Quantile Estimation in Successive Sampling

Housila P. Singh and Ritesh Tailor
School of Studies in Statistics
Vikram University
Ujjain -456010, M.P.
India

Sarjinder Singh
Department of Statistics
St. Cloud State University
St. Cloud, MN 56301-4498
USA

Jong-Min Kim
Statistics, Division of Science and Mathematics
University of Minnesota - Morris
Morris, MN 56267
USA
Email: jongmink@morris.umn.edu

Summary

In successive sampling on two occasions the problem of estimating a finite population quantile has been considered. The theory developed aims at providing the optimum estimates by combining (i) three double sampling estimators viz. ratio-type, product-type and regression-type, from the matched portion of the sample and (ii) a simple quantile based on a random sample from the unmatched portion of the sample on the second occasion. The approximate variance formulae of the suggested estimators have been obtained. Optimal matching fraction is discussed. A simulation study is carried out in order to compare the three estimators and direct estimator. It is found that the performance of the regression-type estimator is the best among all the estimators discussed here.

Keywords: Finite population quantile; Successive sampling; Partial replacement; Auxiliary information.

1. Introduction

The problem of quantile estimation often arises when variables with a highly skewed distribution, such as income, are studied. When there is an extensive literature on the estimation of mean and total in sample surveys, relatively less efforts have been made in the development of efficient procedures for estimating finite population quantiles. It is well known that the use of auxiliary information at the estimation stage can typically increase the precision of estimates of a parameter.

A large number of estimators for estimating population mean based on auxiliary information are available in the literature with their properties under simple random sampling design and other sampling designs. However, few authors including Chambers and Dustan (1986), Kuk and Mak (1989), Rao et al. (1990), Mak and Kuk (1993), Kuk (1993), Rueda et al. (1998), Rueda and Arcoss (2001), Allen et al. (2001), Singh et al. (2001), Singh and Joarder (2002), Singh et al. (2003) and Singh (2003) have discussed the problem of estimating finite population median/quantile using auxiliary information in sample surveys.

The problem of sampling on two successive occasions was first introduced by Jessen (1942) and related review is available in Biradar and Singh (2001). It is to be mentioned that the study relating to environmental issues frequently involves variables with extreme values which influence the value of mean. In such situations the estimation of second quantile assumes importance, as it is not affected by the extreme values. This led authors to consider the problem of estimation of quantile under successive sampling. In this paper we have suggested three estimators viz. (i) ratio-type (ii) product-type and (iii) regression type with their properties.

2. Suggested estimators in successive sampling

Consider a finite population $\Omega = \{U_1, U_2, ..., U_N\}$ of N identifiable units which is supposed to be sampled on two occasions. Assume that size of the population remains unchanged but values of unit changes over two occasions. Let y_i (x_i) be the value of the variate under study for the ith unit of the second (first) occasion. On the first occasion an initial sample of size n_1 is selected by simple random sampling without replacement (SRSWOR) scheme. Out of these n_1 units, m units (called a 'matched' sample) are retained on the second occasion while a fresh simple random sample of size $u = (n - m)$ (called 'unmatched' sample) is drawn without replacement on the second occasion from the remaining $(N - n_1)$ units of the population so that the total sample size at the second occasion becomes $n = (m + u)$. Let y_1, y_2, K, y_N be the values of the population elements U_1, U_2, K, U_N, for the variable of interest y. For any y $(-\infty < y < \infty)$, the population distribution function $F_y(y)$ is defined as the proportion of elements in the population that are less than or equal to y. The finite population β-quantile of y is given by

$$Q_y(\beta) = \inf\{y;\ F_y(y) \geq \beta\} = F_y^{-1}(\beta).$$

The problem under investigation is to estimate the population quantiles $Q_y(\beta)$ of order β $(0 < \beta < 1)$ on the current (second) occasion. Let

$\hat{Q}_x(\beta) = \inf\{x;\ \hat{F}_x(x) \geq \beta\}$

be the sample quantile of order β on the first occasion and $\hat{Q}_y(\beta) = \inf\{y;\ \hat{F}_y(y) \geq \beta\}$ the sample quantile on the current (second) occasion, noting that $\hat{F}_x(x)$ ($\hat{F}_y(y)$) is a monotone nondecreasing function x (y). Denote by $\hat{Q}_{x(m)}(\beta)$ and $\hat{Q}_{y(m)}(\beta)$ be the sample quantiles of the matched sample on the first and second occasions respectively, and $\hat{Q}_{y(u)}(\beta)$ the sample quantile of the unmatched sample on the current occasion. For estimating the population quantile $Q_y(\beta)$ based on successive sampling, two independent estimators can be made. First, based on sample of size u drawn fresh on the current occasion and second based on the sample of size m common to both the occasions. Thus we define the following estimators:

