is examined through a limited simulation study. In chapter 4, this thesis presents some discussion of the estimation of distribution function and associate quantiles under two-stage sampling scheme. Also, the basic idea of estimating population total is extended to estimate population distribution function. A Monto Carlo simulation study is also presented to assess the efficiency of proposed estimator.

Chapter 2
Review of Literature

Single-stage sampling is the most simple sampling scheme in sample surveys. Several methods are considered in the literature for estimating distribution function and quantiles, which can be grouped under three broad headings: (i) design-based approach; (ii) model-dependent approach, or prediction approach; and (iii) model-assisted approach.

However, in practice, multi-stage sampling is also frequently used, especially in surveys of human populations. For example, Kish (1965) described a stratified three-stage sample of dwellings, *i.e.*, sampling counties at the first stage, sampling blocks within selected counties at the second stage and sampling dwellings within chosen blocks at the third stage. For this kind of surveys, there is a wide variety of possible designs in any stage, and there is a much wider variety of possible estimators of interested parameters.

So, this chapter mainly reviews some of papers that are relevant to the estimation of finite population distribution function and quantiles under single-stage sampling design.

2.1 General Estimation Method of Distribution Function

Consider a single-stage sampling scheme defined by $\{S,\ p(\cdot)\}$ and a sample set $s \in S$ with design weight $d_i(s)$ for $i \in s$. Also, we assume availability of auxiliary information, say X, for all of the units in the population. Let $y_{j_1}, ..., y_{j_n}$ be the observed values for Y in set s. We consider below different approaches to estimate the distribution function (1.1) and α^{th} quantile (1.3).

2.1.1 Design-based Approach

Under the design-based approach, $F(t)$ can be estimated by the ratio estimator $\hat{F}_r(t)$, the difference estimator $\hat{F}_d(t)$ and the regression estimator $\hat{F}_{reg}(t)$ (See Rao and Liu (1992) and Rao (1994)).

In the case of a single auxiliary x-variable, denote $g(x_j) = \Delta(t - \hat{R}x_j)$ for $j = 1, ..., N$; where, $\hat{R} = \sum_{i \in s} d_i(s) y_i / \sum_{i \in s} d_i(s) x_i$. Denote $\hat{H} = \sum_{i \in s} d_i(s) \Delta(t - y_i)$, $\hat{G} = \sum_{i \in s} d_i(s) g(x_i)$, $G = \sum_{j \in U} g(x_j)$, and $\hat{B} = cov(\hat{H}, \hat{G}) / v(\hat{G})$. Then, for the sample s, we have

$$\hat{F}_r(t) = \frac{1}{N} \frac{\hat{H}}{\hat{G}} G, \tag{2.1}$$

$$\hat{F}_d(t) = \frac{1}{N} \left\{ \hat{H} + (G - \hat{G}) \right\}, \tag{2.2}$$

and

$$\hat{F}_{reg}(t) = \frac{1}{N} \left\{ \hat{H} + \hat{B}(G - \hat{G}) \right\}. \tag{2.3}$$

The above three estimators have some good properties. For example, $\hat{F}_r(t)$ reduces to $F(t)$ and variance becomes zero when $y_j \propto x_j$ for all $j \in U$. This indicates that the ratio estimator can gain large efficiency when $\Delta(t - y_i)$ and $\Delta(t - Rx_i)$ have a stronger linear relationship. However, in real world, the correlation between

$\Delta(t - y_i)$ and $\Delta(t - \hat{R}x_i)$ is generally weaker than that between y_i and x_i, so here estimator $\hat{F}_r(t)$ actually gains very little efficiency over $\hat{F}(t)$. As for estimators $\hat{F}_d(t)$ and $\hat{F}_{reg}(t)$, they suffer from the same drawback as the ratio estimator $\hat{F}_r(t)$. Since $\hat{F}_{reg}(t)$ involves the computation of $cov(\hat{H}, \hat{G})$ and $v(\hat{G})$, it is computationally more cumbersome than the ratio estimator $\hat{F}_r(t)$. Moreover, all of these three estimators are not model-unbiased under a linear model assumption, although they are asymptotically design-unbiased (Rao, Kovar and Mantel, 1990 and Rao, 1994).

2.1.2 Prediction Approach

This approach, contrary to the design-based approach, assumes that the population Y values are random and the relationship between Y and the auxiliary variable X follows a certain model. The most-common used superpopulation model is given by,

$$y_j = \beta x_j + v(x_j)\varepsilon_j, \quad j = 1, \ ..., \ N, \tag{2.4}$$

where β is an unknown parameter, $v(x_j)$ is a strictly positive function of x_j, and ε_j's are independent and identically distributed (**i.i.d**) random variables with zero means. Under this approach valid inference will be made with respect to the model and irrespective of the sampling design $p(s)$.

At first, note that $F_N(t)$ can be decomposed as follows,

$$F_N(t) = N^{-1}\{\sum_{i\in s} \Delta(t - y_i) + \sum_{j\in \bar{s}} \Delta(t - y_j)\}, \tag{2.5}$$

where s is the same as before, $\bar{s} = \{j;\ j = 1,\ 2,\ ...,\ N$ and $j \notin s\}$. The only unknown in (2.5) is the second term. It is also assumed that the y_j's are independent and the population model holds for the sample, so there will be no sample selection bias (Krieger and Pfeffermann, 1992). Under this assumption, a general estimator (referenced as **CD** estimator) of distribution function, which was suggested by Chambers and Dunstan (1986), is given by

$$F_N^*(t) = N^{-1}\{\sum_{i\in s}\Delta(t-y_i) + \frac{1}{n}\sum_{i\in s,\, j\in \bar{s}}\Delta(\frac{t-b_nx_j}{v(x_j)} - u_{ni})\}, \tag{2.6}$$

where, $b_n = \{\sum_{i\in s} y_ix_i/v^2(x_i)\}\{\sum_{i\in s} x_i^2/v^2(x_i)\}^{-1}$, which is the best linear unbiased estimator (**BLUE**) of β, $u_{ni} = \{v(x_i)\}^{-1}(t-b_nx_i)$. Rao and Liu (1992) noted that, when $v^2(x_j) = \sigma^2x_j$, (2.6) reduces to

$$F_N^*(t) = N^{-1}\{\sum_{i\in s}\Delta(t-y_i) + n^{-1}\sum_{i\in s,\, j\in \bar{s}}\Delta(x_j^{-\frac{1}{2}}(t-\tilde{\beta}_sx_j) - u_{ni})\}, \tag{2.7}$$

where, $\tilde{\beta}_s = n^{-1}\sum_{i\in s} y_i/(n^{-1}\sum_{i\in s} x_i) = \bar{y}_s/\bar{x}_s$, $u_{ni} = x_i^{-\frac{1}{2}}(y_i - \tilde{\beta}_sx_i)$, and the u_{ni}'s are approximately independent with mean zero and variance σ^2.

In general, the model-based estimator of the distribution function will not be the same as those suggested by the conventional design-based approach as in (1.4). But there is one special situation "when the sample is a stratified random sample and the auxiliary information X is a qualitative variable, indicating the stratum in which a population element occurs" under which $F_N^*(t)$ and $\hat{F}_N(t)$ are consistent (Chambers and Dunstan, 1986).

2.1.3 Model-assisted Approach

From earlier discussion, we see that although probability sampling approach is assumption free, the associated inferences have to refer to repeated sampling instead of just the particular sample, s, that has been chosen. Prediction approach, on the other hand, ignores the sampling design completely and the definition of $F_N^*(t)$ depends on (2.4) being the "correct" model for the population. In particular, it assumes that the heteroscedasticity function $v(x)$ is correctly specified. However, in practice, this is unlikely to be the case and in large samples, prediction inferences are also very sensitive to model misspecifications (Hansen, Madow and Tepping, 1983).

To overcome the shortcomings in the above two methods, the model-assisted approach is proposed as an approach providing valid inferences under an assumed model and at the same time protecting against model misspecifications in the sense of providing valid design-based inferences irrespective of the population Y-values. That is, we can only consider design-consistent estimators that are also model-unbiased (at least asymptotically) under an assumed model. So in the case of estimating a finite population distribution function, we may assume that $Y_1, Y_2, ..., Y_N$ come from the superpopulation with model (2.4). Cases of this problem have been treated by Chambers and Dunstan(1986), Kuk(1988), Godambe(1989), Rao, Kovar and Mantel(1990), Chambers (1992), Rao and Liu(1992) and Rao(1994).

Let G denote the model cumulative density function (**c.d.f.**) of ε_j, then under the superpopulation model (2.4),

$$\sum_{j=1}^{N} \phi_j = \sum_{j=1}^{N} [\Delta(t - y_j) - G(\frac{y_j - x_j\beta}{v(x_j)})] \tag{2.8}$$

is a population estimating function, each of those terms has expectation zero. Using estimation function theory, Godambe (1989) arrived at a model-assisted estimator of

$F(t)$ under the special case of $d_i(s) = \pi_i^{-1}$. Rao (1994) extended this to a general design weight $d_i(s)$, where the case $v^2(x_j) = x_j\sigma^2$ was considered.

Under the model (2.4), a predictor of $\Delta(t - y_j)$ is given by

$$\hat{g}(x_j) = \left(\sum_{i\in s} d_i(s)\right)^{-1} \left\{\sum_{i\in s} d_i(s)\Delta\left[x_j^{-\frac{1}{2}}(t - \hat{R}x_j) - e_i\right]\right\}, \tag{2.9}$$

where

$$e_i = x_i^{-\frac{1}{2}}(y_i - \hat{R}x_i). \tag{2.10}$$

Then a model assisted estimator of $F(t)$ based on (2.9) can be given by the difference estimator:

$$\hat{F}_{ma}(t) = N^{-1}\left\{\sum_{i\in s} d_i(s)\Delta(t - y_i) + \left[\sum_{j\in U} \hat{g}(x_j) - \sum_{i\in s} d_i(s)\hat{g}(x_i)\right]\right\}, \tag{2.11}$$

which is model-unbiased (at least asymptotically) under the assumed model, but its asymptotic design-bias is zero only for a subclass of sampling designs. However, this subclass seems to cover a wide variety of sampling design(Godambe, 1989).

For the special case of $d_i(s) = \pi_i^{-1}$, Rao, Kovar and Mantel (1990) also proposed an alternative model-assisted estimator (it can be reference as **RKM** estimator), after noting that the ratio and the difference estimators, $\hat{F}_r(t)$ and $\hat{F}_d(t)$, are not model-unbiased for $F_N(t)$. This estimator is asymptotically model-unbiased and design-unbiased under all designs. Rao and Liu (1992) extended this estimator to the general case of $d_i(s)$ as follows:

$$\hat{F}_{dm}(t) = N^{-1}\{\sum_{i\in s} d_i(s)\Delta(t-y_i) + (\sum_{j\in U}\tilde{g}(x_j) - \sum_{i\in s} d_i(s)\hat{g}_{ic}(x_i))\}, \tag{2.12}$$

where,

$$\hat{g}_{ic}(x_i) = \{\sum_{j\in s} d_j(s/i)\}^{-1}\left\{\sum_{j\in s} d_j(s/i)\Delta\left[x_i^{-\frac{1}{2}}((t-\hat{R}x_i)-e_j)\right]\right\}. \tag{2.13}$$

So, $\hat{F}_{ma}(t)$ and $\hat{F}_{dm}(t)$ are different only in the last term of the formula.

2.2 Robustness of Model-assisted Distribution Function Estimator

From above section, we know that the model-assisted estimator $\hat{F}_{ma}(t)$ and $\hat{F}_{dm}(t)$ have some good properties (Rao, Kovar and Mantel, 1990), but whether or not they are better than $\hat{F}(t)$ as in (1.4) depends on the degree of validity of the

model, this is reflected in the prediction of $\Delta(t-y_j)$. So, some improvements might arise with the use of a more general model and the local fitting method which could be employed to increase the robustness of the estimator.

Chambers, Dorfman and Wehrly (1993) also explored a robust estimation approach via nonparametric regression method against model misspecification. For a general model

$$\phi(y) = \eta(x) + e, \tag{2.14}$$

where $\eta(\cdot)$ is some reasonably smooth function, $\phi(\cdot)$ is some known function and e is the random error with zero mean. This approach works by first getting a robust (although may be inefficient) predictor of population total Φ of ϕ with nonparametric regression, then using bias calibration method to get a more efficient predictor. That is, one smooths $\phi(y)$ against x to obtain $\hat{\eta}_1(x)$ and then smooths $\phi(y) - \hat{\eta}_1(x)$ against x to obtain $\hat{\eta}_2(x)$. So, the final smoothing of $\phi(y)$ against x is defined as $\hat{\eta}_1(x) + \hat{\eta}_2(x)$. In the case of estimating the finite population distribution function of Y, we can take $\phi(Y) = N^{-1}\Delta(t - Y)$, then

$$\Phi = F(t) = N^{-1}\sum_{j=1}^{N}\Delta(t - Y_j) = \frac{n}{N}F_s(t) + (1 - \frac{n}{N})F_{ns}(t). \qquad (2.15)$$

Here, $F_s(t)$ is obtained from (1.6) and $\eta(x, t) = Pr\{Y \leq t/x\}$ which is assumed to be a smooth function of x for any t. Chambers, Dorfman and Wehrly (1993) estimated it by

$$\hat{\eta}(x, t) = \sum_{i \in s}\omega_i(x)\Delta(t - y_i), \qquad (2.16)$$

where the weights ω_i can be calculated using kernel smoothing method (Chambers, Dorfman and Wehrly, 1993). Then the predictor of $F_{ns}(t)$ is:

$$\hat{F}_{ns}(t) = \sum_{j \in \bar{s}}\sum_{i \in s}\omega_i(x_j)\Delta(t - y_i) = \sum_{i \in s}u_i\Delta(t - y_i). \qquad (2.17)$$

And a robust calibrated estimator is given by

$$\tilde{F}_{ns}(t) = \hat{F}_{ns}(t) + \sum_{i \in s}u_i\{\Delta(t - y_i) - \hat{G}(\frac{t - \hat{\mu}(X_i)}{\hat{\sigma}(X_i)})\}, \qquad (2.18)$$

where $u_i's$ are the nonparametric prediction weights defining $\hat{F}_{ns}(t)$, $\hat{\mu}$ and $\hat{\sigma}$ are sample-based estimators of the conditional model expectation $E(Y/X)$ and standard deviation $(VAR(Y/X))^{\frac{1}{2}}$ under the working model for the population. $\hat{G}(\cdot)$ is a sample-based estimator of the distribution function $G(\cdot)$ of the standardized error

$\frac{Y - E(Y/X)}{\sqrt{Var(Y/X)}}$. If we put $u_i = \frac{\pi_i^{-1} - 1}{N - n}$ in (2.18) (where π_i is the inclusion probability), then the resulting estimator is basically the same as that suggested by Rao, Kovar and Mantel (1990). They also suggested that, by choosing the weights u_i to reflect the actual distribution of the sample X values, the robustness of the predictor can be improved, and this heavily depends on the choice of the bandwidth h_n. Some rules of choosing bandwidth are also given in that paper. However, this method may loss some efficiency by obtaining the bias robustness.

