CHAPTER 23

Estimation of Distribution Functions and Quantiles

23.1 INTRODUCTION

In earlier sections, we have considered various methods of estimating the finite population totals, means, and ratios of two variables. Most of these methods, when extended to the estimation of distribution function, yield unsatisfactory results because these estimators may not satisfy the basic properties of the distribution functions. Estimation of distribution functions from the survey data is often a part of the objectives of a survey. In particular, it may be of interest to planners who would want to know the proportion of people living below the poverty line, unemployed, or money spent on education. Furthermore, it is well known that the median is considered to be a more appropriate measure of location than the mean of a skewed distribution such as income. It is also important to estimate the income inequality through the Gini coefficient, which is a function of a distribution function. In this section, different methods of estimation of distribution function have been considered. Estimation of quantiles and medians will also be obtained from the inversion of the estimates of distribution functions.

23.2 ESTIMATION OF DISTRIBUTION FUNCTIONS

The distribution function of a random variable X is defined as $F(x) = P(X \leq x)$. The distribution function $F(x)$ has the following properties:

(i) $F(x) \geq 0$, (ii) nondecreasing, (iii) $F(-\infty) = 0$, (iv) $F(+\infty) = 1$, and (v) $F(x)$is right continuous

(23.2.1)

For a finite population of N units, we define the distribution function $F(t)$ of the study variable y, which is the proportion of the units of the population U whose values are less than or equal to t, i.e.,

$$F(t) = \frac{1}{N} \sum_{i=1}^{N} I(y_i \leq t) \tag{23.2.2}$$

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00023-6

where $I(y_i \le t) = \begin{cases} 1 & \text{if } y_i \le t \\ 0 & \text{if } y_i > t \end{cases}$ and y_i is the value of the study variable y for the ith unit of the population; $i = 1,\ldots, N$.

23.2.1 Design-Based Estimation

Let a sample s of size n be selected from a population U using a sampling design p with $\pi_i(>0)$ and $\pi_{ij}(>0)$ as the inclusion probabilities for the ith, and ith and jth unit ($i \ne j$), respectively. In case the population size N is known, an unbiased estimator of $F(t)$ is given by

$$\widehat{F}(t) = \frac{1}{N}\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i} \tag{23.2.3}$$

For simple random sampling without replacement (SRSWOR) sampling $\pi_i = n/N$ and $\widehat{F}(t)$ reduces to the sample empirical distribution function

$$s(t) = \frac{r(t)}{n} \tag{23.2.4}$$

where $r(t)$ is number of y_i's in the sample less than or equal to t.

The estimator $s(t)$ is admissible and also minimax under certain loss function (Dorfman, 2009). KuK (1988) proposed following alternative estimators based on complementary proportion

$$\widehat{F}_c(t) = 1 - \frac{1}{N}\sum_{i\in s}\frac{1 - I(y_i \le t)}{\pi_i} \tag{23.2.5}$$

and

$$\widehat{F}_w(t) = w\widehat{F}(t) + (1-w)\widehat{F}_c(t) \tag{23.2.6}$$

where w is a suitably chosen weight.

It should be noted that none of the estimators $\widehat{F}(t)$, $\widehat{F}_c(t)$, and $\widehat{F}_w(t)$ mentioned above possess all the properties of the distribution function stated in Eq. (23.2.1). In fact $\widehat{F}(t)$ may exceed unity and $\widehat{F}(\infty) = \frac{1}{N}\sum_{i\in s}\frac{1}{\pi_i}$ may not equal to 1. Details on discussions of the performances of the estimators have been given by Dorfman (2009). To overcome this difficulty, the following alternative Hájek type estimator has been proposed.

$$\widehat{F}_h(t) = \sum_{i\in s}\frac{I(y_i \le t)}{\pi_i}\Big/\sum_{i\in s}\frac{1}{\pi_i} \tag{23.2.7}$$

Though the estimator (23.2.7) is not design unbiased for $F(t)$, it is design consistent and approximately unbiased. It satisfies all the properties of the distribution function stated in Eq. (23.2.1).

23.2.2 Design-Based Estimators Using Auxiliary Information

Rao et al. (1990) considered design-based ratio and difference estimators of the population distribution function $F(t)$ when the values of the auxiliary variable x_i's are known for $i = 1,\ldots, N$. The conventional estimator for the population ratio $R = Y/X$ is

$$\widehat{R} = \left(\sum_{i\in s}\frac{y_i}{\pi_i}\right)\Big/\left(\sum_{i\in s}\frac{x_i}{\pi_i}\right) \tag{23.2.8}$$

where $X = \sum_{i=1}^{N} x_i$ and $Y = \sum_{i=1}^{N} y_i$.

Treating $\widehat{y}_i = \widehat{R}x_i$ as an estimator of y_i, Rao et al. (1990) proposed the following ratio estimator of $F(t)$ as

$$\widehat{F}_R(t) = \frac{1}{N}\widehat{\lambda} \times \left(\sum_{i\in U} I\left(\widehat{y}_i \le t\right)\right) \tag{23.2.9}$$

where

$$\widehat{y}_i = \widehat{R}x_i \quad \text{and} \quad \widehat{\lambda} = \left(\frac{\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i}}{\sum_{i\in s}\frac{I\left(\widehat{y}_i \le t\right)}{\pi_i}}\right) \tag{23.2.10}$$

The proposed ratio estimator $\widehat{F}_R(t)$ is design consistent and exactly equal to the population distribution function $F(t)$ if y_i is proportional to x_i for every $i \in U$. The proposed estimator is expected to gain in efficiency compared to the conventional estimator $\widehat{F}(t)$ if y_i is approximately proportional to x_i. But the estimator $\widehat{F}_R(t)$ suffers from a drawback that it may not possess the desirable properties of the distribution function given in Eq. (23.2.1): it may have a value outside the interval [0, 1]. For a large sample size n, an approximate expression of variance of $\widehat{F}_R(t)$ is given by Rao et al. (1990) as

$$\begin{aligned} Var\left[\widehat{F}_R(t)\right] &\cong V\left(\frac{1}{N}\sum_{i\in s}\frac{Q_i(t)}{\pi_i}\right) \\ &= \frac{1}{2N^2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{Q_i}{\pi_i} - \frac{Q_j}{\pi_j}\right)^2 \end{aligned} \tag{23.2.11}$$

where

$$Q_i = I(y_i \le t) - \lambda I(Rx_i \le t) \text{ and } \lambda = \frac{\sum_{i\in U} I(y_i \le t)}{\sum_{i\in U} I(Rx_i \le t)}. \tag{23.2.12}$$

An approximate unbiased estimator of $V_p\left[\widehat{F}_R(t)\right]$ is

$$\widehat{V}\left[\widehat{F}_R(t)\right] = \frac{1}{2N^2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{\widehat{Q}_i}{\pi_i} - \frac{\widehat{Q}_j}{\pi_j}\right)^2 \tag{23.2.13}$$

where $\widehat{Q}_i = I(y_i \le t) - \widehat{\lambda} I\left(\widehat{y}_i \le t\right)$ and $\widehat{\lambda}$ is given in Eq. (23.2.10).

Rao et al. (1990) also proposed a difference estimator of $F(t)$ as

$$\widehat{F}_D(t) = \frac{1}{N}\left\{\left(\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i}\right) - \left(\sum_{i\in s}\frac{I\left(\widehat{y}_i \le t\right)}{\pi_i} - \sum_{i\in U} I\left(\widehat{y}_i \le t\right)\right)\right\} \tag{23.2.14}$$

An approximate expression of the variance of $\widehat{F}_D(t)$ and its unbiased estimators are respectively given by

$$\begin{aligned} Var\left[\widehat{F}_D(t)\right] &\cong V_p\left(\frac{1}{N}\sum_{i\in s}\frac{G_i(t)}{\pi_i}\right) \\ &= \frac{1}{2N^2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{G_i(t)}{\pi_i} - \frac{G_j(t)}{\pi_j}\right)^2 \end{aligned} \tag{23.2.15}$$

and

$$\widehat{V}\left[\widehat{F}_D(t)\right] = \frac{1}{2N^2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{\widehat{G}_i(t)}{\pi_i} - \frac{\widehat{G}_j(t)}{\pi_j}\right)^2 \tag{23.2.16}$$

where $G_i(t) = I(y_i \le t) - I(Rx_i \le t)$ and $\widehat{G}_i(t) = I(y_i \le t) - I\left(\widehat{R}x_i \le t\right)$.