$$\text{(i)}\quad \hat{Q}_r = \alpha\ \hat{Q}^{(r)}_{y(m)}(\beta) + (1-\alpha)\hat{Q}_{y(u)}(\beta), \quad \text{(ratio-type)} \tag{2.1}$$

$$\text{(ii)}\quad \hat{Q}_p = \delta\ \hat{Q}^{(p)}_{y(m)}(\beta) + (1-\delta)\hat{Q}_{y(u)}(\beta), \quad \text{(product-type)} \tag{2.2}$$

$$\text{(iii)}\quad \hat{Q}_l = \gamma\ \hat{Q}^{(l)}_{y(m)}(\beta) + (1-\gamma)\hat{Q}_{y(u)}(\beta), \quad \text{(regression-type)} \tag{2.3}$$

where α ,δ and γ are suitably chosen scalars,

$$\hat{Q}^{(r)}_{y(m)} = \hat{Q}_{y(m)}(\beta)\left\{\frac{\hat{Q}_x(\beta)}{\hat{Q}_{x(m)}(\beta)}\right\}, \quad \text{(ratio-type)} \tag{2.4}$$

$$\hat{Q}^{(p)}_{y(m)} = \hat{Q}_{y(m)}(\beta)\left\{\frac{\hat{Q}_{x(m)}(\beta)}{\hat{Q}_x(\beta)}\right\}, \quad \text{(product-type)} \tag{2.5}$$

and

$$\hat{Q}^{(l)}_{y(m)} = \hat{Q}_{y(m)}(\beta) + b\left\{\hat{Q}_x(\beta) - \hat{Q}_{x(m)}(\beta)\right\}, \quad \text{(regression-type)} \tag{2.6}$$

with

$$b = \frac{\hat{f}\left(\hat{Q}_{x(m)}(\beta)\right)}{\hat{f}\left(\hat{Q}_{y(m)}(\beta)\right)}\left\{\frac{\hat{P}_{xy(m)}}{\beta(1-\beta)} - 1\right\} \tag{2.7}$$

$\hat{P}_{xy}$ is the proportion of elements in sample such that $y \leq \hat{Q}_{y(m)}(\beta)$ and $x \leq \hat{Q}_{x(m)}(\beta)$, $\hat{f}_x\left(\hat{Q}_{x(m)}(\beta)\right)$ and $\hat{f}_y\left(\hat{Q}_{y(m)}(\beta)\right)$ are the estimates of $f_x(Q_x(\beta))$ and $f_y(Q_y(\beta))$ respectively determined by the method as adopted by Silverman (1986) where $f_y(\cdot)$ ($f_x(\cdot)$) is the derivative of $\tilde{F}_y(\cdot)$ ($\tilde{F}_x(\cdot)$) , the limiting value of $F_y(\cdot)$ ($F_x(\cdot)$) as $N \to \infty$, for instance, see Randles (1982) and Rao et al. (1990). Thus we note that as $N \to \infty$ the distribution of the bivariate variable (x, y) approaches a continuous distribution with marginal densities $f_y(\cdot)$ and $f_x(\cdot)$ for x and y respectively, see Kuk and Mak (1989, p. 264).

The variances of $\hat{Q}_r$, $\hat{Q}_p$ and $\hat{Q}_l$ are respectively given by

$$V\left(\hat{Q}_r\right) = \left[V\left(\hat{Q}_{y(u)}(\beta)\right) + \alpha^2\left\{V\left(\hat{Q}_{y(u)}(\beta)\right) + V\left(\hat{Q}^{(r)}_{y(m)}\right)\right\} - 2\alpha V\left(\hat{Q}_{y(u)}(\beta)\right)\right] \tag{2.8}$$

$$V\left(\hat{Q}_p\right) = \left[V\left(\hat{Q}_{y(u)}(\beta)\right) + \alpha^2\left\{V\left(\hat{Q}_{y(u)}(\beta)\right) + V\left(\hat{Q}^{(p)}_{y(m)}\right)\right\} + 2\alpha V\left(\hat{Q}_{y(u)}(\beta)\right)\right] \tag{2.9}$$

$$V\left(\hat{Q}_l\right) = \left[V\left(\hat{Q}_{y(u)}(\beta)\right) + \alpha^2\left\{V\left(\hat{Q}_{y(u)}(\beta)\right) + V\left(\hat{Q}^{(l)}_{y(m)}\right)\right\} + 2\alpha V\left(\hat{Q}_{y(u)}(\beta)\right)\right] \tag{2.10}$$