2.3 More Distribution Function Estimators

In above section, we saw some distribution function estimators with the utility of complete information of the auxiliary variable X. However, in some situations, we may not have any auxiliary variables, but with some auxiliary information about F or its associate parameters available. This kind of situation has been discussed by Qin and Lawless (1994) who used the profile empirical likelihood-based kernel method to estimate the finite population distribution function under the semiparameter model assumption. That is, they estimated F as the maximum empirical likelihood estimator (**MELE**) $\hat{F}_{mele}$ by maximizing the **profile empirical likelihood** function $L(F)$ (Zhang, 1997) which is defined as

$$L(F) := \max_{\mathbf{p}} \bar{L}(\mathbf{p}) = \max_{\mathbf{p}} \prod_{i \in s} p_i = \prod_{i \in s} \hat{p}_i. \tag{2.19}$$

And

$$\hat{F}_{mele} = \sum_{i \in s} \hat{p}_i \Delta(t - y_i). \tag{2.20}$$

Zhang (1997) established the weak convergence of $\hat{F}_{mele}$ so that this approach not only overcomes the disadvantage of the standard kernel method which can not exploit extra information systematically, but also makes the variance of estimator smaller to arrive at some degree of increasing of efficiency. In this thesis, this situation will not be discussed further.

Another case is that we can not get the complete information about the auxiliary variable X, but still with some summary available information on X. Usually this kind of information can be represented as

$$E_X\{\tau_l(X)\} = 0, \quad l = 1, \ ..., \ c, \tag{2.21}$$

where, $\tau(X) = (\tau_1(X), \ ..., \ \tau_c(X))^T$. Then, how this information can be used to improve the estimation?

Chen and Qin (1993) proposed an empirical likelihood approach to be used in simple random sampling when this kind of auxiliary information (2.21) is available. Their results suggested that this approach has desirable properties when population distribution function and quantiles are estimated under this situation. But, the formulation of their method did not extend to more complex survey designs. Chen and Sitter (1996) generalized it to estimate the parameter which is some function of the distribution function $\zeta = \zeta(F_N)$ under the situation when the sample s is drawn using some sampling design $\{D, p(\cdot)\}$ as follows.

For a finite population U, if the entire finite population is available, the empirical likelihood, originally introduced by Owen (1988, 1990) for constructing confidence regions in nonparametric settings, will be

$$L(F) = \prod_{j \in U} p_j, \tag{2.22}$$

with the corresponding log-likelihood function

$$l(\mathbf{p}) = \sum_{j \in U} \log(p_j), \tag{2.23}$$

where $p_j = P(Y = y_j)$. Suppose that p_j's $(0 \leq p_j \leq 1)$ are unknown and $\sum_{j \in U} p_j \leq 1$. Then, in the presence of auxiliary information (2.21), (2.23) could be maximized subject to

$$\sum_{j=1}^{N} p_j = 1 \text{ and } \sum_{j=1}^{N} p_j \tau_l(X_j) = 0, \text{ for } l = 1, \ldots, c. \tag{2.24}$$

However, since we only have a sample set s of size n of the entire population available, we can view (2.23) as a population total, then using the unified theory of sampling methodology to get a design unbiased estimator of $l(\mathbf{p})$, that is

$$\hat{l}(\mathbf{p}) = \sum_{i \in s} d_i(s) \log p_i, \tag{2.25}$$

where $d_i(s)'$s are the design weights which satisfy $E_s(\sum_{i \in s} d_i(s) \log p_i) = \sum_{i=1}^{N} \log(p_i)$ and E_s stands for the expectation respecting to the sampling design. Chen and Sitter termed (2.25) as **pseudo-empirical likelihood.** For auxiliary information of the form (2.21), the problem reduces to maximizing (2.25) subject to (2.24) with N replaced by n. Using the Lagrange multiplier method, they gave the resulting pseudo empirical maximum likelihood estimator (**PEMLE**) of $F_N(t)$ as

$$\hat{F}_{pe}(t) = \sum_{i \in s} \tilde{p}_i \Delta(t - y_i), \tag{2.26}$$

where

$$\tilde{p}_i = \frac{d_i(s)}{\sum_{i\in s} d_i(s)} \frac{1}{1+\boldsymbol{\lambda}^T \boldsymbol{\tau}(X_i)}, \text{ for } i \in s, \tag{2.27}$$

and $\boldsymbol{\lambda}$ satisfies

$$\sum_{i\in s} \frac{d_i(s)}{\sum_{i\in s} d_i(s)} \frac{\boldsymbol{\tau}(X_i)}{1+\boldsymbol{\lambda}^T \boldsymbol{\tau}(X_i)} = 0. \tag{2.28}$$

2.4 General Estimation Method of Quantile

Compared to the estimation of distribution function, more attention has been given to quantile estimation, for example, McCarthy (1965), Loynes (1966), Sedransk and Meyer (1978), Gross (1980), Sedransk and Smith (1983), Chambers and Dunstan (1986), Rao, Kovar and Mantel (1990), Olsson and Rootzen (1996). Many papers discuss the estimation of the median, for the survey variable of interest under simple random sampling or stratified random sampling, *e.g.*, Gross (1980), Sedransk and Meyer (1978), Sedransk and Smith (1983), without making explicit use of the auxiliary variables in the construction of estimators.

To make use of the auxiliary information, Chambers and Dunstan (1986) considered briefly the estimation of a quantile by inverting a model-based estimator $\hat{F}$ of the distribution function through formula (1.5). When the population size N and the sample size n are large, the estimator suffers the possible bias problem resulting from the model misspecification in addition to intensive computation. This approach also assumes that the complete information about the auxiliary variable is known. But in some applications, especially when the auxiliary information is extracted from the statistical reports or other secondary sources, only a certain summary measures or a grouped frequency distribution of the auxiliary variable X are available. In such situations, Kuk and Mak (1989) proposed three estimators of median: **ratio estimator** $\hat{M}_{YR}$, **position estimator** $\hat{M}_{YP}$ and **stratification estimator** $\hat{M}_{YS}$ under the assumption that only the median M_X of X is known. They also showed that the last two estimators are efficient when the relationship between Y and X departs from the linearity assumption.

Rao, Kovar and Mantel (1990) studied the ratio estimator and regression estimator for a general α^{th} quantile, and showed that their estimators could lead to considerable gains in efficiency over the customary estimator $\hat{q}(\alpha)$ when Y_j is approximately proportional to X_j for $j = 1, ..., N$. Olsson and Rootzen (1996) proposed a quantile estimator for a nonparametric components of variance situation and proved its consistency and asymptotic normality.

Chapter 3
Distribution Function and Quantile Estimator for One-Stage Sampling: Non-parametric Approach

The estimators given in previous chapter assume the linear model as described in (2.4). In this chapter, a general model is considered to obtain estimators similar to the ones that are described in the previous chapter.

3.1 Formulation of the General Estimator of Distribution Function

Suppose that nothing is known about the relationship between the study variable Y and the auxiliary variable X except some population information about X. That is, only the following general model can be assumed.

$$\left\{ \begin{array}{lcl} E(Y/X=x) & = & \eta(x) \ , \\ Var(Y/X=x) & = & v^2(x), \end{array} \right. \tag{3.1}$$

where $\eta(\cdot)$ is unknown, $v(\cdot)$ is some positive function of x. In this case, nonparametric method is frequently used, and several methods have been proposed for estimating $\eta(\cdot)$, for example, kernel, spline, and orthogonal series methods. Fan (1992) developed a design-adaptive nonparametric regression method, which was based on a weighted local linear regression. This method works as follows.

Suppose that the second derivative of $\eta(x_0)$ exists. Using Taylor's expansion in a small neighborhood of a point x_0 gives

$$\eta(x) \cong \eta(x_0) + \eta'(x_0)(x - x_0) = a + b(x - x_0). \tag{3.2}$$

Then estimating $\eta(x_0)$ is equivalent to estimating the intercept a of a local linear regression problem. Now if the study variable Y and auxiliary variable X satisfy (3.1), we can consider such a weighted local linear regression problem: to find a and b by minimizing

$$\sum_{i=1}^{n} \{Y_i - a - b(x_i - x_0)\}^2 \mathbf{K}(\frac{x_0 - x_i}{h_n}),$$

where $\mathbf{K}(\cdot)$ is some kernel function and h_n is the smooth bandwidth. The solution of b will not be discussed in this thesis. Then, the weighted least square solution of a is defined to be the local linear regression smoother, *i.e.*,

$$\hat{\eta}(x_0) = \hat{a} = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i \,, \tag{3.3}$$

with

$$w_i \equiv \mathbf{K}(\frac{x_0 - x_i}{h_n})\{s_{n,2} - (x_0 - x_i)s_{n,1}\}, \tag{3.4}$$

where

$$s_{n,l} = \sum_{i=1}^{n} \mathbf{K}(\frac{x_0 - x_i}{h_n})(x_0 - x_i)^l, \qquad l = 1, 2. \tag{3.5}$$

The bandwidth h_n can be chosen either subjectively by data analysts or objectively by data. Fan (1992) showed that the local linear regression smoother have high asymptotic efficiency (it can be 100% with a suitable choice of kernel and bandwidth) among possible linear smoothers. It adapts to almost all regression settings and does not require any modifications even at the boundary. I explore the use of this method to obtain an estimator for distribution function and quantile.

The model-assisted estimator $\hat{F}_{ma}$ in (2.11) has some advantages, and its efficiency depends largely on the accuracy of the predictor of $\Delta(t-y_j)$. So we can improve the distribution function estimator via improving the predictor of $\Delta(t-y_j)$. From the predictor expression (2.9), we see that it uses the assumption of linearity between Y and X. But in most cases, we can only use the general model (3.1) to describe the relationship between Y and X, this gives a new predictor of $\Delta(t-y_j)$ as

$$\tilde{g}(x_j) = \left(\sum_{i\in s} d_i(s)\right)^{-1} \left\{\sum_{i\in s} d_i(s)\Delta\left[x_j^{-\frac{1}{2}}(t-\hat{\eta}(x_j)) - e_i\right]\right\}, \tag{3.6}$$

where

$$e_i = x_i^{-\frac{1}{2}}\{y_i - \hat{\eta}(x_i)\}, \quad \text{for } i \in s. \tag{3.7}$$

By using an efficient predictor $\hat{\eta}(x_j)$ of Y_j (see Fan, 1992), $\tilde{g}(x_j)$ should be more efficient than $\hat{g}(x_j)$ in (2.9). If we plug $\tilde{g}(x_j)$ into the formula (2.11), we will get a new estimator, denoted as $\hat{F}_{pma}(t)$, of the distribution function as follows:

$$\hat{F}_{pma}(t) = N^{-1}\left\{\sum_{i\in s} d_i(s)\Delta(t-y_i) + \left[\sum_{j\in U}\tilde{g}(x_j) - \sum_{i\in s} d_i(s)\tilde{g}(x_i)\right]\right\}, \tag{3.8}$$

and expect that it has a better performance. In next section, we will present some simulation results to compare the behavior of $\hat{F}_{ma}$, $\hat{F}_{pma}$ and $\hat{F}_{pe}$.

About the estimation of α^{th} quantile of Y, we will estimate it by inverting the distribution function estimator $\hat{F}(t)$ to get $\hat{q}(\alpha)$. Since $\hat{F}(t)$ is not necessarily nondecreasing function of t (Olsson and Rootzén, 1996), to estimate the quantile $q(\alpha)$, we have to modify $\hat{F}(t)$ to $\tilde{F}(t)$ such that $\tilde{F}(t)$ is a nondecreasing function of t, *i.e.*, we have

$$\tilde{F}(t) = \sup\{\hat{F}(y) : y \leq t\}. \tag{3.9}$$

It can be showed that $\tilde{F}(t)$ and $\hat{F}(t)$ have the same limiting distribution function and the above modification would not affect the value of α^{th} quantile $\hat{q}(\alpha)$ (Olsson and Rootzen, 1996).

In the simulation, we will compare some quantile estimators which are obtained from the direct or indirect use of auxiliary information. They are denoted as $\hat{q}_{ma}(\alpha)$, $\hat{q}_{pma}(\alpha)$, $\hat{q}_{pe}(\alpha)$, $\hat{q}_{r}(\alpha)$ and $\hat{q}_{d}(\alpha)$. The former three are obtained from (1.5) when $\hat{F}(t)$ is replaced by $\hat{F}_{ma}(t)$, $\hat{F}_{pma}(t)$ and $\hat{F}_{pe}(t)$ respectively; the later two are the ratio estimator and the difference estimator suggested in Rao, Kovar and Mantel (1990).

3.2 Simulation Study

In this section, a limited simulation is conducted to assess the estimation procedures in earlier section.

Here, two sets of simulated population data are used. Suppose that there are N distinct elements in each population under study. One population data set (Population 1) is generated from a linear model of the form $y_j = \alpha + \beta x_j + v(x_j)\varepsilon_j$; Another population data set (Population 2) comes from a nonparametric model of the form $y_j = \eta(x_j) + v(x_j)\varepsilon_j$ for $j = 1, ..., N$. In both cases, x_j's ($x_j > 0$) and ε_j's are independent, $x_j \backsim N(\mu_X, \sigma_X^2)$, $\varepsilon_j \backsim N(0, 1)$ and $v(x_j) = x_j^{\frac{1}{2}}$ for $j = 1, 2, ..., N$.

No matter which set of population data are being using, the basic procedures are the same. In each simulation, we use the population value $X_1, X_2, ..., X_N$ of the auxiliary variable X and the observed value of Y in a sample set s to estimate the distribution function and associate quantiles of Y through model-assisted approach under linear model assumption ($\hat{F}_{ma}(t)$, $\hat{q}_{ma}(\alpha)$), or under general model assumption ($\hat{F}_{pma}(t)$, $\hat{q}_{pma}(\alpha)$), except when we calculate the pseudo empirical maximum likelihood estimators, $\hat{F}_{el}(t)$ and $\hat{q}_{ma}(\alpha)$, only the population mean $\bar{X}_N$ and the observation value of X in the sample are available as auxiliary information. We will take N_{simu} samples of size n from the simulated population. The procedures are given as follows.