23.2.3 Model-Based Estimators

Let us suppose that the study variable y is related to the auxiliary variable through the superpopulation model

$$y_i = \mu(x_i) + \sqrt{v(x_i)}\,\epsilon_i \tag{23.2.17}$$

where $\mu(x_i)$ is a function of x_i but involves unknown model parameters, $v(x_i)$ is known and positive, and $\in_i$'s are independently and identically distributed random variables with $E_m(\in_i) = 0$ and $V_m(\in_i) = \sigma^2(>0)$. Here E_m and V_m denote expectation and variance operators, respectively, with respect to the model (23.2.17). Chambers and Dunstan (1986) and Rao et al. (1990) used the model (23.2.17) with $\mu(x_i) = \beta x_i$, where β is an unknown parameter.

Let $\widehat{\mu}(x_i)$ be a model-unbiased estimator of $\mu(x_i)$, i.e., $E_m[\widehat{\mu}(x_i)] = \mu(x_i)$. Then $\widehat{r}_i(t) = \dfrac{t - \hat{\mu}(x_i)}{\sqrt{v(x_i)}}$ becomes model-unbiased estimator of $r_i(t) = \dfrac{t - \mu(x_i)}{\sqrt{v(x_i)}}$.

Let $A_i(t)$ be the model expectation of $I(y_i \le t)$ in the sense that

$$A_i(t) = E_m\{I(y_i \le t)\} = P(y_i \le t) = P(\in_i \le r_i(t)) \tag{23.2.18}$$

An approximate model-unbiased estimator of $A_i(t)$ is given by

$$\widehat{A}_i(t) = \frac{1}{n} \sum_{j \in s} I\left\{\widehat{r}_j(y_j) \le \widehat{r}_i(t)\right\} \tag{23.2.19}$$

where $\widehat{r}_j(y_j) = \dfrac{y_j - \hat{\mu}(x_j)}{\sqrt{v(x_j)}}$.

Under the model (23.2.17), Johnson (2003) proposed the following model-based estimator of $F(t)$ as

$$\widehat{F}_J(t) = \frac{1}{N}\left(\sum_{i \in s} I(y_i \le t) + \sum_{i \in U-s} \widehat{A}_i(t)\right) \tag{23.2.20}$$

where $U - s$ denotes the set of nonsampled units.

Chambers and Dunstan (1986) considered the estimator (23.2.20) when $\mu(x_i) = \beta x_i$ and the model parameter β was estimated by the weighted least squares method as

$$\widehat{\beta} = \left(\sum_{i \in s} \frac{y_i x_i}{v(x_i)}\right) \Big/ \left(\sum_{i \in s} \frac{x_i^2}{v(x_i)}\right) \tag{23.2.21}$$

Substituting $\widehat{\mu}(x_i) = \widehat{\beta} x_i$ in the expression (23.2.20), Chambers and Dunstan (1986) estimator is obtained as

$$\widehat{F}_{cd}(t) = \frac{1}{N}\left[\sum_{i \in s} I(y_i \le t) + \sum_{i \in U-s}\left\{\frac{1}{n}\sum_{j \in s} I\left(\frac{y_j - \widehat{\beta} x_j}{\sqrt{v(x_j)}} \le \frac{t - \widehat{\beta} x_i}{\sqrt{v(x_i)}}\right)\right\}\right] \tag{23.2.22}$$

The estimator $\widehat{F}_{cd}$ satisfies all the properties of a distribution function. Although the estimator $\widehat{F}_{cd}(t)$ is independent of the sampling design, it is asymptotically model unbiased and has the property that $\widehat{F}_{cd}(t)$ equals to $F(t)$ when $y_i \propto x_i \ \forall i \in U$. Since the estimator $\widehat{F}_{cd}$ depends on the assumed model, it is highly efficient if the model (23.2.17) holds but if the model is incorrect, the estimator becomes biased and can perform much worse even than the naive estimator $s(t)$.

23.2.4 Model-Assisted Estimators

Rao et al. (1990) proposed the following model assisted difference estimators:

$$\widehat{F}^*(t) = \frac{1}{N}\left[\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i} - \left(\sum_{i\in s}\frac{B_i(t)}{\pi_i} - \sum_{i\in U} B_i(t)\right)\right] \tag{23.2.23}$$

with

$$B_i(t) = \frac{1}{N}\sum_{j\in U} I\left(\frac{y_j - \beta x_j}{\sqrt{\nu(x_j)}} \le \frac{t - \beta x_i}{\sqrt{\nu(x_i)}}\right) \tag{23.2.24}$$

The estimator $\widehat{F}^*(t)$ does not possess all the properties of distribution function. It is both design unbiased and asymptotically model unbiased for $F(t)$. It is calibrated in the sense that $\widehat{F}^*(t)$ reduces $F(t)$ if y_i is exactly proportional to x_i for $\forall i \in U$. The main demerit of this estimator is that it cannot be used in practice as $B_i(t)$'s are unknown. To overcome this difficulty, Rao et al. (1990) proposed the following alternative estimator:

$$\widehat{F}_{rkm}(t) = \frac{1}{N}\left[\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i} - \left(\sum_{i\in s}\frac{\widehat{B}_{ic}(t)}{\pi_i} - \sum_{i\in U}\widehat{B}_i(t)\right)\right] \tag{23.2.25}$$

where

$$\begin{aligned}
\widehat{B}_i(t) &= \left(\sum_{j\in s}\frac{1}{\pi_j}\right)^{-1}\left[\sum_{j\in s}\frac{1}{\pi_j} I\left(\frac{y_j - \widehat{\beta}x_j}{\sqrt{\nu(x_j)}} \le \frac{t - \widehat{\beta}x_i}{\sqrt{\nu(x_i)}}\right)\right] \text{ and} \\
\widehat{B}_{ic}(t) &= \left(\sum_{j\in s}\frac{\pi_i}{\pi_{ij}}\right)^{-1}\left[\sum_{j\in s}\frac{\pi_i}{\pi_{ij}} I\left(\frac{y_j - \widehat{\beta}x_j}{\sqrt{\nu(x_j)}} \le \frac{t - \widehat{\beta}x_i}{\sqrt{\nu(x_i)}}\right)\right]
\end{aligned} \tag{23.2.26}$$

The estimator $\widehat{B}_i(t)$ is asymptotically design unbiased for $B_i(t)$ whereas $\widehat{B}_{ic}(t)$ is asymptotically and conditionally design unbiased for $B_i(t)$ given

$i \in s$. The estimator $\widehat{F}_{rkm}(t)$ is more complex than $\widehat{F}^*(t)$ because it involves second-order inclusion probabilities otherwise it shares the same properties of the estimator $\widehat{F}^*(t)$. Under model misspecification, $\widehat{F}_{rkm}(t)$ performs better than $\widehat{F}_{cd}$.

The asymptotic design variance of $\widehat{F}_{rkm}(t)$ and an estimator of the variance was provided by Rao et al. (1990) as follows:

$$Var\left\{\widehat{F}_{rkm}(t)\right\} = \frac{1}{N^2}\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij}) \left(\frac{I(y_i \le t) - B_i(t)}{\pi_i} - \frac{I(y_j \le t) - B_j(t)}{\pi_j}\right)^2 \tag{23.2.27}$$

$$\widehat{V}\left\{\widehat{F}_{rkm}(t)\right\} = \frac{1}{N^2}\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{I(y_i \le t) - \widehat{B}^*_{ic}(j)}{\pi_i} - \frac{I(y_j \le t) - \widehat{B}^*_{jc}(i)}{\pi_j}\right)^2 \tag{23.2.28}$$

where $\widehat{B}^*_{ic}(j) = \sum_{k\in s}\frac{\pi_{ij}}{\pi_{ijk}} I\left(\frac{y_k - \hat{\beta}x_k}{\sqrt{v(x_k)}} \le \frac{t - \hat{\beta}x_i}{\sqrt{v(x_i)}}\right) \Big/ \left(\sum_{k\in s}\frac{\pi_{ij}}{\pi_{ijk}}\right)$ and $\pi_{ijk}(i \neq j \neq k)$ is the inclusion probability for the ith, jth, and kth unit in the sample.

Dorfman (2009) proposed a modified estimator of $\widehat{F}_{rkm}(t)$ as follows:

$$\begin{aligned}\widehat{F}_{dm}(t) &= \frac{1}{N}\left[\sum_{i\in s} I(y_i \le t) + \sum_{i\in U-s}\left\{\frac{1}{n}\sum_{j\in s} I\left(\frac{y_i - \widehat{\beta}x_j}{\sqrt{v(x_j)}} \le \frac{t - \widehat{\beta}x_i}{\sqrt{v(x_i)}}\right)\right\} + \sum_{i\in s}\left(\frac{1}{\pi_i} - 1\right)R_i\right] \\ &= \widehat{F}_{cd}(t) + \frac{1}{N}\sum_{i\in s}\left(\frac{1}{\pi_i} - 1\right)R_i\end{aligned} \tag{23.2.29}$$

where $R_i = I(y_i \le t) - \frac{1}{n}\sum_{j\in s} I\left(\frac{y_j - \hat{\beta}x_j}{\sqrt{v(x_j)}} \le \frac{t - \hat{\beta}x_i}{\sqrt{v(x_i)}}\right)$.