where the variances of $\hat{Q}_{y(u)}(\beta)$, $\hat{Q}^{(r)}_{y(m)}(\beta)$, $\hat{Q}^{(p)}_{y(m)}(\beta)$ and $\hat{Q}^{(l)}_{y(m)}(\beta)$ to the first degree of approximation (or alternatively, of order n^{-1}) are respectively given by

$$V\left(\hat{Q}_{y(u)}(\beta)\right) = \frac{(1-f_u)}{u}\frac{\beta(1-\beta)}{\left(f_y\left(Q_y(\beta)\right)\right)^2} = \frac{(1-f_u)}{u}A(y,\beta) \tag{2.11}$$

$$V\left(\hat{Q}^{(r)}_{y(m)}(\beta)\right) = A(y,\beta)\left[\frac{(1-f_m)}{m} + \left(\frac{1}{m} - \frac{1}{n_1}\right)\theta(\theta - 2\rho_c)\right] \tag{2.12}$$

$$V\left(\hat{Q}^{(p)}_{y(m)}(\beta)\right) = A(y,\beta)\left[\frac{(1-f_m)}{m} + \left(\frac{1}{m} - \frac{1}{n_1}\right)\theta(\theta + 2\rho_c)\right] \tag{2.13}$$

$$V\left(\hat{Q}^{(l)}_{y(m)}(\beta)\right) = A(y,\beta)\left[\frac{(1-f_m)}{m} - \left(\frac{1}{m} - \frac{1}{n_1}\right)\rho_c^2\right] \tag{2.14}$$

where

$f_u = \frac{u}{N}$, $f_m = \frac{m}{N}$, $A(y,\beta) = \frac{\beta(1-\beta)}{\left(f_y\left(Q_y(\beta)\right)\right)^2}$, $\theta = \frac{Q_y(\beta) f_y\left(Q_y(\beta)\right)}{Q_x(\beta) f_x\left(Q_x(\beta)\right)}$,

$\rho_c = \left\{ \frac{P_{xy}}{\beta(1-\beta)} - 1 \right\}$ and P_{xy} denotes the proportion of units in the population with $x \le Q_x(\beta)$ and $y \le Q_y(\beta)$. (See, Kuk and Mak (1989, p. 268))

Thus we get the variances of $\hat{Q}_r$, $\hat{Q}_p$ and $\hat{Q}_l$ to the first degree of approximation as

$$V\left(\hat{Q}_r\right) = A(y,\beta)\left[\frac{(1-f_u)(1-2\alpha)}{u} + \alpha^2\left\{\frac{(1-f_u)}{u} + \frac{(1-f_m)}{m} + \left(\frac{1}{m} - \frac{1}{n_1}\right)\theta(\theta - 2\rho_c)\right\}\right] \tag{2.15}$$

$$V\left(\hat{Q}_p\right) = A(y,\beta)\left[\frac{(1-f_u)(1-2\delta)}{u} + \delta^2\left\{\frac{(1-f_u)}{u} + \frac{(1-f_m)}{m} + \left(\frac{1}{m} - \frac{1}{n_1}\right)\theta(\theta + 2\rho_c)\right\}\right] \tag{2.16}$$

and

$$V\left(\hat{Q}_l\right) = A(y,\beta)\left[\frac{(1-f_u)(1-2\delta)}{u} + \gamma^2\left\{\frac{(1-f_u)}{u} + \frac{(1-f_m)}{m} - \left(\frac{1}{m} - \frac{1}{n_1}\right)\rho_c^2\right\}\right] \tag{2.17}$$

which are respectively minimized for

$$\alpha = \frac{(1-f_u)/u}{\left[\frac{(1-f_u)}{u} + \frac{(1-f_m)}{m} + \left(\frac{1}{m} - \frac{1}{n_1}\right)\theta(\theta - 2\rho_c)\right]} \tag{2.18}$$

$$\delta = \frac{(1-f_u)/u}{\left[\frac{(1-f_u)}{u} + \frac{(1-f_m)}{m} + \left(\frac{1}{m} - \frac{1}{n_1}\right)\theta(\theta + 2\rho_c)\right]} \tag{2.19}$$

and

$$\gamma=\frac{(1-f_u)/u}{\left[\frac{(1-f_u)}{u}+\frac{(1-f_m)}{m}-\left(\frac{1}{m}-\frac{1}{n_1}\right)\rho_c^2\right]} \tag{2.20}$$