In each simulation, we will use the simple random sampling without replacement ($SRSWOR$) scheme to obtain a sample s of size n out of the population of N elements, because this simple sampling scheme actually plays an important role in practice. So we get the observation data of Y as $\{y_i,\ i \in s\}$, the corresponding auxiliary values are $\{x_i,\ i \in s\}$ and the design weights are $d_i(s) = \frac{N}{n}$ for each sample point $i \in s$.

Distribution function – estimation

Method 1: Model assisted estimator $\hat{F}_{ma}$ proposed in Rao (1994), *i.e.*, the estimator calculated from (2.9) $\sim$ (2.11). This estimator is based on the linear model (2.4) assumption and $v(x_j) = x_j^{\frac{1}{2}}$ with the utility of complete information of auxiliary variable X.

Step 1: Predict y when given x.

Under the above assumptions, β is estimated by the ratio estimate $\hat{\beta} = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i}$, which is a best linear unbiased estimator of β for any sample. Then, when given x_j,

y_j is predicted as $\hat{y}_j = \hat{\beta} x_j$ for each $j \in U$.

Step 2: Calculation of the residual of y_i for each i in s through formula (2.10).

Step 3: Predict $\Delta(t - y_j)$ for each element j in the population U through formula (2.9).

Step 4: Get the value of the distribution function estimator $\hat{F}_{ma}$ at t by using (2.11).

Method 2: Model assisted estimator $\hat{F}_{pma}$ proposed here. This estimator is based on the general model (3.1) and under the assumption that $v(x_j) = x_j^{\frac{1}{2}}$ for $j \in U$ with the utility of complete information of auxiliary variable X.

Step 1: Predict y when given x.

Since the concrete form of the model is unknown, the auxiliary information of X and sample data of Y are used to fit an appropriate model. Here, the local linear regression method (Fan, 1992) is used for this purpose. That is, y_j is predicted through (3.3) when given x_j for each $j \in U$.

Step 2: Calculation of the residuals of y_i for each i in s by using formula (3.7).

Step 3: Predict $\Delta(t - y_j)$ for each element j in the population U by using formula (3.6).

In this process, an appropriate kernel function $K(\cdot)$ and an appropriate bandwidth h_n are needed. Here, the uniform kernel function (see Table 3.3) is firstly used to estimate the distribution function $\hat{F}_{pma}$ and the performance of $\hat{F}_{pma}$ is compared with other distribution function estimators. Later on, the effect of different kernel functions on the estimator $\hat{F}_{pma}$ will also be checked. As for the choice of smooth bandwidth h_n, the cross-validation technique is used. So we choose

$$h_n = \arg\{\min \sum_{i=1}^{n} (y_i - \hat{\eta}_{-i}(x_i))^2\}, \tag{3.10}$$

where $\hat{\eta}_{-i}(\cdot)$ is the regression estimator (3.3) without using the ith observation data (x_i, y_i) of the sample s.

Step 4: Estimate the distribution function $F(t)$ through (3.8).

Note that, in both of the above methods, it is assumed that $v(x_i) = x_i^{\frac{1}{2}}$ and the complete information of X is available.

Method 3: Pseudo empirical maximum likelihood estimator $\hat{F}_{pe}$. This method assumes that only the population mean and the sampled data of X are available as the auxiliary information, but it does not need any model assumption for the relationship between Y and X.

Step 1: Solve the equation (2.28).

Let $\bar{X}_N$ be the population mean of X, then $\tau(x_i) = x_i - \bar{X}_N$. So $c = 1$ in (2.24), and (2.28) becomes

$$\sum_{i \in s} \frac{d_i(s)}{\sum_{i \in s} d_i(s)} \frac{\tau(X_i)}{1 + \lambda\tau(X_i)} = n^{-1} \sum_{i \in s} \frac{\tau(X_i)}{1 + \lambda\tau(X_i)} = 0, \tag{3.11}$$

where $d_i(s) = \frac{N}{n}$. To solve this nonlinear equation for λ, the Newton-Raphson method (Press, Flannery, Teukolsky and Vetterling, 1990) is used here. Plugging one-term Taylor's expansion of (2.27) into the constraint conditions (2.24) gives an initial guess of λ as

$$\lambda_0 = \frac{\sum_{i \in s} d_i(s)\tau(x_i)}{\sum_{i \in s} d_i(s)[\tau(x_i)]^2} = \frac{\sum_{i \in s}(x_i - \bar{X}_N)}{\sum_{i \in s}(x_i - \bar{X}_N)^2}. \tag{3.12}$$

Then starting from this initial guess, an approximate solution of root λ can be obtained quickly.

Step 2: Using the formula (2.27) to calculate the optimal value $\tilde{p}_i$ for $i \in s$.

Step 3: Using the formula (2.26) to calculate the value of distribution function estimator $\hat{F}_{pe}$ at given t.

Also note that, since SRSWOR sampling scheme is being used, $\hat{F}_{pe}$ is actually coincide with $\hat{F}_{mele}$ here.

The α^{th} quantile – estimation

When estimating the quantiles of Y, the assumption is the same as that in the preceding section. If the auxiliary information of X has already been used in the estimation of the distribution function $F(t)$ as above, the distribution function estimator will be directly inverted to estimate the corresponding quantiles of Y. That is, the formula (3.9) and (1.5) will be used to estimate the quantile $q(\alpha)$. The resulting estimators are denoted as $\hat{q}_{ma}(\alpha)$, $\hat{q}_{pma}(\alpha)$ and $\hat{q}_{pe}(\alpha)$ corresponding to $\hat{F}_{ma}$, $\hat{F}_{pma}$ and $\hat{F}_{pe}$ respectively.

However, if the quantile need to be estimated directly from the sampled data of Y with the complete auxiliary information in hand, then it can be estimated by the ratio estimator $\hat{q}_r$ and the difference estimator $\hat{q}_d$ proposed by Rao, Kovar and Mantel (1990) as:

$$\left\{ \begin{array}{lcl} \hat{q}_r & = & \dfrac{\hat{q}_y}{\hat{q}_x} q_x, \\ \hat{q}_d & = & \hat{q}_y + (q_x - \hat{q}_x), \end{array} \right. \tag{3.13}$$

where $\hat{q}_x$, $\hat{q}_y$ are the sample α^{th} quantile of X and Y respectively for a given sample s; q_x is the population α^{th} quantile of X. In this simulation, all these five kinds of quantile estimator of Y will be compared.

Here the relative biases (**RB**) and relative root mean square errors (**RRMSE**) generated in this simulation experiment are used as measurement to compare the performance of different estimators $\hat{\theta}(t)$, which may be either the distribution function estimator $\hat{F}(t)$ or the quantile estimator $\hat{q}(\alpha)$. But for the distribution function estimator, its values only fall in the range between 0 and 1, so the mean square errors (MSE) are also appropriate to measure the efficiency in this case and will be examined too. For each sample $s \in S$, let $\hat{\theta}^r(t)$ denote the estimator $\hat{\theta}(t)$ obtained from the rth simulation sample, and $\theta(t)$ denote the population value of the interested parameter, then

$$RB(\hat{\theta}) := \frac{1}{|\theta(t)|}[\sum_{r=1}^{N_{simu}} \frac{|\hat{\theta}^r(t) - \theta(t)|}{N_{simu}}], \tag{3.14}$$

$$MSE(\hat{\theta}) := \sum_{r=1}^{N_{simu}} \frac{(\hat{\theta}^r(t) - \theta(t))^2}{N_{simu}}, \tag{3.15}$$

and

$$RRMSE(\hat{\theta}) := \frac{1}{|\theta(t)|}[\sum_{r=1}^{N_{simu}} \frac{(\theta^r(t) - \theta(t))^2}{N_{simu}}]^{\frac{1}{2}}. \tag{3.16}$$

<u>Note</u>: Here, we are only interested in the distribution function values at the three quartiles (0.25, 0.50 and 0.75^{th} quantile), so RB and $RRMSE$ are appropriate measures for estimators of distribution function.

A single-stage sampling scheme is usually used in the sampling from a small population, and in most cases, only a very small sample set are observed. So, in this simulation, a small simulated population of size $N = 500$ are generated and small samples of size $n = 30$ are taken out of each population. However, the methods discussed here also apply to population and samples of any other sizes. After $1,000$ times simulation runs, the simulation results become stable, so this section only shows the simulation results of $N_{simu} = 1,000$.

The linear simulated population data come from model $Y_j = -2X_j + X_j^{\frac{1}{2}}\varepsilon_j$, and the nonlinear simulated population data come from model $Y_j = 150+2.5e^{-X_j}+X_j^{\frac{1}{2}}\varepsilon_j$, where, $\varepsilon_j \overset{i.i.d}{\backsim} N(0,1)$ and $X_j \overset{i.i.d}{\backsim} N(10.2, 2.0^2)$ for $j = 1, ..., 500$. The population data of these two data sets are shown in Figure 3.1 and 3.2.

Firstly, uniform kernel function is used to estimate $F(t)$ through (3.8) and the performance of different estimators are compared. Then the effects of different kernels in the process of local regression on the estimator (3.8) of the distribution function are also examined.

Comparison among different estimators of distribution function

The results of the estimation of distribution function $F(t)$ at three different points, $t = 0.25$, 0.50 and 0.75^{th} quantile of Y, are shown in Table 3.1 $\sim$ 3.2. Where F_p stands for the population cumulative distribution function ; $\hat{F}_{ma}$, $\hat{F}_{pma}$ and $\hat{F}_{pe}$ stand for the estimator obtained from method 1, method 2 and method 3, respectively. RB,

MSE, $RRMSE$ stand for the relative biases, mean square errors and relative root of mean square errors of the corresponding distribution function estimator respectively. Figure 3.3a ∽ 3.4c give the plots for the population distribution function and different distribution estimators.

The results indicate that among the three kinds of estimator, if the underlying relationship between the study variable Y and the auxiliary variable X is linear (see Figure 3.1), for the estimation of the distribution function $F(t)$, in terms of biases, as measured by RB, and in terms of efficiency, as measured by MSE and $RRMSE$, $\hat{F}_{ma}$ has the smallest biases and is the most efficient estimator. The pseudo empirical maximum likelihood estimator $\hat{F}_{pe}$ gives the worst performance among these three estimators except that in the median of the population.

In the case of having a nonlinear simulated population data (see Figure 3.2), in terms of biases, as measured by RB, and in terms of efficiency, as measured by MSE and $RRMSE$, $\hat{F}_{pma}$ and $\hat{F}_{pe}$ have comparable performances, but both of them are less biased and more efficient than $\hat{F}_{ma}$.

This indicates that when it is known that there exists a strong linear relationship between Y and X in the underlying population, $\hat{F}_{ma}$ can be used to estimate the distribution function comfortably; but when no indication of strong linear relationship appears, $\hat{F}_{pma}$ and $\hat{F}_{pe}$ may give a better estimation. In most cases, $\hat{F}_{pma}$ is more stable since it makes use of the complete information of the auxiliary variable, but $\hat{F}_{pe}$ has the advantage that it can still be used when only the summary information of the auxiliary variable is available and it is still bias robust.

Figure 3.1: Scatterplot of Linear Simulated Population from Model

$$Y = -2X + X^{\frac{1}{2}}\varepsilon$$

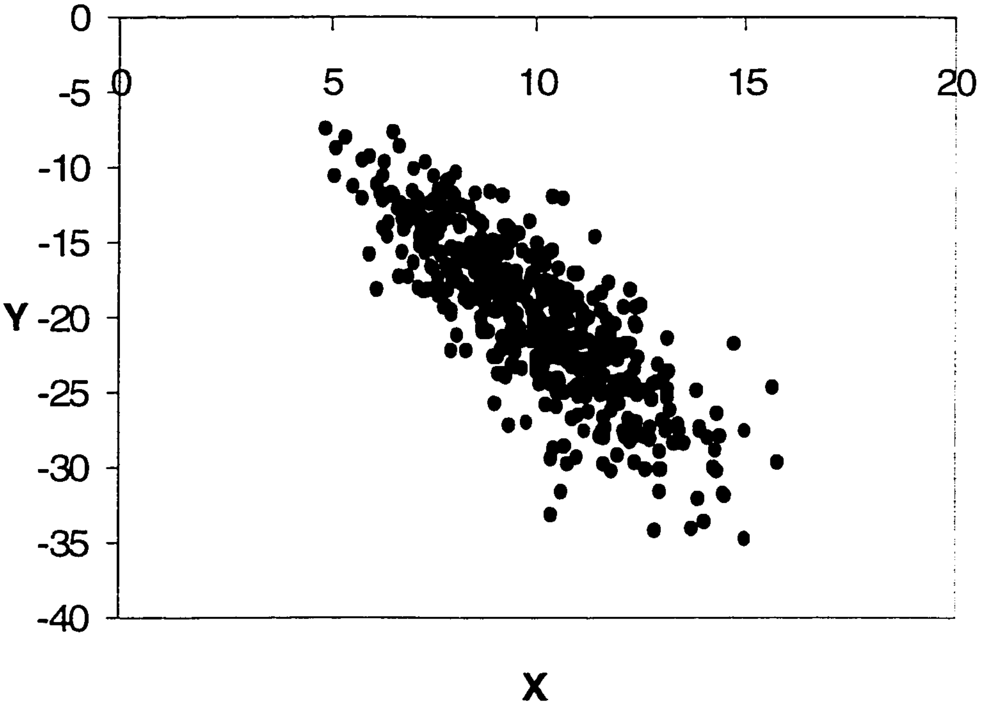


Figure 3.2: Scatterplot of Nonlinear Simulated Population from Model

$$Y = 150 + 2.5e^{-X} + X^{\frac{1}{2}}\varepsilon$$

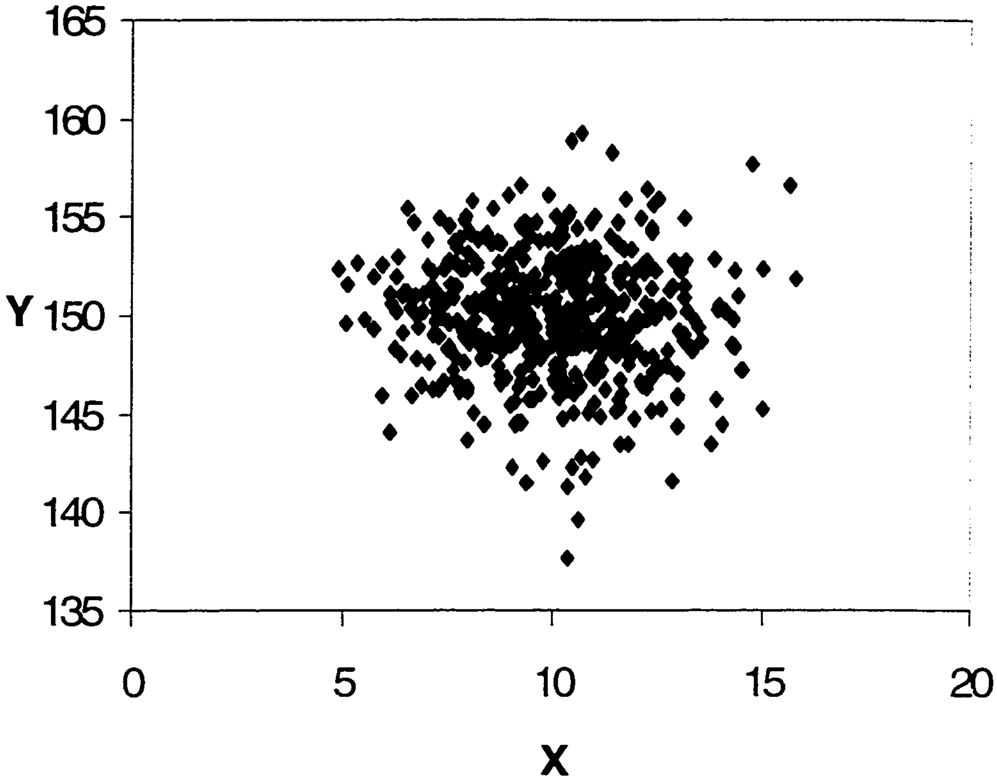


Table 3.1: Simulation Results with Linear Simulated Population Data: Relative Biases (RB), Mean Square Errors (MSE) and Relative Root Mean Square Errors (RRMSE) of Different Distribution Function Estimators (Population Size = 500, Sample Size = 30, Number of Simulation = 1000.)