Both the estimator $\widehat{F}_{dm}(t)$ and $\widehat{F}_{rkm}(t)$ are equally efficient (Dorfman, 2009). $\widehat{F}_{dm}(t)$ has additional advantage as it does not require computation of second-order inclusion probabilities. Wang and Dorfman (1996) proposed an alternative estimator $\widehat{F}_{wdr}(t) = w\widehat{F}_{dm}(t) + (1-w)\widehat{F}_{rkm}(t)$, which is the weighted average of $\widehat{F}_{dm}(t)$ and $\widehat{F}_{rkm}(t)$. Simulation studies showed that the estimator $\widehat{F}_{wdr}(t)$ performs better than $\widehat{F}_{dm}(t)$ and $\widehat{F}_{rkm}(t)$. Mak and Kuk (1993) proposed a modification of $\widehat{F}_{cd}(t)$ estimator as follows:

$$\widehat{F}_{mk}(t) = \frac{1}{N}\left[\sum_{i\in s} I(y_i \le t) + \sum_{i\in U-s}\left\{\frac{1}{n}\sum_{j\in s}\Phi\left(\frac{t-\widehat{\beta}x_i}{\widehat{\sigma}\sqrt{v(x_i)}}\right)\right\}\right]$$

where $\Phi\left(\frac{t-\hat{\beta}x_i}{\hat{\sigma}\sqrt{v(x_i)}}\right)$ is the standard normal distribution function and $\hat{\sigma}^2$ is the weighted list square estimate of σ^2 obtained from the model (23.2.17). The main advantage of using $\widehat{F}_{mk}(t)$ over $\widehat{F}_{cd}(t)$ is ease of computation but their relative performances are not known.

23.2.5 Nonparametric Regression Method

For the model-based and model-assisted estimators, we used parametric regression where a single smooth function $\mu(x)$ was estimated over the entire possible range of the auxiliary variable x. In a nonparametric regression we approximate $\mu(x)$ locally by placing more weight on y_i's corresponding to x_i, which are close to x. In particular, suppose we wish to approximate $\mu(x)$ by a polynomial of degrees p viz.

$$\mu_{lp}(x) = \beta_0 + \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p \tag{23.2.30}$$

then the regression coefficients β_0, β_1,…, β_p may be estimated by using weighted least square method using kernel weight

$$K\left(\frac{x_i - x}{b}\right)$$

attached to the ith observation. The kernel function $K(u)$ is a symmetric positive function of u, which decreases as $|u|$ increases. The parameter $b(>0)$ is the smoothing parameter, which is referred to bandwidth. Thus the normal equations are obtained from the selected sample by minimizing

$$\sum_{i\in s}\left(y_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_p(x_i - x)\right)^2\left\{\frac{1}{\pi_i}K\left(\frac{x_i - x}{b}\right)\right\}$$

with respect to $\beta_0, \beta_1, \ldots, \beta_p$. Here we assume, without loss of generality, that the selected sample s contains first n units of the population. The estimates of $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ are given by

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}_x' \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x' \mathbf{W}_x \mathbf{y}_s \tag{23.2.31}$$

where

$$\mathbf{X}_x = \begin{pmatrix} 1 & x_1 - x & \ldots & (x_1 - x)^p \\ . & . & . & . \\ . & . & . & . \\ 1 & x_n - x & . & (x_n - x)^p \end{pmatrix},$$

$$\mathbf{W}_x = diag\left(\frac{1}{\pi_1} K\left(\frac{x_1 - x}{b}\right) \cdots \frac{1}{\pi_n} K\left(\frac{x_n - x}{b}\right)\right) \text{and } \mathbf{y}_s' = (y_1, \ldots, y_n).$$

The predicted value of y at the point $x = x_i$ is the intercept. So we write

$$\widehat{y}_i = \widehat{\mu}_{lp}(x_i) = \widehat{\beta}_0 = \mathbf{e}_1' \left(\mathbf{X}_{x_i}' \mathbf{W}_{x_i} \mathbf{X}_{x_i}\right)^{-1} \mathbf{X}_{x_i}' \mathbf{W}_{x_i} \mathbf{y} \tag{23.2.32}$$

where $\mathbf{e}_1$ is a $(p+1)$ column vector with 1 as the first element and 0 elsewhere.

23.2.5.1 Nandaraya–Watson Estimator

For $p = 0$, Eq. (23.2.32) reduces to Nandaraya–Watson (1964) estimator:

$$\widehat{\mu}_{lp}(x_i) = \frac{\sum_{i \in s} K\left(\frac{x_i - x}{b}\right) \frac{y_i}{\pi_i}}{\sum_{i \in s} K\left(\frac{x_i - x}{b}\right) \frac{1}{\pi_i}} \tag{23.2.33}$$

The data analyst must choose p and b suitably. For further details, readers are referred to Ruppert et al. (2003).

23.2.5.2 Breidt and Opsomer Estimator

The local polynomial estimator for the population total is given by

$$\widehat{Y}_{bo} = \sum_{i \in s} \frac{y_i - \widehat{\mu}_{lp}(x_i)}{\pi_i} + \sum_{i \in U} \widehat{\mu}_{lp}(x_i) \tag{23.2.34}$$

Now replacing y_i with $I(y_i \leq t)$ in Eq. (23.2.34), we get the Breidt and Opsomer (2000) estimator of the distribution function as

$$\widehat{F}_{lp}(t) = \frac{1}{N} \sum_{i \in s} \frac{I(y_i \leq t) - \widehat{\mu}^*_{lp}(x_i, t)}{\pi_i} + \frac{1}{N} \sum_{i \in U} \widehat{\mu}^*_{lp}(x_i, t) \tag{23.2.35}$$

where $\widehat{\mu}^{*}_{lp}(x_i,t) = \mathbf{e}_1'\left(\mathbf{X}_{x_i}'\mathbf{W}_{x_i}\mathbf{X}_{x_i}\right)^{-1}\mathbf{X}_{x_i}'\mathbf{W}_{x_i}\mathbf{I}_{st}$ and $\mathbf{I}_{st} = (I(y_1 \le t), \ldots, I(y_i \le t), \ldots, I(y_n \le t))'$.

From Breidt and Opsomer (2000), we see that $\widehat{F}_{lp}(t)$ maintains design consistency and asymptotic design-unbiased properties. The estimated variance of $\widehat{F}_{lp}(t)$ was given by Johnson (2003) as

$$\widehat{V}\left(\widehat{F}_{lp}(t)\right) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\left(I(y_i \le t) - \widehat{\mu}^{*}_{lp}(x_i,t)\right)\left(I(y_j \le t) - \widehat{\mu}^{*}_{lp}(x_j,t)\right) \tag{23.2.36}$$

23.2.5.3 Kuo Estimator

Kuo (1988) proposed the following estimator for the distribution function $F(t)$:

$$\widehat{F}_{ko}(t) = \frac{1}{N}\left[\sum_{i\in s} I(y_i \le t) + \sum_{j\in U-s}\sum_{i\in s} w_{ij} I(y_i \le t)\right] \tag{23.2.37}$$

where $w_{ij} = \dfrac{K\left(\dfrac{x_j - x_i}{b}\right)}{\sum\limits_{i\in s} K\left(\dfrac{x_j - x_i}{b}\right)}$, $K(z) = e^{-z^2/2}$ is the standard normal density (kernel).

Dorfman and Hall (1993) provided the expressions of asymptotic bias and variance of $\widehat{F}_{ko}(t)$.

23.2.5.4 Kuk Estimator

Kuk (1993) nonparametric regression estimator for $F(t)$ is given by

$$\widehat{F}_{kk}(t) = \frac{1}{N}\sum_{j\in U}\widehat{R}_j(t) \tag{23.2.38}$$

where

$$\widehat{R}_j(t) = \frac{\sum\limits_{i\in s}\dfrac{1}{\pi_i}w\left(\dfrac{x_j - x_i}{b}\right)W\left(\dfrac{t-y_i}{b}\right)}{\sum\limits_{i\in s}\dfrac{1}{\pi_i}w\left(\dfrac{x_j - x_i}{b}\right)}$$, $W(z) = e^z/(1+e^z)$ is the standard logistic distribution function with density $w(z) = e^z/(1+e^z)^2$, and b is the band with parameter used to control the amount of smoothing.

It should be noted that both the estimators $\widehat{F}_{ko}(t)$ and $\widehat{F}_{kk}(t)$ meet the properties of the distribution function given in Eq. (23.2.1). For more

details, readers are referred to Kuo (1988), Kuk (1993), Dorfman (2009), among others. Dorfman and Hall (1993) considered a design-adjusted version of Kuo (1988) and Rao–Kovar–Mantel (1990) estimators while Silva and Skinner (1995) suggested a poststratified estimator of $F(t)$.