Hence by inserting (2.18) to (2.20) in (2.15) to (2.17) the resulting (minimum) variances of $\hat{Q}_r$, $\hat{Q}_p$ and $\hat{Q}_l$ are respectively give by

$$\min.V\left(\hat{Q}_r\right)=\frac{(1-f_u)}{u}A(y,\beta)\frac{\left[\frac{(1-f_m)}{m}+\left(\frac{1}{m}-\frac{1}{n_1}\right)\theta(\theta-2\rho_c)\right]}{\left[\frac{(1-f_u)}{u}+\frac{(1-f_m)}{m}+\left(\frac{1}{m}-\frac{1}{n_1}\right)\theta(\theta-2\rho_c)\right]}, \tag{2.21}$$

$$\min.V\left(\hat{Q}_p\right)=\frac{(1-f_u)}{u}A(y,\beta)\frac{\left[\frac{(1-f_m)}{m}+\left(\frac{1}{m}-\frac{1}{n_1}\right)\theta(\theta+2\rho_c)\right]}{\left[\frac{(1-f_u)}{u}+\frac{(1-f_m)}{m}+\left(\frac{1}{m}-\frac{1}{n_1}\right)\theta(\theta+2\rho_c)\right]}, \tag{2.22}$$

and

$$\min.V\left(\hat{Q}_l\right)=\frac{(1-f_u)}{u}A(y,\beta)\frac{\left[\frac{(1-f_m)}{m}-\left(\frac{1}{m}-\frac{1}{n_1}\right)\rho_c^2\right]}{\left[\frac{(1-f_u)}{u}+\frac{(1-f_m)}{m}-\left(\frac{1}{m}-\frac{1}{n_1}\right)\rho_c^2\right]}. \tag{2.23}$$

In the next section, we will consider analytical comparisons of the proposed estimators.

3. Analytical comparisons

From (2.11), (2.21), (2.22) and (2.23) we have

$$V\left(\hat{Q}_{y(u)}(\beta)\right)-\min.V\left(\hat{Q}_r\right)=\frac{(1-f_u)^2}{u^2}\frac{A(y,\beta)}{D_r}\geq 0 \tag{3.1}$$

$$V\left(\hat{Q}_{y(u)}(\beta)\right)-\min.V\left(\hat{Q}_p\right)=\frac{(1-f_u)^2}{u^2}\frac{A(y,\beta)}{D_p}\geq 0 \tag{3.2}$$

$$V\left(\hat{Q}_{y(u)}(\beta)\right)-\min.V\left(\hat{Q}_l\right)=\frac{(1-f_u)^2}{u^2}\frac{A(y,\beta)}{D_l}\geq 0 \tag{3.3}$$

$$\min.V\left(\hat{Q}_{y(r)}(\beta)\right)-\min.V\left(\hat{Q}_l\right)=\frac{(1-f_u)}{u}\left(\frac{1}{m}-\frac{1}{n_1}\right)\frac{A(y,\beta)(\theta-\rho_c)^2}{D_lD_r}\geq 0 \tag{3.4}$$

$$\min.V\left(\hat{Q}_{y(p)}(\beta)\right)-\min.V\left(\hat{Q}_l\right)=\frac{(1-f_u)}{u}\left(\frac{1}{m}-\frac{1}{n_1}\right)\frac{A(y,\beta)(\theta+\rho_c)^2}{D_lD_r}\geq 0 \tag{3.5}$$

where

$$D_r=\left[\frac{1-f_u}{u}+\frac{1-f_m}{m}+\left(\frac{1}{m}-\frac{1}{n_1}\right)\theta(\theta-2\rho_c)\right],$$

$$D_p=\left[\frac{1-f_u}{u}+\frac{1-f_m}{m}+\left(\frac{1}{m}-\frac{1}{n_1}\right)\theta(\theta+2\rho_c)\right],$$

$$D_l=\left[\frac{1-f_u}{u}+\frac{1-f_m}{m}-\left(\frac{1}{m}-\frac{1}{n_1}\right)\rho_c^2\right].$$

Expressions (3.1), (3.2) and (3.3) clearly indicate the reduction in variance due to use of $\hat{Q}_r$, $\hat{Q}_p$ and $\hat{Q}_l$ respectively instead of $\hat{Q}_{y(u)}(\beta)$ as an estimator of $\hat{Q}_y(\beta)$. Further the reduction in variance by using the estimator $\hat{Q}_l$ instead of ratio-type estimator $\hat{Q}_r$ and the product-type estimator $\hat{Q}_p$ as estimators of $Q_y(\beta)$ are given by (3.4) and (3.5) respectively. From (3.1) to (3.2) we have the following inequalities

$$\min.V\left(\hat{Q}_l\right)\leq\min.V\left(\hat{Q}_r\right)\leq V\left(\hat{Q}_{y(u)}(\beta)\right) \tag{3.6}$$

and

$$\min.V\left(\hat{Q}_l\right)\leq\min.V\left(\hat{Q}_p\right)\leq V\left(\hat{Q}_{y(u)}(\beta)\right) \tag{3.7}$$

It follows from (3.6) and (3.7) that the regression estimator $\hat{Q}_l$ is better than $\hat{Q}_{y(u)}(\beta)$, $\hat{Q}_r$ and $\hat{Q}_p$.