F_p	0.252	0.502	0.752
$RB(\hat{F}_{ma})$	**0.181576**	**0.099580**	**0.059823**
$RB(\hat{F}_{pma})$	0.197171	0.104532	0.065344
$RB(\hat{F}_{pe})$	0.197316	0.103682	0.065770
$MSE(\hat{F}_{ma})$	**0.003304**	**0.003885**	**0.003155**
$MSE(\hat{F}_{pma})$	0.003858	0.004327	0.003706
$MSE(\hat{F}_{pe})$	0.004014	0.004286	0.003881
$RRMSE(\hat{F}_{ma})$	**0.228111**	**0.124161**	**0.074691**
$RRMSE(\hat{F}_{pma})$	0.246488	0.131031	0.080958
$RRMSE(\hat{F}_{pe})$	0.251428	0.130407	0.082846

Table 3.2: Simulation Results with Nonlinear Simulated Population Data: Relative Biases (RB), Mean Square Errors (MSE) and Relative Root Mean Square Errors (RRMSE) of Different Distribution Function Estimators (Population Size = 500, Sample Size = 30, Number of Simulation = 1000.)

F_p	0.252	0.502	0.752
$RB(\hat{F}_{ma})$	0.297976	0.162664	0.096423
$RB(\hat{F}_{pma})$	0.256258	0.142478	**0.081615**
$RB(\hat{F}_{pe})$	**0.252221**	**0.142223**	0.083809
$MSE(\hat{F}_{ma})$	0.008937	0.010601	0.008348
$MSE(\hat{F}_{pma})$	0.006503	**0.008131**	**0.005884**
$MSE(\hat{F}_{pe})$	**0.006225**	0.008214	0.006146
$RRMSE(\hat{F}_{ma})$	0.375137	0.205103	0.121496
$RRMSE(\hat{F}_{pma})$	0.320012	**0.179624**	**0.102006**
$RRMSE(\hat{F}_{pe})$	**0.313081**	0.180538	0.104247

Figure 3.3a: Plot of Population Distribution Function F_p and Model Assisted Estimator $\hat{F}_{ma}$ with Linear Simulated Population

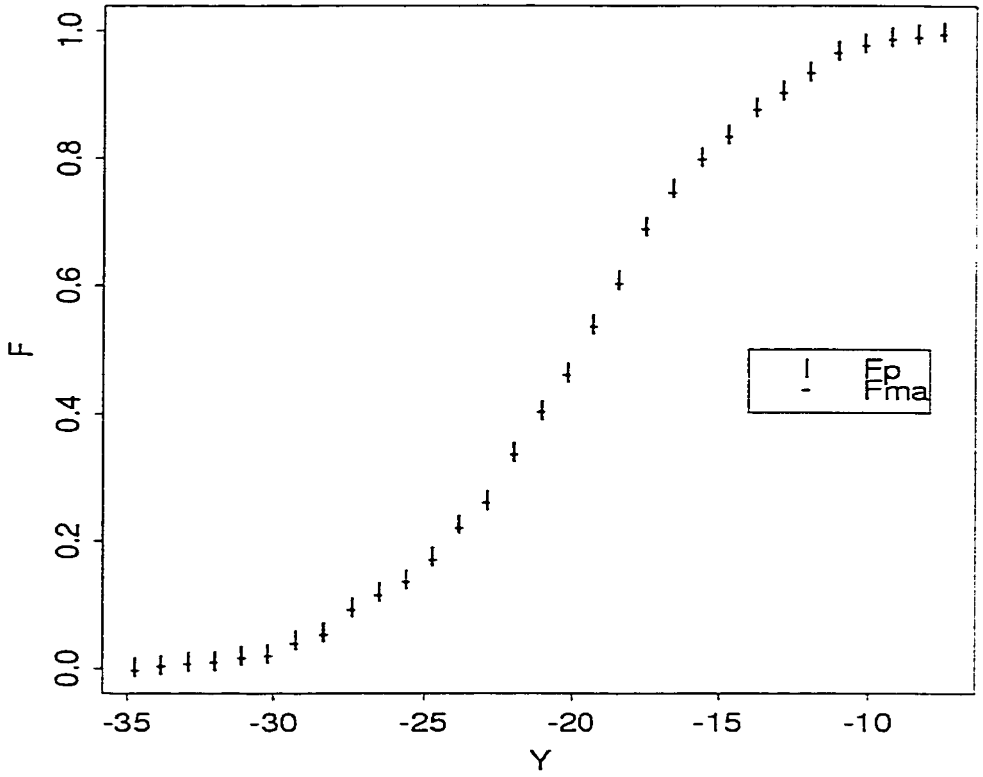


Figure 3.3b: Plot of Population Distribution Function F_p and Proposed Estimator $\hat{F}_{pma}$ with Linear Simulated Population

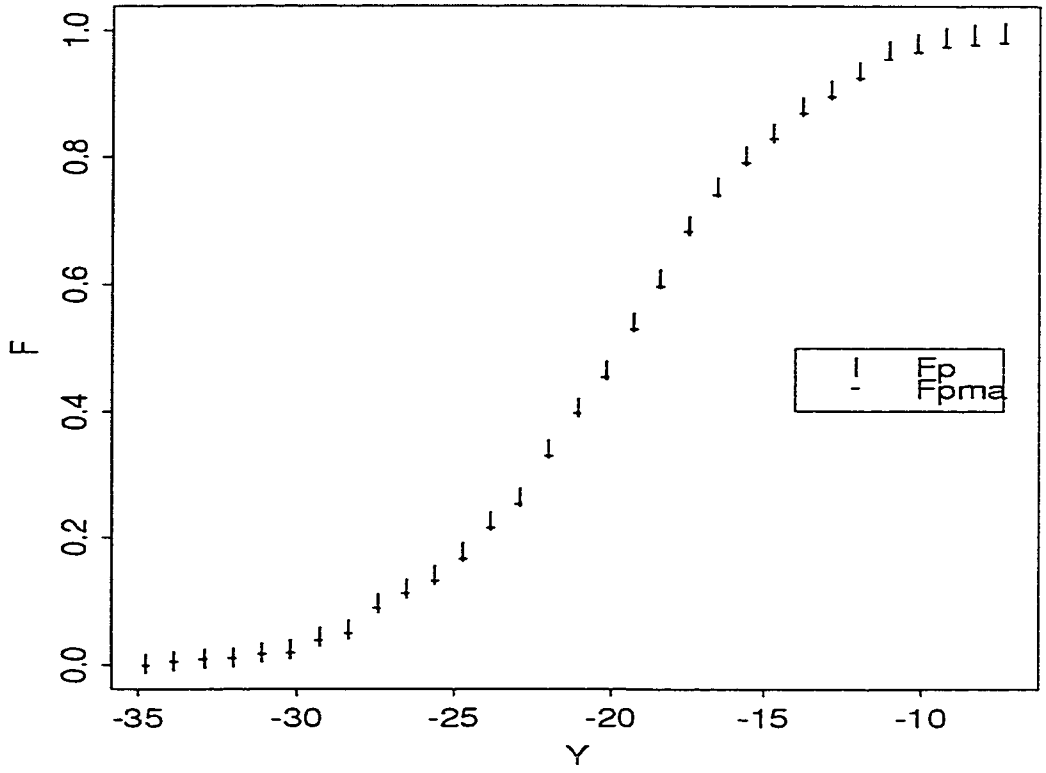


Figure 3.3c: Plot of Population Distribution Function F_p and Pseudo Empirical Likelihood Estimator $\hat{F}_{pe}$ with Linear Simulated Population

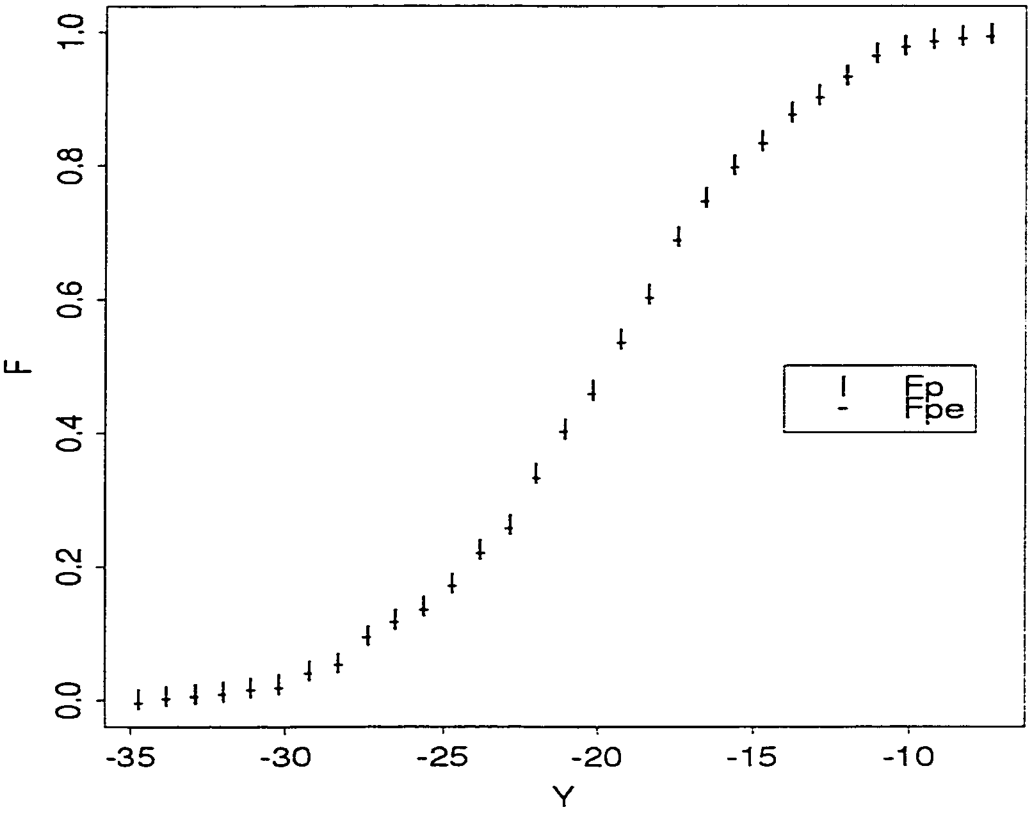


Figure 3.4a: Plot of Population Distribution Function F_p and Model Assisted Estimator $\hat{F}_{ma}$ with Nonlinear Simulated Population

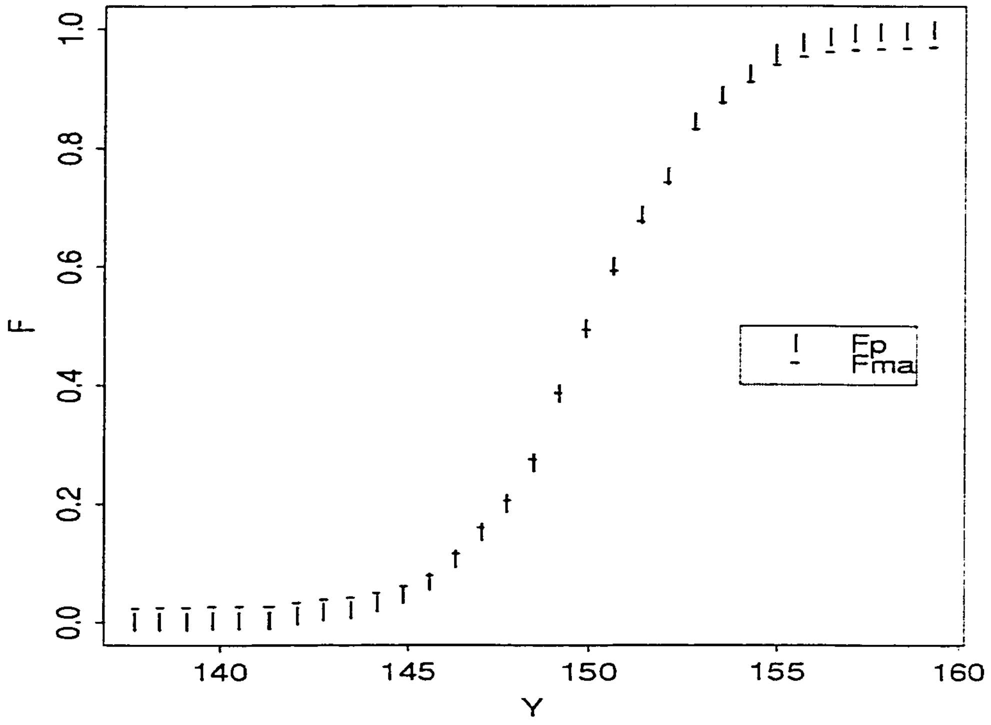


Figure 3.4b: Plot of Population Distribution Function F_p and Proposed Estimator $\hat{F}_{pma}$ with Nonlinear Simulated Population