23.2.6 Calibration Method

In Deville and Särndal's (1992) calibration method, the Horvitz–Thompson estimator $\widehat{\overline{Y}}_{ht} = \frac{1}{N}\sum_{i\in s} d_i y_i$ with $d_i = 1/\pi_i$ was calibrated as

$$\widehat{\overline{Y}}_c = \sum_{i\in s} w_i y_i$$

The weights w_i's were chosen to minimize the distance

$$\Phi = \sum_{i\in s} \frac{(w_i - d_i)^2}{d_i q_i} \tag{23.2.39}$$

subject to the calibrating constraints $\sum_{i\in s} w_i x_i = X$. Here x_i's are the values of the auxiliary variable x with known total X and q_i are suitably chosen weights. Details have been given in Section 9.7. Suppose that the study variable y is related to the auxiliary variable through the following superpopulation model

$$y_i = \mu(x_i, \theta) + \sqrt{v(x_i)}\,\epsilon_i$$

where θ is an unknown model parameter and ϵ_i's are independently identically distributed with mean zero and variance σ^2.

Let $\widehat{\theta}$ be a suitable estimator of θ obtained from the selected sample using some standard procedure so that $\widehat{\mu}_i = \mu\left(x_i, \widehat{\theta}\right)$ is an estimator for $\mu_i = \mu(x_i, \theta)$. Minimizing Eq. (23.2.39) subject to the calibrating constraints (i) $\sum_{i\in s} w_i = N$ and (ii) $\sum_{i\in s} w_i \widehat{\mu}_i = \sum_{i\in U} \widehat{\mu}_i$, Wu and Sitter (2001) derived the calibrated estimator for the population total Y as

$$\widehat{\overline{Y}}_c = \widehat{\overline{Y}}_{ht} - \widehat{B}\left(\frac{1}{N}\sum_{i\in s}\frac{\widehat{\mu}_i}{\pi_i} - \frac{1}{N}\sum_{i\in U}\widehat{\mu}_i\right) \tag{23.2.40}$$

where $$\widehat{B} = \left(\sum_{i\in s} d_i q_i (\widehat{\mu}_i - \overline{\mu})(y_i - \overline{y})\right)\left(\sum_{i\in s} d_i q_i (\widehat{\mu}_i - \overline{\mu})^2\right)^{-1},$$

$\overline{y} = \sum_{i\in s} d_i q_i y_i \big/ \sum_{i\in s} d_i q_i$, and $\overline{\mu} = \sum_{i\in s} d_i q_i \widehat{\mu}_i \big/ \sum_{i\in s} d_i q_i$.

For estimating distribution function $F(t)$, Wu and Sitter (2001) replaced y_i by $I(y_i \le t)$, $\widehat{\mu}_i$ by

$$\widehat{G}_i = \frac{1}{n}\sum_{j\in s} I\left(\frac{y_j - \hat{\mu}_j}{\sqrt{v(x_j)}} \le \frac{t - \hat{\mu}_i}{\sqrt{v(x_i)}}\right)$$ and substituted $q_i = 1$ in the expression (23.2.40). The resultant model calibrated estimator was $F(t)$ obtained as

$$\widehat{F}_{ws}(t) = \widehat{F}_{ht}(t) - \widehat{B}_{ws}\left(\frac{1}{N}\sum_{i\in s}\frac{\widehat{G}_i}{\pi_i} - \frac{1}{N}\sum_{i\in U}\widehat{G}_i\right) \tag{23.2.41}$$

where $\widehat{F}_{ht}(t) = \dfrac{1}{N}\displaystyle\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i}$, $\widehat{B}_{ws} = \dfrac{\left(\displaystyle\sum_{i\in s} d_i\left(\hat{G}_i - \overline{G}\right)\left(I(y_i \le t) - \overline{I}\right)\right)}{\left(\displaystyle\sum_{i\in s} d_i\left(\hat{G}_i - \overline{G}\right)^2\right)}$,

$\overline{G} = \dfrac{\displaystyle\sum_{i\in s} d_i\hat{G}_i}{\displaystyle\sum_{i\in s} d_i}$, and $\overline{I} = \dfrac{\displaystyle\sum_{i\in s} d_i I(y_i \le t)}{\displaystyle\sum_{i\in s} d_i}$.

Wu and Sitter (2001) also proposed pseudoempirical likelihood estimator of $F(t)$ as

$$F^*_{ws}(t) = \sum_{i\in s}\widehat{p}_i I(y_i \le t) \tag{23.2.42}$$

where the weights $\widehat{p}_i$ are obtained by maximizing pseudoempirical likelihood

$$l(p) = \sum_{i\in s} p_i I(y_i \le t)$$

subject to (i) $p_i \ge 0$, (ii) $\displaystyle\sum_{i\in s} p_i = 1$, and (iii) $\displaystyle\sum_{i\in s} p_i\widehat{G}_i = \frac{1}{N}\sum_{i\in U}\widehat{G}_i$.

Since $p_i \ge 0$, $F^*_{ws}(t)$ is a genuine distribution function. The calibrated estimator $\widehat{F}_{ws}$ and the pseudocalibrated estimator F^*_{ws} are asymptotically equivalent.

Rueda et al. (2007) considered a vector of auxiliary variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iq})'$, which are known for $i = 1,\ldots, N$. They defined a pseudovariable $g_i = \widehat{\boldsymbol{\beta}}'\mathbf{x}_i$ for $i = 1,\ldots, N$, where

$\widehat{\boldsymbol{\beta}} = \left(\sum_{i \in s} d_i q_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \sum_{i \in s} d_i q_i \mathbf{x}_i \gamma_i$ and q_i are known positive constants unrelated to $d_i = 1/\pi_i$. The proposed calibrated estimator is

$$\widehat{F}_r = \frac{1}{N} \sum_{i \in s} w_i I(\gamma_i \le t) \tag{23.2.43}$$

where the weights w_i were derived by minimizing the chi-square distance $\Phi_s = \sum_{i \in s} \frac{(w_i - d_i)^2}{d_i q_i}$ subject to the constraints

$$\frac{1}{N} \sum_{i \in s} w_i I(g_i \le t_j) = F_g(t_j) = \frac{1}{N} \sum_{i \in U} I(g_i \le t_j),\ j = 1, \ldots, p$$

for suitably chosen p points $t_1 < t_2 < \cdots < t_p$. Rueda et al. (2007), showed by simulation studies based on the actual data, that their estimator $\widehat{F}_r$ performs similar to $\widehat{F}_{cd}$ but better than $\widehat{F}_{rkm}$.

23.2.7 Method of Poststratification

Silva and Skinner (1995) poststratified the initial sample s into H strata $U_1, \ldots, U_h, \ldots, U_H$ where the unit $i \in U_h$ if $x_{(h-1)} < x_i \le x_{(h)}$ for $h = 1, \ldots, H$, $-\infty = x_{(0)} < x_{(1)} < x_{(2)} < \cdots < x_{(H)} = \infty$. Let s_h be the sample of size n_h from the hth stratum and $n = \sum_{h=1}^{H} n_h$. Silva and Skinner (1995) assumed that N_h, the stratum size of U_h, is known and n is so large that the probability of $n_h = 0$ is zero. Under this assumption, they proposed the following estimator of $F(t)$ as

$$\widehat{F}_{ss}(t) = \sum_{h=1}^{H} W_h \widehat{F}^*_h(t) \tag{23.2.44}$$

where $W_h = \frac{N_h}{N}$ and $\widehat{F}^*_h(t) = \left(\sum_{i \in s_h} \frac{I(\gamma_i \le t)}{\pi_i} \right) \Big/ \left(\sum_{i \in s_h} \frac{1}{\pi_i} \right)$

It can be easily checked that the estimator $\widehat{F}_{ss}(t)$ possesses all the properties of the distribution function stated in Eq. (23.2.1). Following Rao et al. (1990), Silva and Skinner (1995) derived the approximate expression of variance of $\widehat{F}_{ss}(t)$ as

$$Var\left\{ \widehat{F}_{ss}(t) \right\} \cong \frac{1}{N^2} \frac{1}{2} \sum_{i \ne} \sum_{j \in U} (\pi_i \pi_j - \pi_{ij}) \times \left(\frac{I(\gamma_i \le t) - F^*_{h(i)}(t)}{\pi_i} - \frac{I(\gamma_j \le t) - F^*_{h(j)}(t)}{\pi_j} \right)^2 \tag{23.2.45}$$

where $h(k)$ is the poststratum to which the unit k belongs and $F_h^*(t) = \frac{1}{N_h}\sum_{i\in U_h} I(y_i \le t)$ is the population distribution function of the hth stratum $h = 1,\ldots, H$.