In the case $n_1 = n$ (i.e. sample sizes are same at both the occasions) and assume that the population size N is large enough so that $f_u \approx 0$, $f_m \approx 0$, the expression in (2.11), (2.15), (2.16) and (2.17) respectively reduce to

$$V\left(\hat{Q}_{y(u)}(\beta)\right) = \frac{A(y,\beta)}{u}, \tag{3.8}$$

$$V\left(\hat{Q}_r\right) = \frac{A(y,\beta)}{mnu}\left[mn(1-2\alpha)+\alpha^2\left\{n^2+u^2\theta(\theta-2\rho_c)\right\}\right], \tag{3.9}$$

$$V\left(\hat{Q}_p\right) = \frac{A(y,\beta)}{mnu}\left[mn(1-2\delta)+\delta^2\left\{n^2+u^2\theta(\theta+2\rho_c)\right\}\right], \tag{3.10}$$

and

$$V\left(\hat{Q}_l\right) = \frac{A(y,\beta)}{mnu}\left[mn(1-2\gamma)+\gamma^2\left\{n^2-u^2\rho_c^2\right\}\right]. \tag{3.11}$$

Thus the variance expressions (3.9) to (3.11) are respectively minimized with (2.18) to (2.20) for

$$\alpha = \frac{mn}{n^2+u^2\theta(\theta-2\rho_c)} \tag{3.12}$$

$$\delta = \frac{mn}{n^2+u^2\theta(\theta+2\rho_c)} \tag{3.13}$$

and

$$\gamma = \frac{mn}{n^2-u^2\rho_c^2} \tag{3.14}$$

Thus the resulting variances of $\hat{Q}_r$, $\hat{Q}_p$ and $\hat{Q}_l$ are respectively given by

$$\min.V^*\left(\hat{Q}_r\right) = A(y,\beta)\frac{[n+u\theta(\theta-2\rho_c)]}{\left[n^2+u^2\theta(\theta-2\rho_c)\right]} \tag{3.15}$$

$$\min.V^*\left(\hat{Q}_p\right) = A(y,\beta)\frac{[n+u\theta(\theta+2\rho_c)]}{\left[n^2+u^2\theta(\theta+2\rho_c)\right]} \tag{3.16}$$

and

$$\min.V^*\left(\hat{Q}_l\right)=A(y,\beta)\frac{\left[n-u\rho_c^2\right]}{\left[n^2-u^2\rho_c^2\right]} \tag{3.17}$$

Minimization of (3.15), (3.16) and (3.17) with respect to u give the optimum value of u respectively as

$$u=\frac{n}{1+\sqrt{1+\theta(\theta-2\rho_c)}} \tag{3.18}$$

$$u=\frac{n}{1+\sqrt{1+\theta(\theta+2\rho_c)}} \tag{3.19}$$

and

$$u=\frac{n}{1+\sqrt{1-\rho_c^2}} \tag{3.20}$$

Thus the resulting values of $\min.V^*\left(\hat{Q}_r\right)$, $\min.V^*\left(\hat{Q}_p\right)$ and $\min.V^*\left(\hat{Q}_l\right)$ are respectively given by

$$\left(\min.V^*\left(\hat{Q}_r\right)\right)_{\text{opt}}=\frac{A(y,\beta)}{2n}\left[1+\sqrt{1+\theta(\theta-2\rho_c)}\right] \tag{3.21}$$

$$\left(\min.V^*\left(\hat{Q}_p\right)\right)_{\text{opt}}=\frac{A(y,\beta)}{2n}\left[1+\sqrt{1+\theta(\theta+2\rho_c)}\right] \tag{3.22}$$

and

$$\left(\min.V^*\left(\hat{Q}_l\right)\right)_{\text{opt}}=\frac{A(y,\beta)}{2n}\left[1+\sqrt{1-\rho_c^2}\right] \tag{3.23}$$

Further the variance of the direct estimator $\hat{Q}_{y(n)}(\beta)$ to the first degree of approximation is given by

$$V\left(\hat{Q}_{y(n)}^{(\beta)}\right)=\frac{(1-f_n)}{n}A(y,\beta) \tag{3.24}$$

when the population size N is very large so that $f_n\approx 0$, we get

$$V\left(\hat{Q}^{(\beta)}_{y(n)}\right)=\frac{A(y,\beta)}{n}$$

(3.25)