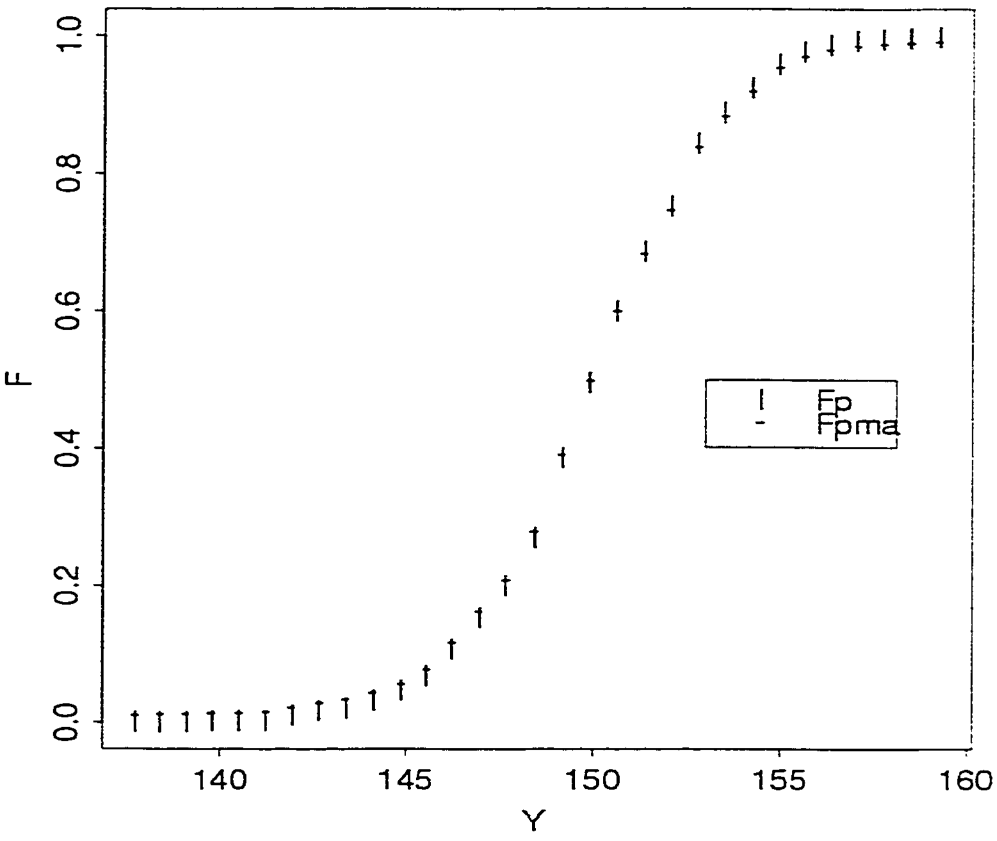
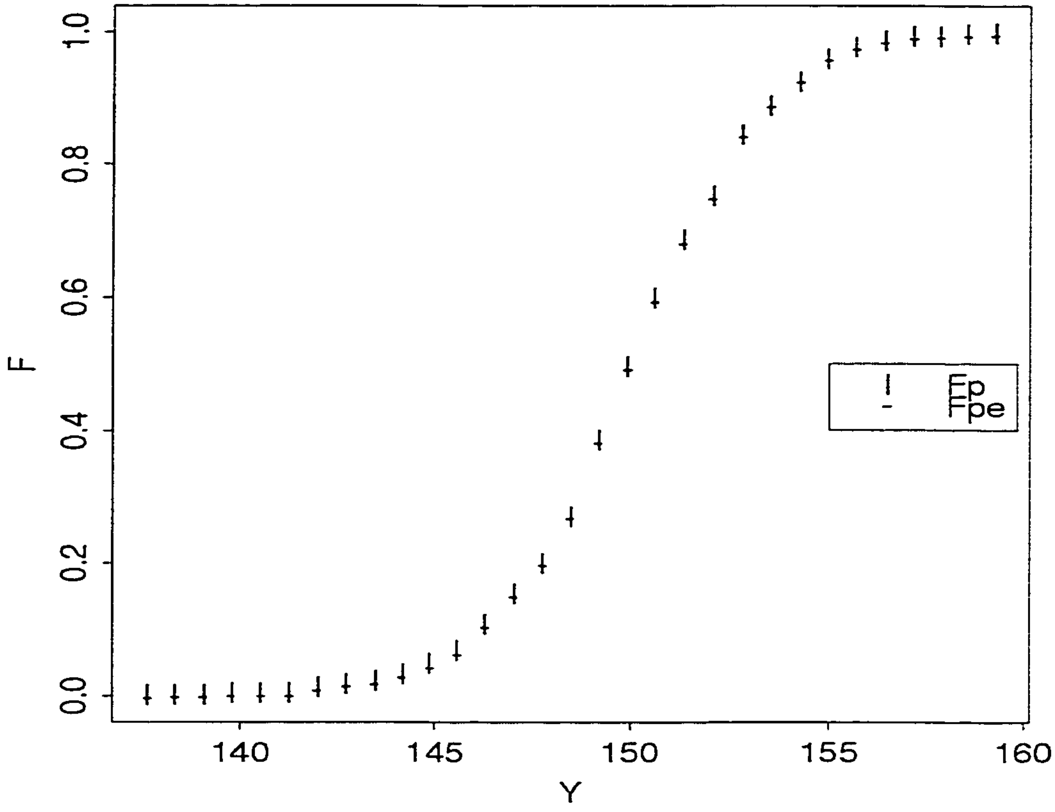


Figure 3.4c: Plot of Population Distribution Function F_p and Pseudo Empirical Likelihood Estimator $\hat{F}_{pe}$ with Nonlinear Simulated Population



The effect of kernel on the model-assisted estimator

When the proposed model-assisted estimator $\hat{F}_{pma}$ is used to estimate the population distribution function and associate quantiles, an appropriate kernel function is needed. And the associate quantile esimator $\hat{q}_{pma}$ is just the inverse of the distribution function estimator $\hat{F}_{pma}$. So in this section, only effects of kernel functions on the proposed distribution function estimator are being examined by comparing the following seven common kernels:

Table 3.3: Common Kernel Function

Name	Function
Uniform	$\frac{1}{2}I(\|u\| \leq 1)$
Triangle	$(1-\|u\|)I(\|u\| \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)I(\|u\| \leq 1)$
Quartic	$\frac{15}{16}(1-u^2)^2I(\|u\| \leq 1)$
Triweight	$\frac{35}{32}(1-u^2)^3I(\|u\| \leq 1)$
Cosinus	$\frac{\pi}{4}\cos(\frac{\pi}{2}u)I(\|u\| \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$

The estimation results are given in Table 3.4 $\backsim$ 3.5.

Table 3.4: **Relative Biases (RB) and Relative Root Mean Square Errors (RRMSE) of $\hat{F}_{pma}$ at 0.25, 0.50 and 0.75th quantile of Y under Different Kernels with Linear Simulated Population**

Kernel Type	α	**RB** ($\hat{\mathbf{F}}_{pma}(t)$ at $t = \alpha^{th}$ quantile)	**RRMSE** ($\hat{\mathbf{F}}_{pma}(t)$ at $t = \alpha^{th}$ quantile)
Uniform	0.25	0.197171	0.246488
	0.50	**0.104532**	0.131031
	0.75	**0.065344**	**0.080958**
Triangle	0.25	**0.196066**	**0.245604**
	0.50	0.105077	0.130627
	0.75	0.065938	0.081446
Epanechnikov	0.25	0.196504	0.245687
	0.50	0.104650	**0.130567**
	0.75	0.065620	0.081175
Quartic	0.25	0.197204	0.247660
	0.50	0.105155	0.131291
	0.75	0.065949	0.081499
Triweight	0.25	0.197888	0.248405
	0.50	0.105654	0.132177
	0.75	0.066328	0.082121
Cosinus	0.25	0.196497	0.245935
	0.50	0.104901	0.130454
	0.75	0.065759	0.081319
Gaussian	0.25	0.206267	0.259046
	0.50	0.108356	0.136323
	0.75	0.065838	0.082678

Table 3.5: Relative Biases (RB) and Relative Root Mean Square Errors (RRMSE) of $\hat{F}_{pma}$ at 0.25, 0.50 and 0.75th quantile of Y under Different Kernels with Nonlinear Simulated Population

Kernel Type	α	**RB** ($\hat{\mathbf{F}}_{pma}(t)$ at $t = \alpha^{th}$ quantile)	**RRMSE** ($\hat{\mathbf{F}}_{pma}(t)$ at $t = \alpha^{th}$ quantile)
Uniform	0.25	**0.256258**	**0.320012**
	0.50	**0.142478**	**0.179624**
	0.75	**0.081615**	**0.102006**
Triangle	0.25	0.260487	0.323601
	0.50	0.143233	0.179927
	0.75	0.082615	0.103401
Epanechnikov	0.25	0.259692	0.322613
	0.50	0.143080	0.179817
	0.75	0.082315	0.102867
Quartic	0.25	0.261260	0.324810
	0.50	0.143631	0.180126
	0.75	0.082880	0.103578
Triweight	0.25	0.261561	0.325889
	0.50	0.143937	0.180817
	0.75	0.083200	0.103901
Cosinus	0.25	0.259759	0.322867
	0.50	0.143231	0.179971
	0.75	0.082427	0.103087
Gaussian	0.25	0.297373	0.374064
	0.50	0.162365	0.204364
	0.75	0.095283	0.120422

The above results demonstrate that no matter which kind of model the underlying population data truly comply to, in terms of biases and efficiency, as measured by RB, MSE and $RRMSE$, the estimator of distribution function obtained from (3.8) has a better performance when the Uniform kernel function is used. Especially, when the underlying relationship between Y and X is nonlinear, the use of Uniform kernel function produces an estimator with the smallest bias and the highese efficiency compared to other kernel functions.

However, when any one of the kernel function 1 ~ 7 is used in the process of estimating distribution function, the estimator $\hat{F}_{pma}$ always has better performance than $\hat{F}_{ma}$ when the underlying distribution function is nonlinear. So in the following the Uniform kernel function is still used to obtain $\hat{q}_{pma}$.

Comparison among different quantile estimators

In this simulation, five kinds of α^{th} quantile estimator: $\hat{q}_{ma}$, $\hat{q}_{pma}$, $\hat{q}_{pe}$, $\hat{q}_{r}$, $\hat{q}_{d}$ are compared. The first three stand for the inversion of $\hat{F}_{ma}$, $\hat{F}_{pma}$ and $\hat{F}_{pe}$; the latter two are the ratio estimator, the difference estimator with the auxiliary information used, respectively. The results are given in Table 3.6 ~ 3.7.

For the linear simulated population data, in terms of biases and efficiency, as measured by RB and $RRMSE$, estimator $\hat{q}_{ma}$ has the best performance among these five estimators, but at the 0.75^{th} quantile of Y, the bias of $\hat{q}_{ma}$ is larger than that of ratio estimator $\hat{q}_{r}$. $\hat{q}_{pe}$ has better performance than $\hat{q}_{pma}$ except in the case of the 0.25^{th} quantile. Generally, the first three estimators have smaller biases, as measured by RB, and more efficiency, as measured$RRMSE$ than the last two estimators except at the 0.75^{th} quantile, $\hat{q}_r$ is better than $\hat{q}_{pe}$ and $\hat{q}_{pma}$. So the first three estimators are appropriate to be used. In particular, $\hat{q}_{ma}$ is encouraged to be used when strong linear relationship exists in the underlying population.

For the nonlinear simulated population data, in terms of biases and efficiency, as measured by RB and $RRMSE$, estimator $\hat{q}_{pma}$ is less biased and more efficient than $\hat{q}_{ma}$, and even less biased than $\hat{q}_{pe}$ in the median (0.5^{th} quantile) of Y. However, $\hat{q}_{pe}$ has the highest efficiency at all of these three quantiles, and in the 0.25^{th} and 0.75^{th} quantiles, $\hat{q}_{pe}$ also has the smallest bias. In the 0.5^{th} and 0.75^{th} quantiles, $\hat{q}_r$ has a higher efficiency than $\hat{q}_{pma}$. So in the case of nonlinear population as an interested population, $\hat{q}_{pe}$ and $\hat{q}_{pma}$ are suggested to be used when compared to $\hat{q}_{ma}$.

Table 3.6: Simulation Results with Simulated Linear Population Data: Relative Biases (RB) and Relative Root Mean Square Errors (RRMSE) of Different α^{th} Quantile $q(\alpha)$ Estimators (Population Size $= 500$, Sample Size $= 30$, Simulation Times $= 1000$.)

α	0.25	0.50	0.75
q_p	-23.2445	-19.5589	-16.5630
$RB(\hat{q}_{ma})$	**0.036742**	**0.034524**	0.058889
$RB(\hat{q}_{pma})$	0.039491	0.036819	0.064503
$RB(\hat{q}_{pe})$	0.039821	0.035828	0.063712
$RB(\hat{q}_{r})$	0.056092	0.037986	**0.058158**
$RB(\hat{q}_{d})$	0.048555	0.038254	0.062231
$RRMSE(\hat{q}_{ma})$	**0.044007**	**0.043401**	**0.071810**
$RRMSE(\hat{q}_{pma})$	0.047948	0.045569	0.080772
$RRMSE(\hat{q}_{pe})$	0.048059	0.044996	0.078924
$RRMSE(\hat{q}_{r})$	0.071937	0.047087	0.071958
$RRMSE(\hat{q}_{d})$	0.061875	0.047420	0.077631

Table 3.7: Simulation Results with Simulated Nonlinear Population Data: Relative Biases (RB) and Relative Root Mean Square Errors (RRMSE) of Different α^{th} Quantile $q(\alpha)$ Estimators (Population Size = 500, Sample Size = 30, Simulation Times = 1000.)