An estimator of variance of $\widehat{F}_{ss}(t)$ was presented by Silva and Skinner (1995) as

$$\widehat{Var}\left\{\widehat{F}_{ss}(t)\right\} \cong \frac{1}{N^2}\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{I(y_i \le t) - \widehat{F}^*_{h(i)}}{\pi_i} - \frac{I(y_j \le t) - \widehat{F}^*_{h(j)}(t)}{\pi_j}\right)^2 \tag{23.2.46}$$

where $\widehat{F}^*_{h(i)}$ is an estimate of $F^*_{h(i)}$.

23.2.8 Empirical Comparison of the Estimators

Rao et al. (1990) did comprehensive studies of the performances of the design-based estimators $\widehat{F}_h(t), \widehat{F}_R(t), \widehat{F}_D(t), \widehat{F}^*(t)$ and the model-based estimator $\widehat{F}_{cd}(t)$ based on a population of sugarcane farms considered by Chambers and Dunstan (1986). The gross value of cane was treated as the study variable y while area under sugarcane was taken as auxiliary variable x. The population was found to obey the model (23.2.17) with $v(x) = \sqrt{x}$. Under simple random sampling, the relative biases of $\widehat{F}_D(t)$ and $\widehat{F}_{rkm}(t)$ are negligible and much less than that of $\widehat{F}_R(t)$. Whence efficiency is concerned, $\widehat{F}_{rkm}(t)$ is more efficient than $\widehat{F}_D(t)$ and $\widehat{F}(t)$. The ratio estimator $\widehat{F}_R(t)$ is less efficient than $\widehat{F}_D(t)$. The model-based estimator $\widehat{F}_{cd}(t)$ was found to be more efficient than the design-based estimators possibly because the population obeys the model. Rao et al. (1990) also considered the Hansen et al. (1983) population with $N = 14{,}000$ units to study the effect of model misspecification and computed relative bias and relative root mean square errors of the estimators $\widehat{F}_h(t), \widehat{F}_R(t)$, $\widehat{F}_D(t), \widehat{F}_{cd}(t)$, and $\widehat{F}_{rkm}(t)$. It was found that design-based estimators have negligible bias, even less than 1%, whereas the model-based estimator $\widehat{F}_{cd}(t)$ has a much larger relative bias of about 20%. The design-based estimator $\widehat{F}_{rkm}(t)$ was found to be much more efficient than the other design-based estimators $\widehat{F}_R(t)$ and $\widehat{F}_D(t)$ whereas the model-based estimator $\widehat{F}_{cd}(t)$ was found least efficient.

Silva and Skinner (1995) conducted simulation studies to compare performances of the estimators $\widehat{F}_h(t)$, $\widehat{F}_{ss}(t)$, $\widehat{F}_{ko}(t)$, $\widehat{F}_{kk}(t)$, $\widehat{F}_{cd}(t)$, and $\widehat{F}_{rkm}(t)$ based on two populations. The first population was the sugarcane farm data mentioned above, which was originally considered by Chambers and Dunstan (1986) and later Rao et al. (1990) and Kuk (1993). The second population comprises of 430 farms with 50 or more beef cattle, which was originally used by Chambers et al. (1993) and then Kuk (1993). Population 1 is a good working model for Eq. (23.2.17) as stated earlier whereas Population 2 is not. As far as relative bias is concerned, for both the populations, $\widehat{F}_{ss}(t)$ was found to have least bias while $\widehat{F}_{rkm}(t)$ occupied second place. $\widehat{F}_{ko}(t)$, $\widehat{F}_{kk}(t)$, and $\widehat{F}_{cd}(t)$ were found to have relatively high absolute relative biases. For Population 1, $\widehat{F}_{cd}(t)$ had least mean squared error (MSE) as expected because the model fits the population. The second and third places were occupied by $\widehat{F}_{kk}(t)$ and $\widehat{F}_{rkm}(t)$, respectively. Surprisingly, for Population 2, $\widehat{F}_{cd}(t)$ also had the smallest MSE followed by $\widehat{F}_{kk}(t)$. The estimators $\widehat{F}_{ss}(t)$ and $\widehat{F}_{ko}(t)$ were found to have moderate MSE. For further details, interested readers are referred to Rao et al. (1990) and Silva and Skinner (1995).

23.3 ESTIMATION OF QUANTILES

The αth quantile $\theta_y(\alpha), 0 < \alpha < 1$ of a finite population vector $\mathbf{y} = (y_1, \ldots, y_N)$ is defined as

$$\theta_y(\alpha) = inf\{t : F_y(t) \geq \alpha\} \tag{23.3.1}$$

where $F_y(t)$ is the distribution function of y. In case $\widehat{F}_y(t)$, an estimator of $F_y(t)$, is a monotonic nondecreasing function of t, the customary estimator of $\theta_y(\alpha)$ is obtained as

$$\widehat{\theta}_y(\alpha) = inf\{t : \widehat{F}_y(t) \geq \alpha\} \tag{23.3.2}$$

Let $\widehat{F}_x(t)$ be the customary estimator of $F_x(t)$. In case the population αth quantile $\theta_x(\alpha)$ of x is known, the ratio estimator of $\theta_y(\alpha)$ is given by

$$\widehat{\theta}_{ry}(\alpha) = \frac{\widehat{\theta}_y(\alpha)}{\widehat{\theta}_x(\alpha)}\theta_x(\alpha) \tag{23.3.3}$$

Similarly, a difference estimator of $\theta_y(\alpha)$ is given by

$$\widehat{\theta}_{dy}(\alpha) = \widehat{\theta}_y(\alpha) - \widehat{R}\left\{\widehat{\theta}_x(\alpha) - \theta_x(\alpha)\right\} \tag{23.3.4}$$

where $\widehat{R} = \dfrac{\sum_{i\in s} y_i/\pi_i}{\sum_{i\in s} x_i/\pi_i}$ is a consistent estimator of the population ratio $R = Y/X$.

Both the estimators $\widehat{\theta}_{ry}(\alpha)$ and $\widehat{\theta}_{dy}(\alpha)$ reduce to $\theta_y(\alpha)$ if $y_i \propto x_i \; \forall i \in U$. In this case the variances of the estimators become zero. Hence the estimators $\widehat{\theta}_{ry}(\alpha)$ and $\widehat{\theta}_{dy}(\alpha)$ are expected to produce a considerable gain in efficiency over $\widehat{\theta}_y(\alpha)$ if y_i is approximately proportional to x_i. Rao et al. (1990) derived the variances of $\widehat{\theta}_{ry}(\alpha)$ and $\widehat{\theta}_{dy}(\alpha)$ and also their unbiased estimators. These are omitted here because of their complexities.

23.4 ESTIMATION OF MEDIAN

The median of a variable y is obtained by substituting $\alpha = 1/2$ in Eq. (23.3.1). Thus the population median of y and x are, respectively,

$$\begin{aligned} \widetilde{\mu}_y &= \theta_y(1/2) = inf\{t : F_y(t) \geq 1/2\} \text{ and} \\ \widetilde{\mu}_x &= \theta_x(1/2) = inf\{t : F_x(t) \geq 1/2\} \end{aligned} \tag{23.4.1}$$

If the population median $\widetilde{\mu}_x = \theta_x(1/2)$ of the auxiliary variable x is known, then the ratio estimator of the population median $\widetilde{\mu}_y$ is obtained from Eq. (23.3.3) as

$$\widehat{\widetilde{\mu}}_{ry} = \widehat{\theta}_{ry}(1/2) = \frac{\widehat{\widetilde{\mu}}_y}{\widehat{\widetilde{\mu}}_x}\widetilde{\mu}_x \tag{23.4.2}$$

where $\widehat{\widetilde{\mu}}_x$ and $\widehat{\widetilde{\mu}}_y$ are the sample medians of x and y, respectively.

Kuk and Mak (1989) proposed the following modifications of the ratio estimator of $\widehat{\widetilde{\mu}}_{ry}$ based on simple random sampling.

23.4.1 Position Estimator and Stratification Estimator

At first, let us arrange y_i's, $i \in s$ in order of magnitude as $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$. Let i_0 be the number of observations of y_i's less than or equal to the population median $\widetilde{\mu}_y$, i.e., $y_{(i_0)} \leq \widetilde{\mu}_y < y_{(i_0+1)}$. Clearly, i_0 is unknown because $\widetilde{\mu}_y$ is unknown. Let $p = i_0/n$, then $\widetilde{\mu}_y$ is approximately the sample pth quantile $\widehat{\theta}_y(p)$. Suppose $\widehat{p}$ is an estimator of p, then $\widehat{\theta}_y(\widehat{p})$ is an estimator of the median $\widetilde{\mu}_y$. $\widehat{\widetilde{\mu}}_y$, the sample median of y can be viewed as the special estimator $\widehat{\theta}_y(\widehat{p})$ with $\widehat{p} = 1/2$. To estimate $\widehat{p}$, Kuk and Mak (1989) considered the following two-way classified table.