From (3.21), (3.22), (3.23), and (3.25) we have

$$V\left(\hat{Q}_{y(n)}(\beta)\right)-\left(\min.V^{*}\left(\hat{Q}_r\right)\right)_{\text{opt}}=\frac{A(y,\beta)}{2n}\frac{\theta(\theta-2\rho_c)}{\left\{1+\sqrt{1+\theta(\theta-2\rho_c)}\right\}}>0 \text{ if } \rho_c>\frac{\theta}{2} \tag{3.26}$$

$$V\left(\hat{Q}_{y(n)}(\beta)\right)-\left(\min.V^{*}\left(\hat{Q}_p\right)\right)_{\text{opt}}=-\frac{A(y,\beta)}{2n}\frac{\theta(\theta+2\rho_c)}{\left\{1+\sqrt{1+\theta(\theta+2\rho_c)}\right\}}>0 \text{ if } \rho_c<-\frac{\theta}{2} \tag{3.27}$$

$$\left(\min.V^{*}\left(\hat{Q}_r\right)\right)_{opt}-\left(\min.V^{*}\left(\hat{Q}_l\right)\right)_{\text{opt}}=\frac{A(y,\beta)}{2n}\frac{(\theta-\rho_c)^2}{\left\{\sqrt{1+\theta(\theta-2\rho_c)}+\sqrt{1-\rho_c^2}\right\}}\geq 0 \tag{3.28}$$

and

$$\left(\min.V^{*}\left(\hat{Q}_p\right)\right)_{\text{opt}}-\left(\min.V^{*}\left(\hat{Q}_l\right)\right)_{\text{opt}}=\frac{A(y,\beta)}{2n}\frac{(\theta+\rho_c)^2}{\left\{\sqrt{1+\theta(\theta+2\rho_c)}+\sqrt{1-\rho_c^2}\right\}}\geq 0 \tag{3.29}$$

From (3.26) to (3.29) we have the following inequalities

$$\left(\min.V^{*}\left(\hat{Q}_l\right)\right)_{\text{opt}}\leq\left(\min.V^{*}\left(\hat{Q}_r\right)\right)_{\text{opt}}\leq V\left(\hat{Q}_{y(n)}\right) \text{, when } \rho_c>\frac{\theta}{2} \tag{3.30}$$

and

$$\left(\min.V^{*}\left(\hat{Q}_l\right)\right)_{\text{opt}}\leq\left(\min.V^{*}\left(\hat{Q}_p\right)\right)_{\text{opt}}\leq V\left(\hat{Q}_{y(n)}\right)\text{, when } \rho_c<-\frac{\theta}{2} \tag{3.31}$$

Finally we conclude that the regression-type estimator $\hat{Q}_l$ has least variance and hence more efficient than $\hat{Q}_r$, $\hat{Q}_p$ and $\hat{Q}_{y(n)}$.

For $u=0$ (complete matching) or $u=n$ (no matching) the variance in (3.15), (3.16) and (3.17) reduce to:

$$\min.V^{*}\left(\hat{Q}_r\right)=\min.V^{*}\left(\hat{Q}_p\right)=\min.V\left(\hat{Q}_l\right)=V\left(\hat{Q}_{y(n)}(\beta)\right)=\frac{A(y,\beta)}{n} \tag{3.32}$$

Thus in this case all the estimators $\hat{Q}_{y(n)}(\beta)$, $\hat{Q}_r$, $\hat{Q}_p$ and $\hat{Q}_l$ are equally efficient. Further from (3.23) and (3.24), the efficiency of $\hat{Q}_l$ with respect to direct estimator $\hat{Q}_{y(n)}$ is given by

$$E = 2\left[1+\sqrt{1-\rho_c^2}\right]^{-1} \tag{3.33}$$

We also note from (3.20) that

$$\frac{u}{n} = 1-\lambda = \mu = \left(1+\sqrt{1-\rho_c^2}\right)^{-1}$$

Thus we have

$$\frac{m}{n} = \lambda = \sqrt{\left(1-\rho_c^2\right)}\left\{1+\sqrt{1-\rho_c^2}\right\}^{-1} \tag{3.34}$$

Equation (3.34) shows that the optimum percentage to be matched decreases with increasing value of ρ_c. For $\rho_c = 1$ implies $P_{xy} = 2\beta(1-\beta)$, this percentage is 0 and 50 respectively. However, for $m = 0$, b in (2.6) cannot be obtained and the results derived above are invalid. It is expected that ρ_c lies between 0.50 and 1. The percentage gain in efficiency $(E-1)100\%$ increases with increasing value of ρ_c(for optimum matching percentage)

Remark 3.1. For $\beta = 0.5$, the studies of this paper reduce to the estimation of population median in successive sampling.

4. Simulation study: Three Quartiles of abortion cases

For the purpose of simulation study we consider the situation of a population consisting of $N = 50$ states, and let y_i represent the number of abortions during 2000 and x_i be the number of abortions during 1992 in the i$^{\text{th}}$ state. Table 4.1 gives the descriptive statistics of number of abortions during 1992 and 2000.