α	0.25	0.50	0.75
q_p	148.341	149.908	151.988
$RB(\hat{q}_{ma})$	0.004443	0.004151	0.004928
$RB(\hat{q}_{pma})$	0.003776	**0.003554**	0.004089
$RB(\hat{q}_{pe})$	**0.003638**	0.003710	**0.004034**
$RB(\hat{q}_r)$	0.047821	0.035158	0.037198
$RB(\hat{q}_d)$	0.041321	0.035177	0.041383
$RRMSE(\hat{q}_{ma})$	0.058523	0.053216	0.063124
$RRMSE(\hat{q}_{pma})$	0.050050	0.045433	0.051267
$RRMSE(\hat{q}_{pe})$	**0.004770**	**0.004760**	**0.005054**
$RRMSE(\hat{q}_r)$	0.060386	0.044779	0.045555
$RRMSE(\hat{q}_d)$	0.051528	0.044488	0.050660

Chapter 4
Distribution Function and Quantile Estimation in Two-stage Sampling

In this chapter, we consider the estimation of distribution function and α^{th} quantile of a finite population under two-stage sampling scheme. In practice, under the most simple SRSWOR/SRSWOR sampling plan, there are already several conventional estimators of distribution function developed under design-based inference methods. However, the estimation under a superpopulation model is considered in this chapter. Section 4.1 gives a brief introduction to the conventional estimators. In section 4.2, along the same lines of chapter 3 and extending the results for estimating population total to population distribution function, a distribution function estimator is proposed. Some of alternative estimators are given in section 4.3. Finally, in section 4.4, the proposed estimator is compared with conventional estimators under SRSWOR/SRSWOR sampling plan through simulation study.

4.1 Two-stage Distribution Function Estimator in Conventional Theory

Consider a finite population which contains K elements and are arranged in N clusters, the ith cluster is known to contain M_i elements, so that $\sum_{i=1}^{N} M_i = K$. Then the population can be represented as $U = \{Y_{ij},\ i = 1, ..., N;\ j = 1, ..., M_i\}$. First a sample s of size n units is taken out of N clusters at the first stage and then a sample s_i of size m_i for $i \in s$ is chosen at the second stage. According to the conventional sampling theory, there are several types of estimator for two-stage cluster sampling. For the most common and simple design: simple random sampling without replacement (**SRSWOR**) in both stages, the usual estimators are (see Royal, 1976):

(i) the expansion estimator

$$\bar{F}_e = \frac{\bar{T}_e}{K} = \frac{\sum_{i \in s} m_i \bar{\Delta}_{s_i}}{k}, \tag{4.1}$$

(ii) the "unbiased" estimator

$$\bar{F}_u = \frac{\bar{T}_u}{K} = \frac{\sum_{i \in s} M_i \bar{\Delta}_{s_i}}{n\bar{M}}, \tag{4.2}$$

(iii) the "ratio-type" estimator

$$\bar{F}_r = \frac{\bar{T}_r}{K} = \frac{\sum_{i \in s} M_i \bar{\Delta}_{s_i}}{n\bar{M}_s}, \tag{4.3}$$

where,

$$k = \sum_{i \in s} m_i, \tag{4.4}$$

$$\bar{M} = \frac{K}{N}, \tag{4.5}$$

$$\bar{M}_s = \frac{1}{n} \sum_{i \in s} M_i, \tag{4.6}$$

and

$$\bar{\Delta}_{s_i} = \frac{1}{m_i} \sum_{j \in s_i} \Delta(t - y_{ij}), \text{ for } i \in s. \tag{4.7}$$

Remark: The estimator $\bar{F}_u$ is unbiased only if $\bar{M}_s$ happens to equal $\bar{M}$ (Royal, 1976). Where as, in general, $\bar{F}_e$ and $\bar{F}_r$ are biased estimator.

In the next section, estimation of distribution function under a superpopulation model will be consdered.

4.2 Proposed Method of Estimation: Distribution Function in Two-stage Sampling

Assume the population U is generated from a two-stage superpopulation model:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, ..., N, \quad j = 1, ..., M_i. \tag{4.8}$$

By denoting $E_m(\cdot)$ for expectation operators under the above model, it is assumed that (see Scott and Smith, 1969)

$$\begin{aligned} & E_m(Y_{ij}|\mu_i) = \mu_i, \\ & E_m[(Y_{ij} - \mu_i)(Y_{i'j'} - \mu_{i'})|\mu_i, \mu_{i'}] = \begin{cases} \sigma_\varepsilon^2, & i = i', j = j', \\ 0, & \text{otherwise.} \end{cases} \\ & E_m(\mu_i) = \mu, \\ & E_m[(\mu_i - \mu)(\mu_{i'} - \mu)] = \begin{cases} \tau^2, & i = i', \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{4.9}$$

So, $\{Y_{ij}\}$'s are identically distributed random variables with finite population distribution function $F_Y(t)$, which is defined as

$$F_Y(t) := \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \Delta(t - Y_{ij}). \tag{4.10}$$

In current case, the population has nothing to do with the time of sampling. It is easy to see that it satisfies the assumption A in Olsson and Rootzèn's paper (1996), *i.e.*,

(i) The finite population distribution function $F_Y(t)$ of Y is (piecewise) continuous.

(ii) The N clusters are independent, so the vectors $(Y_{i1},...,Y_{iM_i})$, $i = 1,...,N$, are independent.

(iii) For each i, the correlation coefficient between $\Delta(t - Y_{ij})$ and $\Delta(t - Y_{ij'})$ depends only on t, so we can denote it as $\rho_\Delta(t)$.

Now suppose that it is needed to estimate the distribution function $F_Y(t)$ of Y and the associate α^{th} quantile. A two-stage sampling scheme is conducted to get the observation data. That is, at the first stage, n out of N clusters are chosen through some sampling scheme to get a primary sample s; at the second stage, m_i out of M_i elements of each selected cluster $i \in s$ are chosen under designed sampling scheme to get a subsample s_i. Since $K = \sum_{i=1}^N M_i$, then the same decomposition method as that in chapter 2 can be used to decompose the finite population distribution function $F_Y(t)$ as follows:

$$\begin{aligned} F_Y(t) &= \frac{1}{K}\{\sum_{i\in s}\sum_{j=1}^{M_i}\Delta(t-Y_{ij}) + \sum_{i\in \bar{s}}\sum_{j=1}^{M_i}\Delta(t-Y_{ij})\} \\ &= \frac{1}{K}\{\sum_{i\in s}\sum_{j\in s_i}\Delta(t-Y_{ij}) + \sum_{i\in s}\sum_{j\in \bar{s}_i}\Delta(t-Y_{ij}) \\ &\quad +\sum_{i\in \bar{s}}\sum_{j=1}^{M_i}\Delta(t-Y_{ij})\} \qquad (4.11) \\ &= \frac{1}{K}\{T_I + T_{II} + T_{III}\}. \end{aligned}$$

Notice that only the first term T_I in (4.11) is completely known after the sampling, so the other two terms, T_{II} and T_{III}, have to be predicted from the observation data in order to obtain the estimator of the finite population distribution function $F_Y(t)$. When all the population cluster sizes, M_i $(i = 1, ..., N)$, are known, estimating $F_Y(t)$ is equivalent to estimating $T_\Delta(t)$ – the population total of $\Delta(t - Y)$, which is defined as

$$T_\Delta(t) := \sum_{i=1}^{N} \sum_{j=1}^{M_i} \Delta(t - Y_{ij}). \tag{4.12}$$

Hence the prediction method (Royall, 1976) can be used to get an unbiased estimator of T_Δ. Firstly, observe that

$$E[\Delta(t - Y_{ij}) - F_Y(t)] = 0, \tag{4.13}$$

and

$$Cov\{(\Delta(t - Y_{ij}), \Delta(t - Y_{i'j'}))|U\} = \begin{cases} 0, & \text{for } i \neq i'; \\ \rho_\Delta \sigma^2_\Delta, & \text{for } i = i' \quad j \neq j'; \\ \sigma^2_\Delta, & \text{for } i = i' \quad j = j', \end{cases} \tag{4.14}$$

where,

$$\rho_\Delta = \rho_\Delta(t) := \frac{E\{[\Delta(t - Y_{ij}) - F_Y(t)][\Delta(t - Y_{ij'}) - F_Y(t)]|U\}}{E\{[\Delta(t - Y_{ij}) - F_Y(t)]|U\}^2}, \text{ for } 1 \leq j \neq j' \leq M_i, \tag{4.15}$$

which is the common correlation coefficient between $\Delta(t - Y_{ij})$ and $\Delta(t - Y_{ij'})$ within cluster i, and

$$\sigma^2_\Delta(t) := Var(\Delta(t - Y_{ij})|U) = F_Y(t)(1 - F_Y(t)), \tag{4.16}$$

is the common variance of $\Delta(t - Y_{ij})$. Then under the above model, by assuming ρ_Δ is known, $T_\Delta(t)$ (see (4.12)) can be predicted as,

$$\begin{aligned} \hat{T}_\Delta(t) = \; & \textstyle\sum_{i \in s} \sum_{j \in s_i} \Delta(t - y_{ij}) \\ & + \textstyle\sum_{i \in s} (M_i - m_i)[\omega_i \bar{\Delta}_{s_i} + (1 - \omega_i)\hat{F}_0] \\ & + \textstyle\sum_{i \in \bar{s}} M_i \hat{F}_0, \end{aligned} \tag{4.17}$$

where

$$\omega_i = \frac{\rho_\Delta m_i}{1 - \rho_\Delta + m_i \rho_\Delta}, \text{ for } i \in s. \tag{4.18}$$

Hence, the resulting estimator for $F_Y(t)$, which is termed as **predictive estimator**, is given by:

$$\begin{aligned} \hat{F}_g(t) &= \frac{1}{K}\hat{T}_\Delta(t) \qquad (4.19) \\ &= \frac{1}{K}\{\sum_{i \in s} \sum_{j \in s_i} \Delta(t - y_{ij}) + \sum_{i \in s}(M_i - m_i)[\omega_i \bar{\Delta}_{s_i} + (1 - \omega_i)\hat{F}_0] + \sum_{i \in \bar{s}} M_i \hat{F}_0\}, \end{aligned}$$

where,

$$\hat{F}_0 = \sum_{i \in s} u_i \bar{\Delta}_{s_i},$$

with the weight given by

$$u_i = \frac{m_i/(1 - \rho_\Delta + m_i \rho_\Delta)}{\sum_{i \in s} m_i/(1 - \rho_\Delta + m_i \rho_\Delta)}, \quad \text{for } i \in s. \tag{4.20}$$

Along the lines of Royall (1976), it can be shown that the above estimator $\hat{F}_g(t)$ is the best unbiased estimator for the finite population parameter $F_Y(t)$ when σ^2_Δ and ρ_Δ are known, and the mean square error of $\hat{F}_g(t)$ is

$$
\begin{aligned}
MSE(\hat{F}_g(t)) &= \frac{\sigma_\Delta^2}{K^2}\{(1-\rho_\Delta)(K-k)+\rho_\Delta\sum_{i\in\bar{s}}M_i^2 \\
&\quad +\rho_\Delta\sum_{i\in s}\frac{(M_i-m_i)^2(1-\rho_\Delta)}{1-\rho_\Delta+m_i\rho_\Delta} \\
&\quad +\frac{[K-\sum_{i\in s}m_i(1-\rho_\Delta+M_i\rho_\Delta)/(1-\rho_\Delta+m_i\rho_\Delta)]^2}{\sum_{i\in s}m_i/(1-\rho_\Delta+m_i\rho_\Delta)}\}. \quad (4.21)
\end{aligned}
$$

In practice, values of σ_Δ^2 and ρ_Δ are usually unknown. However, the estimate for σ_Δ^2 and ρ_Δ can be used to obtain an approximately optimal estimator of the interested parameter.

A simple estimator of $\rho_\Delta(t)$, given in Olsson and Rootžen(1996), is as follows. Denote $m = \#\{i : m_i \neq 1 \text{ and } i \in s\}$ and set

$$\hat{\sigma}_\Delta^2(t) = \frac{1}{m}\sum_{i\in s \text{ and } m_i\neq 1}\frac{\sum_{j\in s_i}\{\Delta(t-y_{ij})-\bar{F}(t)\}^2}{m_i}. \quad (4.22)$$

So

$$\hat{\rho}_\Delta(t) = \frac{1}{m\hat{\sigma}_\Delta^2(t)}\sum_{i\in s \text{ and } m_i\neq 1}\frac{\sum_{1\leq j\neq j'\leq m_i}\{\Delta(t-y_{ij})-\bar{F}(t)\}\{\Delta(t-y_{ij'})-\bar{F}(t)\}}{m_i(m_i-1)}, \quad (4.23)$$

with

$$\bar{F}(t) = \frac{1}{n}\sum_{i\in s}\frac{1}{m_i}\sum_{j\in s_i}\Delta(t-y_{ij}). \quad (4.24)$$

4.3 Alternative Estimators

This section gives some discussion on alternative estimators of the finite distribution function in two-stage sampling. The corresponding quantile estimators can be obtained from formula (1.5) with different distribution function estimator plugged in.

At first, for the easy use of the predictive estimator (4.19), the simple estimator $\bar{F}(t)$ of the distribution function is used in the process of estimating $\sigma^2_\Delta(t)$ and $\rho_\Delta(t)$. In practice, for simplicity, $\bar{F}(t)$ can be replaced by $\bar{F}_1(t)$,

$$\bar{F}_1(t) = \frac{1}{k}\sum_{i\in s}\sum_{j\in s_i}\Delta(t-y_{ij}), \tag{4.25}$$

when it is reasonable to ignore the dependence within each cluster. When all the sample size $m_i's$ are equal, they will result in the same estimator (Royall, 1976).

Some additional estimators can be obtained as special cases of the predictive estimator (4.19). If $\rho_\Delta = 0$, then $\omega_i = 0$ and $u_i = \frac{m_i}{k}$. So $\hat{F}_0 = \sum_{i\in s} u_i\bar{\Delta}_{s_i} = \bar{F}_e$, in this case the optimal estimator (4.19) becomes

$$\hat{F}_{g0} = \frac{1}{K}\{\sum_{i\in s}\sum_{j\in s_i}\Delta(t-y_{ij}) + \sum_{i\in s}(M_i - m_i)\bar{F}_e + \sum_{i\in\bar{s}} M_i\bar{F}_e\}. \tag{4.26}$$

This is actually the conventional expansion estimator $\bar{F}_e$.