	$x \le \widetilde{\mu}_x$	$x > \widetilde{\mu}_x$	Total
$y \le \widetilde{\mu}_y$	P_{11}	P_{12}	P_{10}
$y > \widetilde{\mu}_y$	P_{21}	P_{22}	P_{20}
Total	P_{01}	P_{02}	1

In the table above, P_{ij} denotes the proportion of units in the population that belongs to the (i, j)th cell. Let n_x be the number of observations in the sample s with x_i's less than or equal to $\widetilde{\mu}_x$. In case P_{ij} are known, an estimate of p is given by

$$\begin{aligned} \widehat{p} &= \frac{1}{n}\left(n_x \frac{P_{11}}{P_{01}} + (n - n_x)\frac{P_{12}}{P_{02}}\right) \\ &\cong \frac{2}{n}\left(n_x P_{11} + (n - n_x)\left(\frac{1}{2} - P_{11}\right)\right) \end{aligned} \tag{23.4.3}$$

where $P_{0j} = P_{1j} + P_{2j} \cong 1/2$ for $j = 1, 2$.

In practice, P_{ij} are generally unknown. So, we estimate P_{ij} by p_{ij}, the proportion of units in the sample that fall in the (i, j)th cell, i.e., p_{11} is the proportion of observation in the class $x \le \widehat{\widetilde{\mu}}_x$(=sample median of x) and $y \le \widehat{\widetilde{\mu}}_y$. Now replacing P_{11} by p_{11} in Eq. (23.4.3), we find an estimate of p as

$$\widehat{p}_1 \cong \frac{2}{n}\left(n_x p_{11} + (n - n_x)\left(\frac{1}{2} - p_{11}\right)\right) \tag{23.4.4}$$

Thus an estimator of $\widetilde{\mu}_y$ is given by

$$\widehat{\widetilde{\mu}}_{py} = \widehat{\theta}_y(\widehat{p}_1) \tag{23.4.5}$$

The estimator $\widehat{\widetilde{\mu}}_{py}$ was termed as the "position estimator" by Kuk and Mak (1989).

For a given value of $y = t$, let $\widetilde{F}_{1y}(t)$ be the proportion of those units in the sample with $x \le \widetilde{\mu}_x$, which have y-values less than equal to t and $\widetilde{F}_{2y}(t)$ be the proportion of those units in the sample with $x > \widetilde{\mu}_x$, which have y-values less than equal to t. Then $F_y(t)$ can be estimated as

$$\begin{aligned} \widetilde{F}_y(t) &= \frac{1}{N}\left[N_x \widetilde{F}_{1y}(t) + (N - N_x)\widetilde{F}_{2y}(t)\right] \\ &\cong \frac{1}{2}\left[\widetilde{F}_{1y}(t) + \widetilde{F}_{2y}(t)\right] \end{aligned} \tag{23.4.6}$$

where N_x is the number of units in the population with $x \le \widetilde{\mu}_x$.

Since $\widetilde{F}_y(t)$ is a distribution function, an estimator of the median $\widetilde{\mu}_y$ is obtained as

$$\widehat{\widetilde{\mu}}_{sy} = inf\left\{\gamma : \widetilde{F}_y(t) \geq \frac{1}{2}\right\} \tag{23.4.7}$$

Kuk and Mak (1989) called the estimator $\widehat{\widetilde{\mu}}_{sy}$ as the "stratification estimator."

23.4.2 Comparison of the Efficiencies

Gross (1980) and Kuk and Mak (1989) derived the asymptotic distributions of the median estimators $\widehat{\widetilde{\mu}}_y$, $\widehat{\widetilde{\mu}}_{ry}$, $\widehat{\widetilde{\mu}}_{py}$, and $\widehat{\widetilde{\mu}}_{sy}$ when $N \to \infty$, $n \to \infty$ and $n/N \to f$, $0 \leq f \leq 1$. They assumed that as $N \to \infty$, the bivariate distribution of (x, y) approaches a continuous distribution with marginal densities $f_x(x)$ and $f_y(y)$ for x and y, respectively. Gross (1980) proved that the sample median $\widehat{\widetilde{\mu}}_y$ is consistent and asymptotically normally distributed with mean $\widetilde{\mu}_y$ and variance

$$Var\left(\widehat{\widetilde{\mu}}_y\right) = \frac{(1-f)}{4n}\left\{f_y\left(\widetilde{\mu}_y\right)\right\}^{-2} \tag{23.4.8}$$

Kuk and Mak (1989) derived the following results:

(i) $\widehat{\widetilde{\mu}}_{ry}$ is asymptotically normal with mean $\widetilde{\mu}_y$ and variance

$$Var\left(\widehat{\widetilde{\mu}}_{ry}\right) = \frac{(1-f)}{n}\left[\frac{1}{4}\left(f_y\left(\widetilde{\mu}_y\right)\right)^{-2} + \frac{1}{4}\widetilde{R}^2\left(f_x(\widetilde{\mu}_x)\right)^{-2} - 2\widetilde{R}\left(\frac{f_y\left(\widetilde{\mu}_y\right)}{f_x(\widetilde{\mu}_x)}\right)^{-1}\left(P_{11} - \frac{1}{4}\right)\right] \tag{23.4.9}$$

where $\widetilde{R} = \widetilde{\mu}_y/\widetilde{\mu}_x$.

(ii) $\widehat{\widetilde{\mu}}_{py}$ and $\widehat{\widetilde{\mu}}_{sy}$ both asymptotically follow the same distribution that is normal with mean $\widetilde{\mu}_y$ and variance

$$Var\left(\widehat{\widetilde{\mu}}_{py}\right) = Var\left(\widehat{\widetilde{\mu}}_{sy}\right) = \frac{2(1-f)P_{11}(1-2P_{11})}{n}\left(f_y\left(\widetilde{\mu}_y\right)\right)^{-2} \tag{23.4.10}$$

where P_{11} is the proportion of units in the population with $x \leq \widetilde{\mu}_x$ and $y \leq \widetilde{\mu}_y$. The probability P_{11} can be regarded as a measure of concordance.

It is important to note that if $P_{11} = 1/2$, then the asymptotic variances $Var\left(\widehat{\widetilde{\mu}}_{py}\right)$ and $Var\left(\widehat{\widetilde{\mu}}_{sy}\right)$ are both equal to zero. The expressions (23.4.8) and (23.4.10) show that the estimators $\widehat{\widetilde{\mu}}_{py}$ and $\widehat{\widetilde{\mu}}_{sy}$ are asymptotically more efficient than the sample median because

$$Var\left(\widehat{\widetilde{\mu}}_{sy}\right) - Var\left(\widehat{\widetilde{\mu}}_{y}\right) = \frac{1}{4}(4P_{11} - 1)^2 \geq 0 \tag{23.4.11}$$

The estimator $\widehat{\widetilde{\mu}}_{ry}$ becomes asymptotically more efficient than the sample median $\widehat{\widetilde{\mu}}_{y}$ if

$$\rho_c > \frac{1}{2}\left(\widetilde{R}\frac{f_y\left(\widetilde{\mu}_y\right)}{f_x(\widetilde{\mu}_x)}\right) \tag{23.4.12}$$

where $\rho_c = 4(P_{11} - 1/4)$ varies from -1 to 1 as P_{11} increases from 0 to 1/2.

23.4.3 Further Generalization

Let the vector of the auxiliary variable $\mathbf{x} = (x_1,\ldots, x_N)$ be known and the range of the auxiliary variable be partitioned into r mutually exclusive and exhaustive class intervals $(a_0, a_1]$, $(a_1, a_2],\ldots, (a_{r-1}, a_r]$ with $a_0 = 0$, $a_r = \infty$, and $a_j = \theta_x(\alpha_j)$. Let P_{1j} be the proportion of the units in the population with $y \leq \widetilde{\mu}_y$ and x falling in the class $(a_{j-1}, a_j]$ for $j = 1,\ldots, r$ and let p_{1j} be the proportion of units in s with $y \leq \widehat{\widetilde{\mu}}_y$ and x in $\left(\widehat{a}_{j-1}, \widehat{a}_j\right]$ where $\widehat{a}_j = \widehat{\theta}_x(\alpha_j)$. Also, let P_{0j} be the proportion of units in the population with $x \in (a_{j-1}, a_j]$ and p_{oj} be the proportion in the sample with $x \in \left(\widehat{a}_{j-1}, \widehat{a}_j\right]$, whereas n_{xj} is the number of units in the sample with $x \in (a_{j-1}, a_j]$.