Table 4.1. Descriptive Statistics of number of abortions during 1992 and 2000.

	Abortions 1992	*Abortions 2000*
Mean	30.6	26.3
Standard Error	7.3	6.0
Median	14.5	12
Mode	7	6
Standard Deviation	51.3	42.8
Kurtosis	18.0	13.1
Skewness	3.9	3.4
Minimum	1	1
Maximum	304	236
Count	50	50

The value of the correlation between the number of abortions during 1992 and 2000 is found to be $\rho_{xy} = 0.987$. The following graph in Fig 4.1(a) and Fig 4.1 (b) shows that distribution of number of abortions in different states is skewed towards right.

One reason of skewness may be the distribution of population in different states, that is, the states having large populations are expected to have large number of abortion cases. Thus skewness of the data indicates that the use of three quartiles may be a good measure of central locations than mean in such a situation.

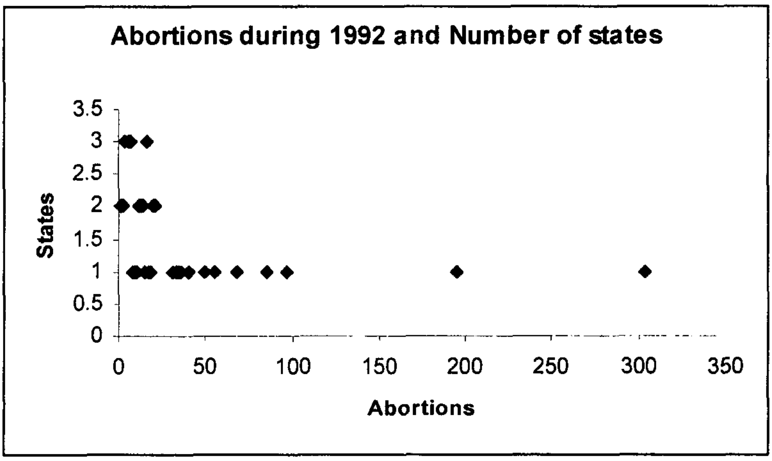

Fig. 1 (a) Abortions during 1992 versus number of states.

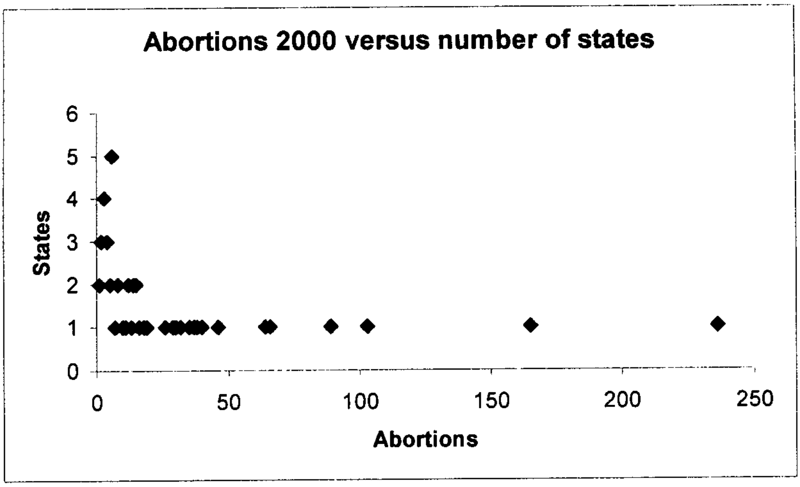

Fig. 2 (b) Abortions during 20000 versus number of states.

We selected 5000 samples of $n_1 = 20$ using without replacement sampling and only the number of abortions during 1992 among the selected states was noted. Thus $\hat{Q}_{x(n_1)}(l)_{|k}$, $l = 1, 2, 3$ and $k = 1,2,....,5000$ sample quantiles were computed. From each one of the selected 5000 samples, we decided to retain $m = 5$ states in each sample, and we selected new $u = n - m = 10 - 5 = 5$ states out of $N - n_1 = 50 - 20 = 30$ states using without replacement sampling. From the m units retained in the sample, we computed $\hat{Q}_{x(m)}(l)_{|k}$, $\hat{Q}_{y(m)}(l)_{|k}$ with $l = 1, 2, 3$ for $k = 1,2,....,5000$; and from the new unmatched units selected on the second occasion we also computed $\hat{Q}_{y(u)}(l)_{|k}$ with $l = 1, 2, 3$ for $k = 1,2,....,5000$. We decided to select a parameter Φ between 0.1 and 0.9 with a step of 0.1. Then the relative efficiencies of the ratio type estimators, for $l = 1, 2, 3$,