When $\rho_\Delta = 1$, then $\omega_i = 1$, $u_i = \frac{1}{n}$, $\hat{F}_0 = \sum_{i\in s} u_i\bar{\Delta}_{s_i} := \bar{F}_p$. In this case, the optimal estimator can be represented as

$$\hat{F}_{g1} = \frac{1}{K}\{\sum_{i\in s}\sum_{j\in s_i}\Delta(t-y_{ij}) + \sum_{i\in s}(M_i - m_i)\bar{\Delta}_{s_i} + \sum_{i\in\bar{s}} M_i\bar{F}_p\}. \tag{4.27}$$

This also suggests that the conventional estimators $\bar{F}_e$ as in (4.1) and $\bar{F}_r$ as in (4.3) be replaced in the third part of the predictive estimator, that gives

$$\hat{F}_{ge} = \frac{1}{K}\{\sum_{i\in s}\sum_{j\in s_i}\Delta(t-y_{ij}) + \sum_{i\in s}(M_i - m_i)\bar{\Delta}_{s_i} + \sum_{i\in \bar{s}} M_i\bar{F}_e\}, \tag{4.28}$$

and

$$\hat{F}_{gr} = \frac{1}{K}\{\sum_{i\in s}\sum_{j\in s_i}\Delta(t-y_{ij}) + \sum_{i\in s}(M_i - m_i)\bar{\Delta}_{s_i} + \sum_{i\in \bar{s}} M_i\bar{F}_r\}. \tag{4.29}$$

Another form with $\bar{F}_p$ used in the third part of the predictive estimator can be given as:

$$\hat{F}_{gp} = \frac{1}{K}\{\sum_{i\in s}\sum_{j\in s_i}\Delta(t-y_{ij}) + \sum_{i\in s}(\bar{M}_s - m_i)\bar{\Delta}_{s_i} + \sum_{i\in \bar{s}} M_i\bar{F}_p\}. \tag{4.30}$$

And, actually, $\hat{F}_{gr}$ and $\hat{F}_{gp}$ are just the predictive estimator representation of $\bar{F}_r$ and $\bar{F}_p$.

The difference between (4.27) and (4.30) only exists in the second term which predicts $\sum_{i\in s}\sum_{j\in \bar{s}_i}\Delta(t-Y_{ij})$.

In the next section, results from a Monto Carlo simulation will be presented to compare estimators $\bar{F}_e$ as in (4.1), $\bar{F}_r$ as in (4.3) and $\bar{F}_u$ as in (4.2) with $\hat{F}_g$ as in (4.19) and $\hat{F}_{ge}$ as in (4.28) under SRSWOR/SRSWOR sampling scheme.

4.4 Simulation Study

In this section, the simulation study for some special cases of two-stage sampling process is presented.

Since two-stage sampling scheme is usually used in moderate or large populations. Here, three large simulated populations are generated using two-stage sampling procedure. In each population, it is assumed that there are $N = 200$ clusters and the elements have a common correlation coefficient, $\rho(Y_{ij}, Y_{ij'})$ within each cluster, for $j \neq j'$ and $i = 1, ..., N$. The cluster sizes in each population considered in this study are:

$$M_i = \begin{cases} 80, & \text{for } i = 1, ..., 50; \\ 100, & \text{for } i = 51, ..., 100; \\ 120, & \text{for } i = 101, ..., 200. \end{cases} \tag{4.31}$$

Finally we generate the data using the following procedure.

Step 1: Generate cluster means $\mu_1, \mu_2, ..., \mu_N$ of each cluster such that $\mu_i \sim N(\mu, \sigma_\mu^2)$ independently, where $\mu = 100.0$, and

$$\sigma_\mu = \begin{cases} 0.81, & \text{Population 1,} \\ 0.91, & \text{Population 2,} \\ 50.91, & \text{Population 3.} \end{cases} \tag{4.32}$$

Step 2: Within each cluster i, $\varepsilon_{i1}, ..., \varepsilon_{iM_i}$ are generated such that $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, for $j = 1, ..., M_i$, where

$$\sigma_\varepsilon = \begin{cases} 100.7, & \text{Population 1,} \\ 0.9, & \text{Population 2,} \\ 0.9, & \text{Population 3.} \end{cases} \tag{4.33}$$

Step 3: Generate the population data of size $K = 21,000$ $\{Y_{ij}, i = 1, ..., N; j = 1, ..., M_i\}$ as:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \text{ for } i = 1, ..., N; \; j = 1, ..., M_i. \tag{4.34}$$

Then, note that in each population, the correlation coefficient within cluster i is:

$$\rho_Y = \rho(Y_{ij}, Y_{ij'}) = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\varepsilon^2}, \text{ for } 1 \leq j \neq j' \leq M_i \text{ and } i = 1, ..., N. \tag{4.35}$$

In this simulation study, we consider three kinds of population in which every cluster has a correlation coefficient satisfying: i) $\rho_\Delta \approx 0$; ii) $0 < \rho_\Delta < 1$; and iii) $\rho_\Delta \approx 1$ by assigning ρ_Y satisfying i), ii) and iii) too. The three simulated populations generated in section 4.4.1 have the correlation coefficients: $\rho_Y = 6.46968e^{-5}$, $\rho_Y = 0.505525$ and $\rho_Y = 0.999688$ respectively. But in the process of estimating distribution function, σ_Δ and ρ_Δ are assumed to be unknown. They can be estimated through (4.22) and (4.23). The estimates $\hat{\sigma}_\Delta$ and $\hat{\rho}_\Delta$ are plugged in the predictive estimator to obtain corresponding distribution function estimate $\hat{F}_g(t)$. We consider two cases: case (i) with moderate sample size within cluster:

$$m_i = \begin{cases} 15, & \text{for } i = 1, ..., 5; \\ 20, & \text{for } i = 6, ..., 15; \quad \text{for } i \in s, \\ 25, & \text{for } i = 16, ..., 20; \end{cases} \tag{4.36}$$

so that $k = 400$; and case (ii) with small sample size within cluster:

$$m_i = \begin{cases} 10, & \text{for } i = 1, ..., 5; \\ 12, & \text{for } i = 6, ..., 15; \quad \text{for } i \in s, \\ 15, & \text{for } i = 16, ..., 20; \end{cases} \tag{4.37}$$

so that $k = 245$.

In both cases small sample is selected using two-stage sampling plan by first selecting 20 clusters from 200 clusters using SRSWOR. Within ith selected cluster, m_i units, as defined in (4.36) for case 1 and in (4.37) for case 2, are selected using SRSWOR.

After $10,000$ simulation runs under above mentioned three populations, simulation results become stable, then the relative biases (RB), relative root mean square errors ($RRMSE$) and relative efficiency (RE) with respect to $\bar{F}_u$ (or $\bar{q}_u$) are computed for the proposed estimators using $N_{simu} = 10,000$. RB and $RRMSE$ are computed using (3.14) and (3.16) respectively; while RE is computed as follows.

Let $\bar{\theta}_u$ denote the unbiased estimator of θ, $\hat{\theta}$ denote any estimator of θ, then we define

$$RE(\hat{\theta}) = \text{relative efficiency of } \hat{\theta} \text{ with respect to } \bar{\theta}_u := \frac{RRMSE(\bar{\theta}_u)}{RRMSE(\hat{\theta})}. \qquad (4.38)$$

The above mentioned measures are computed based on $10,000$ runs and results for distribution functions are tabulated in Table 4.1a, 4.2a, 4.3a for population 1, 2 and 3, respectively; for quantiles in Table 4.1b, 4.2b and 4.3b, respectively. Turning to case 2, only estimation of distribution function has been considered and results are tabulated in Table 4.4 $\sim$ 4.6 for population 1 to 3 respectively. The minimum value of RB and $RRMSE$, and maximum value of RE will appear in bold font.

The simulation results in case 1 indicate that, although the estimates $\hat{\sigma}^2_\Delta$ and $\hat{\rho}_\Delta$ rather than the population values for them are used in the estimation of distribution function, the predictive estimator $\hat{F}_g$ still has better performance than the conventional ones, especially (when $\rho_Y < \rho_\Delta$),

i) when ρ_Y is close to zero, hence ρ_Δ is also close to zero, $\hat{F}_g$ attains its optimal value at $\hat{F}_{g0}$ which is also the conventional expansion estimator $\bar{F}_e$, and the alternative predictive estimator $\hat{F}_{ge}$ also performs better than the conventional estimator $\bar{F}_u$, $\bar{F}_r$.

ii) when $0 < \rho_\Delta < 1$, $\hat{F}_g$ is still better than $\bar{F}_\epsilon$, $\bar{F}_u$ and $\bar{F}_r$ in terms of biases, as measured by RB; efficiency, as measured by $RRMSE$; and relative efficiency with respect to $\bar{F}_u$, as measured by RB.

iii) when ρ_Δ is close to the unity, the results show that $\hat{F}_g$ does attain the optimum $\hat{F}_{g1}$.

For the estimation of quantiles, basically, $\hat{q}_g$ has almost all of the good characteristics of $\hat{F}_g$, except when ρ_Δ is close to the unity, for the estimation of lower (upper) tail quantiles, the bias, as measured by RB, of $\hat{q}_g$ is bigger than that of $\bar{q}_e$.

The results of distribution function estimation in case 2 (Table 4.4 ~ 4.6) show that $\hat{F}_g$ still has a better performance than the conventional estimators of distribution function under sampling scheme SRSWOR/SRSWOR when $0 < \rho_Y < 1$, but the relative efficiency (RE) decreases. So, based on the simulation results, the predictive estimators $\hat{F}_g$, $\hat{q}_g$ are suggested to be used under two stage sampling plan SRSWOR/SRSWOR, especially when a moderate sample data set is available.

Table 4.1a: Estimation of Distribution Function under Two-stage Sampling Scheme ($k = 400$, $\rho_Y = 6.46968e^{-5}$): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiency (RE) with respect to $\bar{F}_u$

F	0.250048	0.500048	0.750048
$RB(\bar{F}_e)$	**0.068441**	**0.039291**	**0.023295**
$RB(\bar{F}_u)$	0.074332	0.047439	0.035329
$RB(\bar{F}_r)$	0.068998	0.039596	0.023494
$RB(\hat{F}_g)$	0.068551	0.039335	0.023321
$RB(\hat{F}_{ge})$	0.068480	0.039310	0.023308
$RB(\hat{F}_{g1})$	0.069834	0.040066	0.023623
$RRMSE(\bar{F}_e)$	**0.085812**	**0.049308**	**0.029165**
$RRMSE(\bar{F}_u)$	0.092797	0.059348	0.044277
$RRMSE(\bar{F}_r)$	0.086313	0.049619	0.029419
$RRMSE(\hat{F}_g)$	0.085891	0.049353	0.029185
$RRMSE(\hat{F}_{ge})$	0.085821	0.049315	0.029177
$RRMSE(\hat{F}_{g1})$	0.087296	0.050144	0.029589
$RE(\bar{F}_e)$	**1.081399**	**1.203618**	**1.518155**
$RE(\bar{F}_r)$	1.075122	1.196074	1.505048
$RE(\hat{F}_g)$	1.080404	1.202521	1.517115
$RE(\hat{F}_{ge})$	1.081285	1.203447	1.517531
$RE(\hat{F}_{g1})$	1.063015	1.183551	1.496401

Table 4.1b: Estimation of Quantile under Two-stage Sampling Scheme ($k = 400$, $\rho_Y = 6.46968e^{-5}$): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiency (RE) with respect to $\bar{q}_u$

α	0.25	0.50	0.75
q	33.397644	100.530327	168.462555
$RB(\bar{q}_e)$	**0.166154**	**0.049270**	**0.032852**
$RB(\bar{q}_u)$	0.180235	0.060242	0.049784
$RB(\bar{q}_r)$	0.167222	0.049772	0.033106
$RB(\hat{q}_g)$	0.166879	0.049517	0.032910
$RB(\hat{q}_{ge})$	0.166549	0.049429	0.032915
$RB(\hat{q}_{g1})$	0.168613	0.050382	0.033175
$RRMSE(\bar{q}_e)$	**0.206488**	**0.062255**	**0.040994**
$RRMSE(\bar{q}_u)$	0.223631	0.075853	0.062570
$RRMSE(\bar{q}_r)$	0.207851	0.062800	0.041396
$RRMSE(\hat{q}_g)$	0.207310	0.062432	0.041100
$RRMSE(\hat{q}_{ge})$	0.206955	0.062408	0.041120
$RRMSE(\hat{q}_{g1})$	0.209678	0.063453	0.041590
$RE(\bar{q}_e)$	**1.083022**	**1.218424**	**1.526321**
$RE(\bar{q}_r)$	1.075920	1.207850	1.511499
$RE(\hat{q}_g)$	1.078728	1.214970	1.522384
$RE(\hat{q}_{ge})$	1.080578	1.215437	1.521644
$RE(\hat{q}_{g1})$	1.066545	1.195420	1.504448

Table 4.2a: Estimation of Distribution Function under Two-stage Sampling Scheme ($k = 400$, $\rho_Y = 0.505525$): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiency (RE) with respect to $\bar{F}_u$

F	0.250048	0.500048	0.750048
$RB(\bar{F}_e)$	0.172682	0.105269	0.057185
$RB(\bar{F}_u)$	0.175874	0.108024	0.062422
$RB(\bar{F}_r)$	0.174016	0.105159	0.056874
$RB(\hat{F}_g)$	**0.171048**	**0.104064**	0.056440
$RB(\hat{F}_{ge})$	0.172744	0.105213	0.057132
$RB(\hat{F}_{g1})$	0.171082	0.104071	**0.056426**
$RRMSE(\bar{F}_e)$	0.215411	0.131488	0.071282
$RRMSE(\bar{F}_u)$	0.219305	0.134922	0.077948
$RRMSE(\bar{F}_r)$	0.216766	0.131261	0.070910
$RRMSE(\hat{F}_g)$	**0.213544**	0.129717	**0.070404**
$RRMSE(\hat{F}_{ge})$	0.215439	0.131400	0.071209
$RRMSE(\hat{F}_{g1})$	0.213599	**0.129681**	0.070407
$RE(\bar{F}_e)$	1.018077	1.026116	1.093516
$RE(\bar{F}_r)$	1.011713	1.027891	1.099253
$RE(\hat{F}_g)$	**1.026978**	1.040126	**1.107153**
$RE(\hat{F}_{ge})$	1.017945	1.026804	1.094637
$RE(\hat{F}_{g1})$	1.026714	**1.040415**	1.107106

Table 4.2b: Estimation of Quantile under Two-stage Sampling Scheme ($k = 400$, $\rho_Y = 0.505525$): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiency (RE) with respect to $\bar{q}_u$