Proportions in population

	$(a_0, a_1]$	—	$(a_{j-1}, a_j]$	—	$(a_{r-1}, a_r]$
$y \leq \widetilde{\mu}_y$	P_{11}	—	P_{1j}	—	P_{1r}
$y > \widetilde{\mu}_y$	P_{21}	—	P_{2j}	—	P_{2r}
Total	P_{01}	—	P_{0j}	—	P_{0r}

Proportions in sample

	$(a_0, a_1]$	—	$(a_{j-1}, a_j]$	—	$(a_{r-1}, a_r]$
$y \leq \widehat{\widetilde{\mu}}_y$	p_{11}	—	p_{1j}	—	p_{1r}
$y > \widehat{\widetilde{\mu}}_y$	p_{21}	—	p_{2j}	—	p_{2r}
Total	p_{01}	—	p_{0j}	—	p_{0r}

The estimated proportion of y's in the sample that is less than equal to $\widetilde{\mu}_y$ is given by Kuk and Mak (1989) as

$$\begin{aligned}\widetilde{p}_{g1} &= \frac{1}{n}\sum_{j=1}^{r} n_{xj}\frac{p_{1j}}{p_{0j}} \\ &= \frac{1}{n}\sum_{j=1}^{r} n_{xj}\frac{p_{1j}}{(\alpha_j - \alpha_{j-1})}\end{aligned} \tag{23.4.13}$$

Finally, the estimated populated median is

$$\widehat{\widetilde{\mu}}_{gpy} = \widehat{\theta}_y\left(\widetilde{p}_{g1}\right) \tag{23.4.14}$$

Let $\widehat{F}_{yj}(t)$ be the proportion of the units in the sample with $x \in (a_{j-1}, a_j]$ that have y-values less than or equal to t. Then $F_y(t)$ may be estimated by

$$\widehat{F}_y(t) = \sum_{j=1}^{r}(\alpha_j - \alpha_{j-1})\widehat{F}_{yj}(t)$$

Consequently the estimated median is

$$\widehat{\widetilde{\mu}}_{gsy} = inf\left\{y : \widehat{F}_y(t) \geq 1/2\right\} \tag{23.4.15}$$

Kuk and Mak (1989) showed that the asymptotic variances of $\widehat{\widetilde{\mu}}_{gpy}$ and $\widehat{\widetilde{\mu}}_{gsy}$ are the same and equal to

$$V_{ar}\left(\widehat{\widetilde{\mu}}_{gpy}\right) = V_{ar}\left(\widehat{\widetilde{\mu}}_{gsy}\right) = \frac{1-f}{n}\frac{1}{\left(f_y\left(\widetilde{\mu}_y\right)\right)^2}\left(\frac{1}{2} - \sum_{j=1}^{r}\frac{P_{1j}^2}{(\alpha_j - \alpha_{j-1})}\right) \tag{23.4.16}$$

23.4.4 Empirical Comparison

Kuk and Mak (1989) compared efficiencies of the estimators $\widehat{\widetilde{\mu}}_y, \widehat{\widetilde{\mu}}_{ry}, \widehat{\widetilde{\mu}}_{py}$, and $\widehat{\widetilde{\mu}}_{sy}$ empirically using four populations named Hospitals, Counties 70, Villages, and Factories. The first two populations were used by Royall and Cumberland (1981), where x and y are well correlated and a linearity relationship holds. For the population Villages (Murthy, 1967), where x (area in 1951) and y (number of households in 1961) are poorly correlated, the probability of concordance P_{11} is also low. For the population Factories (Murthy, 1967), x (number of workers) and y (output) are not linearly

related. From the first two populations, 500 independent samples of suitable sizes were selected by SRSWOR method while from each of the remaining other two populations, 1000 independent samples of suitable sizes were selected. Empirical studies reveal that for the populations Hospitals and Counties where x and y are linearly related, all three estimators $\widehat{\widetilde{\mu}}_{ry}$, $\widehat{\widetilde{\mu}}_{py}$, and $\widehat{\widetilde{\mu}}_{sy}$ have much lower mean square errors than that of the sample median $\widehat{\widetilde{\mu}}_{y}$. For the population Villages where x and y are poorly related, the ratio estimator $\widehat{\widetilde{\mu}}_{ry}$ performed worse than the sample median $\widehat{\widetilde{\mu}}_{y}$. But the other two estimators $\widehat{\widetilde{\mu}}_{py}$ and $\widehat{\widetilde{\mu}}_{sy}$ fare better than $\widehat{\widetilde{\mu}}_{y}$. For the nonlinear population Factories, $\widehat{\widetilde{\mu}}_{py}$ and $\widehat{\widetilde{\mu}}_{sy}$ still performed better than $\widehat{\widetilde{\mu}}_{y}$ where the ratio estimator $\widehat{\widetilde{\mu}}_{ry}$ is again outperformed by the sample median. Thus the efficiencies of the position estimator $\widehat{\widetilde{\mu}}_{py}$ and stratification estimator $\widehat{\widetilde{\mu}}_{sy}$ do not depend on the validity of linearity assumption and hence it is much safer to use, than the ratio estimator $\widehat{\widetilde{\mu}}_{ry}$.

23.5 CONFIDENCE INTERVAL FOR DISTRIBUTION FUNCTION AND QUANTILES

The conventional $(1-\alpha)100\%$ confidence interval of the distribution function $F(t)$ is

$$\left(\widehat{F}(t) - z_{\alpha/2}\sqrt{\widehat{V}\left[\widehat{F}(t)\right]},\ \widehat{F}(t) + z_{\alpha/2}\sqrt{\widehat{V}\left[\widehat{F}(t)\right]}\right) \tag{23.5.1}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile from $N(0, 1)$ and $\widehat{V}\left[\widehat{F}(t)\right]$ is an estimator of $V\left[\widehat{F}(t)\right]$. The validity of the confidence interval obviously depends on the asymptotic normality of the distribution

$$\frac{\widehat{F}(t) - F(t)}{\sqrt{\widehat{V}\left[\widehat{F}(t)\right]}}$$

which is justified when the sample size is large. However, for small to moderate sample size, because of range constraint $0 \le \widehat{F}(t) \le 1$, the distribution of $\widehat{F}(t)$ for t at large or small quantiles is usually not symmetric. For small and moderate sample size, the performance of the interval (23.5.1) is very often unsatisfactory; coverage probability is generally lower than the nominal value and two tail probabilities are unbalanced (Wu, 1999).

Chen and Wu (2002) proposed an alternative method. In this method a smooth and monotone function g is chosen so that the distribution of

$\widehat{W} = \widehat{W}(t) = g\left[\widehat{F}(t)\right]$ is better approximated by the normal distributions. Two such popular transformations are logit transformation and complementary log−log transformation: $\widehat{W} = \log\left(\frac{\hat{F}(t)}{1 - \hat{F}(t)}\right)$ and $\log\left(-\log\left\{\widehat{F}(t)\right\}\right)$. Noting that $\widehat{W}$ asymptotical normal with mean $W = W(t) = g[F(t)]$ and variance $V\left(\widehat{W}\right) = \left(g'\{F(t)\}\right)^2 V\left\{\widehat{F}(t)\right\}$, we can find a $(1 - \alpha)100\%$ confidence interval of W as

$$\left(\widehat{W} - z_{\alpha/2}\sqrt{\widehat{V}\left(\widehat{W}\right)},\ \ \widehat{W} + z_{\alpha/2}\sqrt{\widehat{V}\left(\widehat{W}\right)}\right)$$

where $\widehat{V}\left(\widehat{W}\right) = \left(g'\left\{\widehat{F}(t)\right\}\right)^2 \widehat{V}\left\{\widehat{F}(t)\right\}$ is a suitable estimator of $V\left(\widehat{W}\right)$.

Finally, transformed confidence interval for $F(t)$ is obtained as

$$\left(g^{-1}\left\{\widehat{W} - z_{\alpha/2}\sqrt{\widehat{V}\left(\widehat{W}\right)}\right\},\ g^{-1}\left\{\widehat{W} + z_{\alpha/2}\sqrt{\widehat{V}\left(\widehat{W}\right)}\right\}\right) \quad (23.5.2)$$

On the basis of simulation studies, Wu (1999) showed that performance of the intervals (23.5.1) and (23.5.2) is similar when t is in the middle range of quantiles but Eq. (23.5.2) is much better when t is small or large quantiles.