$$\hat{Q}_R(l)|_k = \Phi\ \hat{\mathrm{Q}}_{y(m)}(\mathrm{l})|_\mathrm{k} \left\{ \frac{\hat{Q}_{x(n_1)}(l)|_k}{\hat{Q}_{x(m)}(l)|_k} \right\} + (1 - \Phi)\hat{Q}_{y(u)}(l)|_k, \tag{3.1}$$

with respect to $\hat{Q}_{y(u)}(l)_{|k}$ are given by

$$\mathrm{RE}(l)=\frac{\sum_{k=1}^{5000}\left[\hat{Q}_{y(u)}(l)|_k-Q_y(l)\right]^2}{\sum_{k=1}^{5000}\left[\hat{Q}_R(l)|_l-Q_y(l)\right]^2}\times 100 \text{ for } l=1,2,3 \qquad (3.2)$$

where $Q_y(l)$ for $l=1,2,3$ denotes the l^{th} population quartile.

Table 4.1. Relative efficiency of the ratio type estimators.

Φ	$RE(1)$	$RE(2)$	$RE(3)$
0.1	111.90	110.39	114.23
0.2	124.63	121.02	131.34
0.3	135.82	131.68	150.78
0.4	146.90	141.36	172.22
0.5	155.15	148.57	197.33
0.6	159.32	156.02	223.28
0.7	160.05	157.96	249.17
0.8	158.30	159.49	272.42
0.9	146.72	154.92	281.85

The relative efficiency of the ratio type estimator of the first quartile $Q_y(1)$ ranges from 111% to 160% with median efficiency being 146%; the relative efficiency of the estimator of second quartile $Q_y(2)$ ranges from 110% to 159% with median efficiency being 148%; and the relative efficiency of the estimator of third quartile $Q_y(3)$ ranges from 114% to 281% with median efficiency of 197%. It is interesting to note that the relative efficiency of the first and second quartiles behaves in the same fashion that as the value of Φ increases from 0.1 to 0.9 the relative efficiency increases at first and then starts decreasing, whereas the relative efficiency of the estimator of the third quartile goes on increasing. It is not obvious to find its reason, but one reason may be that data is skewed to right, and the ratio type adjustment may be making more sense than the simple sample quartile estimator. We acknowledge that more simulation may be performed in future studies as pointed out by one of the learned referees.

Conclusion

To our knowledge, this is a first attempt to estimate finite population quartiles using successive sampling. The analytical and empirical results support the fact that estimation of three quartiles using successive sampling is feasible, which was ignored by the survey statisticians in the past.

Acknowledgements

The authors are thankful to the Editor, Associate Editor and the two referees for the valuable comments on the original version of this manuscript which lead to substantial improvement. The authors also acknowledge the use of free access to the data by the Statistical Abstracts of the United States 2006 used here to evaluate one methodology over the other.

References

Allen, J., Singh, H.P., Singh, S. and Smarandache, F. (2002). A general class of estimators of population median using two auxiliary variables in double sampling. *INTERSTAT.*

Biradar, R.S. and Singh, H.P. (2001). Successive sampling using auxiliary information on both the occasions. *Calcutta Statistical Association Bulletin*, 51, 243-251.

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Expt. Stat. Res. Bull.*, No. 304.

Kuk, A.Y.C. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika*, 80, 385-392.

Kuk, A.Y.C. and Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *J. Royal Statist. Soc.*, B, 51, 261—269.

Mak, T.K. and Kuk, A.Y.C. (1993). A new method for estimating finite population quantiles using auxiliary information. *Canadian J. Statist.*, 21, 29-38.

Randles, R.H. (1982). On the asymptotic normality of Statistics with estimated parameters. *Annals of Statistics*, 10, 462-474.

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.

Rueda Garcia, M., Arcos Cebrian, A. and Artes Rodriguez, E. (1998). Quantile interval estimation in finite population using a multivariate ratio estimator. *Metrika*, 47, 203-213.

Rueda Garcia A. and Arcos Cebrian, A. (2001). On estimating the median from survey data using multiple auxiliary information. *Metrika*, 54, 59-76.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall, London.

Singh, H.P., Singh, S. and Puertas, M.S. (2003). Ratio type estimators for the median of finite populations. *Allgemeines Statistisches Archiv*, 87, 369-382.

Singh, S. (2003). *Advanced sampling theory with applications: How Michael 'Selected' Amy.* Kluwer Academic Publishers, The Netherlands.

Singh, S. and Joarder, A.H. (2002). Estimation of the distribution function and median in two phase sampling. *Pakistan J. Statist.*, 18(2), 301-319.

Singh, S., Joarder, A.H., and Trcay, D.S. (2001). Median estimation using double sampling. *Australian and NewZealand J. Statist.*, 43(1), 33-46.