α	0.25	0.50	0.75
q	99.166565	100.025986	100.860977
$RB(\bar{q}_e)$	0.001710	0.001640	0.001656
$RB(\bar{q}_u)$	0.001741	0.001687	0.001806
$RB(\bar{q}_r)$	0.001722	0.001642	0.001646
$RB(\hat{q}_g)$	**0.001676**	**0.001628**	0.001638
$RB(\hat{q}_{ge})$	0.001709	0.001638	0.001653
$RB(\hat{q}_{g1})$	**0.001676**	**0.001628**	**0.001637**
$RRMSE(\bar{q}_e)$	0.002164	0.002048	0.002050
$RRMSE(\bar{q}_u)$	0.002201	0.002105	0.002249
$RRMSE(\bar{q}_r)$	0.002174	0.002049	0.002042
$RRMSE(\hat{q}_g)$	0.002120	**0.002027**	**0.002032**
$RRMSE(\hat{q}_{ge})$	0.002161	0.002046	0.002048
$RRMSE(\hat{q}_{g1})$	**0.002119**	**0.002027**	**0.002032**
$RE(\bar{q}_e)$	1.017098	1.027832	1.097073
$RE(\bar{q}_r)$	1.012420	1.027330	1.101371
$RE(\hat{q}_g)$	1.038208	**1.038481**	**1.106791**
$RE(\hat{q}_{ge})$	1.018510	1.028837	1.098145
$RE(\hat{q}_{g1})$	**1.038697**	**1.038481**	**1.106791**

Table 4.3a: Estimation of Distribution Function under Two-stage Sampling Scheme ($k = 400$, $\rho_Y = 0.999688$): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiency (RE) with respect to $\bar{F}_u$

F	0.250048	0.500048	0.750048
$RB(\bar{F}_e)$	0.293659	0.171238	0.098635
$RB(\bar{F}_u)$	0.295310	0.171568	0.101058
$RB(\bar{F}_r)$	0.293439	0.170194	0.098281
$RB(\hat{F}_g)$	**0.286248**	**0.167078**	**0.095189**
$RB(\hat{F}_{ge})$	0.293534	0.171064	0.098562
$RB(\hat{F}_{g1})$	0.286251	**0.167078**	**0.095189**
$RRMSE(\bar{F}_e)$	0.367103	0.213609	0.124315
$RRMSE(\bar{F}_u)$	0.369085	0.214062	0.126911
$RRMSE(\bar{F}_r)$	0.366856	0.212643	0.123786
$RRMSE(\hat{F}_g)$	**0.362295**	0.210097	0.121256
$RRMSE(\hat{F}_{ge})$	0.366892	0.213408	0.124199
$RRMSE(\hat{F}_{g1})$	0.362297	**0.210096**	**0.121255**
$RE(\bar{F}_e)$	1.005399	1.002121	1.020882
$RE(\bar{F}_r)$	1.006076	1.006673	1.025245
$RE(\hat{F}_g)$	**1.018742**	1.018872	1.046637
$RE(\hat{F}_{ge})$	1.005977	1.003065	1.021836
$RE(\hat{F}_{g1})$	1.018736	**1.018877**	**1.046645**

Table 4.3b: Estimation of Quantile under Two-stage Sampling Scheme (k = 400, ρ_Y = 0.999688): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiency (RE) with respect to $\bar{q}_u$

α	0.25	0.50	0.75
q	67.616135	98.518211	137.165283
$RB(\bar{q}_e)$	**0.169759**	0.104066	**0.080850**
$RB(\bar{q}_u)$	0.173338	0.101742	0.085348
$RB(\bar{q}_r)$	0.172874	0.101360	0.082424
$RB(\hat{q}_g)$	0.171031	**0.098280**	0.081960
$RB(\hat{q}_{ge})$	0.170669	0.102661	0.081919
$RB(\hat{q}_{g1})$	0.171032	0.099484	0.081503
$RRMSE(\bar{q}_e)$	**0.210285**	0.136378	**0.103265**
$RRMSE(\bar{q}_u)$	0.215490	0.134474	0.108429
$RRMSE(\bar{q}_r)$	0.214798	0.133775	0.104848
$RRMSE(\hat{q}_g)$	0.212625	**0.129826**	0.104510
$RRMSE(\hat{q}_{ge})$	0.212308	0.134689	0.104586
$RRMSE(\hat{q}_{g1})$	0.212627	0.131263	0.103892
$RE(\bar{q}_e)$	**1.024752**	0.986039	**1.050007**
$RE(\bar{q}_r)$	1.003222	1.005225	1.034154
$RE(\hat{q}_g)$	1.013474	**1.035802**	1.037499
$RE(\hat{q}_{ge})$	1.014988	0.998404	1.036745
$RE(\hat{q}_{g1})$	1.013465	1.024462	1.043670

Table 4.4: Estimation of Distribution Function under Two-stage Sampling Scheme ($k = 245$, $\rho_Y = 6.46968e^{-5}$): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiency (RE) with respect to $\bar{F}_u$

F	0.250048	0.500048	0.750048
$RB(\bar{F}_e)$	0.088167	0.050435	0.029300
$RB(\bar{F}_u)$	0.092417	0.056881	0.039350
$RB(\bar{F}_r)$	0.088424	0.050633	0.029415
$RB(\hat{F}_g)$	0.088235	0.050473	0.029316
$RB(\hat{F}_{ge})$	**0.088155**	**0.050430**	**0.029298**
$RB(\hat{F}_{g1})$	0.089240	0.051104	0.029650
$RRMSE(\bar{F}_e)$	0.110900	0.063187	**0.036686**
$RRMSE(\bar{F}_u)$	0.116210	0.071352	0.049415
$RRMSE(\bar{F}_r)$	0.111221	0.063432	0.036899
$RRMSE(\hat{F}_g)$	0.110963	0.063239	0.036703
$RRMSE(\hat{F}_{ge})$	**0.110880**	**0.063182**	0.036690
$RRMSE(\hat{F}_{g1})$	0.112131	0.064061	0.037101
$RE(\bar{F}_e)$	1.047881	1.129220	**1.346972**
$RE(\bar{F}_r)$	1.044857	1.124858	1.339196
$RE(\hat{F}_g)$	1.047286	1.128291	1.346348
$RE(\hat{F}_{ge})$	**1.048070**	**1.129309**	1.346825
$RE(\hat{F}_{g1})$	1.036377	1.113813	1.331905

Table 4.5: Estimation of Distribution Function under Two-stage Sampling Scheme (k = 245, ρ_Y = 0.505525): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiences (RE) with respect to $\bar{F}_u$

F	0.250048	0.500048	0.750048
$RB(\bar{F}_e)$	0.178606	0.107402	0.058506
$RB(\bar{F}_u)$	0.182375	0.110442	0.063959
$RB(\bar{F}_r)$	0.180690	0.107715	0.058560
$RB(\hat{F}_g)$	**0.178118**	**0.106991**	**0.058283**
$RB(\hat{F}_{ge})$	0.178720	0.107369	0.058479
$RB(\hat{F}_{g1})$	0.178286	0.107055	0.058336
$RRMSE(\bar{F}_e)$	0.223048	0.134246	0.073079
$RRMSE(\bar{F}_u)$	0.227667	0.138483	0.080056
$RRMSE(\bar{F}_r)$	0.225493	0.134839	0.073194
$RRMSE(\hat{F}_g)$	**0.222609**	**0.133681**	**0.072808**
$RRMSE(\hat{F}_{ge})$	0.223182	0.134238	0.073054
$RRMSE(\hat{F}_{g1})$	0.222833	0.133757	0.072870
$RE(\bar{F}_e)$	1.020709	1.031561	1.095472
$RE(\bar{F}_r)$	1.009641	1.027025	1.093751
$RE(\hat{F}_g)$	**1.022721**	**1.035921**	**1.099550**
$RE(\hat{F}_{ge})$	1.020096	1.031623	1.095847
$RE(\hat{F}_{g1})$	1.021693	1.035333	1.098614

Table 4.6: Estimation of Distribution Function under Two-stage Sampling Scheme ($k = 245$, $\rho_Y = 0.999688$): Relative Biases (RB), Relative Root Mean Square Errors (RRMSE) and Relative Efficiency (RE) with respect to $\bar{F}_u$

F	0.250048	0.500048	0.750048
$RB(\bar{F}_e)$	0.293583	0.168522	0.098635
$RB(\bar{F}_u)$	0.296829	0.169886	0.101908
$RB(\bar{F}_r)$	0.295322	0.168556	0.099076
$RB(\hat{F}_g)$	**0.287794**	0.165206	0.095962
$RB(\hat{F}_{ge})$	0.293490	0.168420	0.098590
$RB(\hat{F}_{g1})$	0.287799	**0.165205**	**0.095961**
$RRMSE(\bar{F}_e)$	0.368370	0.210824	0.123406
$RRMSE(\bar{F}_u)$	0.372042	0.212882	0.127397
$RRMSE(\bar{F}_r)$	0.370401	0.211188	0.123952
$RRMSE(\hat{F}_g)$	**0.365755**	**0.208841**	0.121674
$RRMSE(\hat{F}_{ge})$	0.368392	0.210752	0.123397
$RRMSE(\hat{F}_{g1})$	0.365758	0.208842	**0.121673**
$RE(\bar{F}_e)$	1.009968	1.009762	1.032340
$RE(\bar{F}_r)$	1.004430	1.008021	1.027793
$RE(\hat{F}_g)$	**1.017189**	**1.019350**	1.047036
$RE(\hat{F}_{ge})$	1.009908	1.010107	1.032416
$RE(\hat{F}_{g1})$	1.017181	1.019345	**1.047044**

REFERENCES

Chambers, R. L. and Dunstan, R. (1986). *Estimating distribution functions from survey data.* Biometrika. $73, 497 - 604$.

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993). *Bias robust estimation in finite populations using nonparametric calibration.* Journal of the American Statistical Association. 88, $268 - 277$.

Chen, J. and Qin, J. (1993). *Empirical likelihood estimation for finite populations and the effective usage of auxiliary in auxiliary information.* Biometrika. 80, $107 - 116$.

Chen, J. and Sitter, R. R. (1996). *A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys.* Simon Fraser University, Burnaby.

Chen, S. X. (1997). *Empirical likelihood-based kernel density estimation.* Australian Journal of Statistics. 39, $47 - 56$.

Cheng, K .F., Lin, P. E. (1981). *Nonparametric estimation of a regression function: limiting distribution.* Australian Journal of Statistics. 23, $186 - 195$.

Cochran, W. G. (1977). *Sampling techniques (Third Edition).* New York: John Wiley & Sons.

Dielman, D., Lowry, C., and Pfaffenberger, R. (1994). *A comparison of quantile estimators.* Communications in Statistics. Part B–Simulation and Computation. $23, 355 - 371$.

Fan, J. Q. (1992). *Design-adaptive nonparametric regression.* Journal of the American Statistical Association. 87, 998 – 1004.

Fransisco, C. A. and Fuller, W. A. (1991). *Quantile estimation with a complex survey design.* The Annals of Statistics. 19, 454 – 469.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. Journal of the Royal Statistical Society. $B17, 269 - 278$.

Godambe, V. P. (1989). *Estimation of cumulative distribution of a survey population.* Technical Report STAT-89-17, University of Waterloo.

Gross, S. T. (1980). *Median estimation in sample surveys.* Proc Survey Res. Meth. Sect., Am. Statist. Assoc. *181-184.*

Hansen, M. H., Madow, W. G. and Tepping, B. J. (1983). *An evaluation of model-dependent and probability-sampling inferences in sample surveys.* Journal of the American Statistical Association. *78,776-793.*

Kish, L. (1965). *Survey sampling.* Wiley, New York.

Krieger, A. M. and Pfeffermann, P. (1992). *Maximum likelihood estimation from complex surveys.* Unpublished Technical Report, Department of Statistics, University of Pennsylvania.

Kuk, A. Y. C. (1988). *Estimation of distribution functions and medians under sampling with unequal probabilities.* Biometrika. 75, 97 – 103.

Kuk, A. Y. C. and Mak, T. K. (1989). *Median estimation in the presence of auxiliary information.* Journal of the Royal Statistical Society. $B51, 261 - 269$.

Loynes, R. M. (1966). *Some aspects of the estimation of quantiles.* Journal of the Royal Statistical Society. *B28, 497-512.*

McCarthy, P. G. (1965). *Stratified sampling and distribution-free confidence intervals for a median.* Journal of the American Statistical Association. $60, 772 - 783$.

Olsson, J. and Rootzén, H. (1996). *Quantile estimation from repeated measurements.* Journal of the American Statistical Association. 91, $1560 - 1565$.

Owen, A. B. (1988). *Empirical likelihood ratio confidence intervals for a single functional.* Biometrika. *75,237-249.*

Owen, A. B. (1990). *Empirical likelihood confidence regions.* The Annals of Statistics. *18,90-120.*

Press, W. H., Flannery, B. P., Teukolsky Saul A. and Vetterling W. T. (1990). *Numerical Recipes, The art of Scientific Computing (Fortran Version).* Cambridge University Press.

Qin J. and Lawless J. (1994). *Empirical likelihood and general estimating equations.* The Annals of Statistics. *22:300-325.*

Royall, R. M. (1976). *The linear least-squares prediction approach to two stage sampling.* Journal of the American Statistical Association. 71, $657 - 670$.

Rao, J. N. K. (1975). *Unbiased variance estimation for multistage designs.* Sankhya. *C37, 133-139.*

Rao, J. N. K., Kovar, J. G. and Mantel, H. J. (1990). *On estimating distribution functions and quantiles from survey data using auxiliary information.* Biometrika. $77, 365 - 375$.

Rao, J. N. K. and Liu, J. (1992). *On estimating distribution functions from sample survey data using supplementary information at the estimation stage.* In Saleh, A. K. M. E. (ed.) Nonparametric statistics and related topics. Elsevier Science Publishers, Amsterdam. $pp\ 399 - 407$.

Rao, J. N. K. (1994). *Estimating totals and distribution functions using auxiliary information at the estimation stage.* Journal of Official Statistics. $10, 153 - 165$.

Scott, A. and Smith, T. M. F. (1969). *Estimation in multi-stage surveys.* Journal of the American Statistical Association. $64, 830 - 840$.

Sedransk, J. and Meyer, J. (1978). *Confidence intervals for the quantiles of a finite population: simple random stratified random sampling.* Journal of the Royal Statistical Society. $B40, 239 - 252$.

Sedransk, J. and Smith, P. (1983). *Lower bounds for confidence coefficients for confidence intervals for finite population quantiles.* Comm. Statist. Theory Meth. $12, 1329 - 1344$.

Thompson, M. E. (1997). *Theory of sample surveys.* St Edmundsbury Press, Bury St Edmunds, Suffolk.

Zhang, B. (1997). *Estimating a distribution function in the presence of auxiliary information*. Metrika. 46, 245 − 251.