Let $\widehat{\theta}(\alpha)$ be an estimator of the $\alpha(0 < \alpha < 1)$th quantile $\theta(\alpha)$ obtained from the inversion of estimated distribution function $\widehat{F}(t)$, i.e., $\widehat{F}\left\{\widehat{\theta}(\alpha)\right\} = \alpha$. Furthermore, let $L(t)$ and $U(t)$ be the lower and upper $(1 - \alpha)100\%$ confidence intervals of $F(t)$ obtained from any of the formulas (23.5.1) and (23.5.2). Then following Woodruff (1952), the confidence interval of the $\alpha(0 < \alpha < 1)$th quantile $\theta(\alpha)$ is obtained as

$$\left(\widehat{F}^{-1}\left[L\left\{\widehat{\theta}(\alpha)\right\}\right],\ \ \widehat{F}^{-1}\left[U\left\{\widehat{\theta}(\alpha)\right\}\right]\right) \quad (23.5.3)$$

Obviously the confidence interval based on the formula (23.5.2) is expected to perform better than that on Eq. (23.5.1).

Godambe and Thompson (1999) used estimating functions for determination of confidence intervals of distribution functions and quantiles while Chen and Sitter (1999), Wu and Sitter (2001), among others, used empirical likelihood methods. Details have been given in Chapters 22 and 25, respectively.

23.6 CONCLUDING REMARKS

In surveys, estimation of distribution function and quantiles often play an important role. Rao et al. (1990) showed how one can use auxiliary information x to construct customary design-based ratio $\widehat{F}_R(t)$ and difference estimator $\widehat{F}_D(t)$ of the population distribution function $F(t)$, which can improve upon $\widehat{F}(t)$, the conventional estimator of the distribution function $F(t)$. The proposed estimators $\widehat{F}_R(t)$ and $\widehat{F}_D(t)$ are asymptotically design unbiased but not model unbiased. Chambers and Dunstan (1986) proposed a model-based estimator $\widehat{F}_{cd}(t)$, which is efficient if the assumed model is valid, but this estimator can perform poorly under model misspecification. Rao et al. (1990) also proposed model-based estimators $\widehat{F}^{*}(t)$ and $\widehat{F}_{rkm}(t)$, which are both asymptotically design unbiased and model-unbiased under the model (23.2.17) with $v(x_i) = \sqrt{x_i}$. Moreover, $\widehat{F}_{rkm}(t)$ can be modified using multiauxiliary variables. Poststratification estimator $\widehat{F}_{ss}(t)$ was proposed by Silva and Skinner (1995) and it retains the property of asymptotic design unbiasedness. To compute Kuk's (1993) estimator $\widehat{F}_{kk}(t)$, the values of the study variable y need to be scaled since the bandwidth that is used to control the smoothing is the same for both the variables x and y. The estimators $\widehat{F}_R(t)$, $\widehat{F}_D(t)$, $\widehat{F}^{*}(t)$, $\widehat{F}_{cd}(t)$, and $\widehat{F}_{ss}(t)$ are advantageous because the variance estimators are readily available and easy to compute. Although the estimator of variance of $\widehat{F}_{rkm}(t)$ is readily available, it can be extremely complex for a varying probability sampling scheme because of the involvement of third-order inclusion probabilities π_{ijk} that are difficult to compute. Extensive simulation studies were done by Chambers and Dunstan (1986), Rao et al. (1990), Chambers et al. (1993), Kuk (1993), and Silva and Skinner (1995), among others, to compare performances among the proposed estimators. The simulation studies do not finger out any particular estimator as the best in all situations. However, $\widehat{F}_{rkm}(t)$ seems to perform well in most of the situations irrespective of the failure of the model and so it is safe to be used.

Estimators of the distribution functions can be extended to estimation of population quantiles and medians. Kuk and Mak (1989) proposed the alternative position estimator $\widehat{\widetilde{\mu}}_{py}$ and stratification estimator $\widehat{\widetilde{\mu}}_{sy}$ to estimate the median under SRSWOR sampling. Empirical studies of Kuk and Mak (1989) reveal that the estimators $\widehat{\widetilde{\mu}}_{py}$ and $\widehat{\widetilde{\mu}}_{sy}$ are more efficient than the sample median $\widehat{\widetilde{\mu}}_{y}$ and the ratio estimator $\widehat{\widetilde{\mu}}_{ry}$ even if the relation between x and y is not linear. The use of empirical likelihood method for estimating

finite population distribution function was considered by Owen (1988), Chen and Sitter (1999), Wu and Sitter (2001). The proposed estimators use auxiliary information effectively at the estimation stage and possess attractive properties. Godambe and Thompson (1999) used the method of estimating function whereas Chen and Wu (2002) used calibrated pseudoempirical likelihood methods for determining confidence intervals. Further discussions are given by Johnson (2003), Drofman (2009), among others.

23.7 EXERCISES

23.7.1 Let a sample of size 10 be selected from a population of size 50 by simple random sampling with replacement method.

Selected sample	1	2	3	4	5	6	7	8	9	10
y-values	10	8	6	15	3	8	5	6	3	1

(i) Sketch the empirical distribution function.
(ii) Estimate the 25th, 50th, and 80th percentiles from the graph. Estimate the standard errors of the estimators used. Determine 95% confidence interval of the population median.

23.7.2 A sample of size of size 8 is selected from a finite population of size 50 using Poisson sampling scheme. The following table gives the units selected in the sample, y-values, and inclusion probabilities (π_i) of the selected units.

Selected sample	1	2	3	4	5	6	7	8
y-values	15	10	5	4	6	15	20	15
π_i	0.10	0.15	0.20	0.25	0.15	0.10	0.25	0.10

Sketch the distribution function using the formulae.

(i) $\widehat{F}(t) = \frac{1}{N}\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i}$

(ii) $\widehat{F}_h(t) = \sum_{i\in s}\frac{I(y_i \le t)}{\pi_i}\Big/\sum_{i\in s}\frac{1}{\pi_i}$

(iii) $\widehat{F}_c(t) = 1 - \frac{1}{N}\sum_{i\in s}\frac{1 - I(y_i \le t)}{\pi_i}$

Discuss the appropriateness of the formulae. From the sketches, obtain estimates of the population median. Compute standard errors of the estimators used.

23.7.3 A sample of 10 plants is selected at random from 40 plants of a garden. The following table gives the height (y) of 10 plants along with the diameter (x) of all the 40 plants in that garden. Fit distribution functions over the data using the following methods:

(i) Ratio method: $\widehat{F}_R(t) = \frac{1}{N}\widehat{\lambda}\sum_{i\in U}\frac{I(\hat{y}_i \le t)}{\pi_i}$

where $\widehat{\lambda} = \left(\frac{\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i}}{\sum_{i\in s}\frac{I(\hat{y}_i \le t)}{\pi_i}}\right)$ and $\widehat{y}_i = \left(\frac{\sum_{i\in s} y_i}{\sum_{i\in s} y_i}\right)x_i$

(ii) Difference method:

$$\widehat{F}_D(t) = \frac{1}{N}\left[\sum_{i\in s}\frac{I(y_i \le t)}{\pi_i} - \left\{\sum_{i\in s}\frac{I(\hat{y}_i \le t)}{\pi_i} - \sum_{i\in U} I(\widehat{y}_i \le t)\right\}\right]$$

From the fitted distribution functions, estimate median and 80th percentiles of the heights of the plants. Also estimate standard errors of the estimators used.

Plants	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Diameter (in cm)	15	20	30	25	30	40	15	12	20	30
Height (in cm)	80	90	120	100	100					

Plants	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Diameter (in cm)	15	20	30	25	30	40	15	12	20	30

Plants	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Diameter (in cm)	10	20	18	28	50	60	45	20	25	30

Plants	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
Diameter (in cm)	15	25	20	10	10	30	25	40	60	30

23.7.4 Continuation of Exercise 27.7.3. Assume that y and x are related to the model $y_i = \beta x_i + \in_i$, where $\in_i$'s are independent with $E_m(\in_i) = 0$ and $V_m(\in_i) = \sigma^2 x_i$. Estimate the median height of the plant using the Chamber and Dunstan estimator

$$\widehat{F}_{cd}(t) = \frac{1}{N}\left[\sum_{i\in s} I(y_i \le t) - \sum_{i\in U-s}\left\{\frac{1}{n}\sum_{j\in s} I\left(\frac{\left(y_i - \widehat{\beta}x_j\right)}{\sqrt{v(x_j)}} \le \frac{\left(t - \widehat{\beta}x_j\right)}{\sqrt{v(x_j)}}\right)\right\}\right]$$

where $\widehat{\beta} = \sum_{j\in s}\frac{y_i x_i}{\sqrt{v(x_j)}} \Big/ \sum_{j\in s}\frac{x_i^2}{\sqrt{v(x_j)}}$ and $v(x_j) = \sigma^2 x_j^2$.

23.7.5 The following table relates to the daily wages in (US$) of 40 factory workers selected at random from 120 workers.

120	80	160	200	400	550	300	400	500	150
140	90	150	180	300	250	200	180	150	250
270	380	920	150	200	225	420	100	120	150
80	75	150	200	500	590	250	140	150	200

Estimate the median wage of the factory workers and obtain 90% confidence interval of the median wages.