ESTIMATION OF THE DISTRIBUTION FUNCTION WITH A COMPLEX SURVEY

Carol A. Francisco and Wayne A. Fuller, Iowa State University

Abstract

Estimation of the cumulative distribution function and related statistics, such as the median and interquartile range, is considered. Large sample properties of estimators constructed from stratified cluster samples are presented. The PC CARP computer algorithm is discussed.

1. Introduction

There is an extensive literature on quantile estimation, most of it for simple random sampling. Extension of results derived under an assumption of simple random sampling from an absolutely continuous distribution to the complex sampling designs used in finite population sampling has met with limited success. Woodruff (1952) proposed using a weighted sample median to estimate the population median, where the weight assigned to each observation is proportional to the inverse of its selection probability. Using the approach taken by Maritz and Jarrett (1978), Gross (1980) derived a small-sample estimator of the variance of the weighted sample median estimator for stratified sampling without replacement from a finite population.

A number of authors have investigated model-free procedures for constructing exact $100(1 - \alpha)$ percent confidence intervals for quantiles in finite populations. Inferences from the sample to the finite population are based upon confidence intervals which take into account the sampling scheme. Thompson (1936), Wilks (1962), and Konijn (1973) have given design-based confidence intervals for the sample median when simple random sampling from a finite population is assumed. Meyer (1972) and Sedransk and Meyer (1978) investigated three exact confidence interval procedures for quantiles when sampling is from a stratified population.

For more than two strata, the confidence interval procedures proposed by Meyer (1972) and Sedransk and Meyer (1978) become very complex. Some authors have determined lower bounds for confidence coefficients as a means of dealing with this problem (McCarty 1965, Smith and Sedransk 1983). Other approaches to inference from the sample to the finite population are based on information about the distribution of values of the characteristic under study in the finite population (Ericson 1969, Binder 1982, Chambers and Dunstan 1986).

In this paper, a theoretical basis for the confidence set procedure proposed by Woodruff (1952) will be developed. In Section 2, an estimator of the population distribution function will be given and used to define a weighted quantile estimator. Large sample confidence sets for the p-th quantile and the interquartile range are given in Sections 3.1 and 3.2 for single-stage stratified cluster sampling. Incorporation of the proposed procedures into the PC CARP survey data analysis computer program is described in Section 4. Results of Monte Carlo studies also are given.

2. Estimation Procedures

2.1 Empirical Distribution Function

Let $U(N) = \{u(hij): h = 1, \ldots, L, i = 1, \ldots, N(h), j = 1, \ldots, M(hi)\}$ be a finite population which is divided into L strata. Let

$$N = \sum_{h=1}^{L} N_h \text{ , and } M = \sum_{h=1}^{L} \sum_{i=1}^{N_h} M_{hi} \text{ ,}$$

where N is the total number of clusters (primary sampling units), $N(h)$ is the number of clusters in stratum h, M is the number of elemental units in $U(N)$, and $M(hi)$ is the number of elemental units in cluster i of stratum h. Let $Y(hij)$ be the value of a characteristic Y associated with the j-th elemental unit in the i-th cluster of the h-th stratum.

Define the finite population distribution function for Y by

$$F_N(x) = M^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} I\{Y_{hij} \leq x\} \text{ ,} \qquad (2.1)$$

where

$$I\{Y_{hij} \leq x\} = 1 \quad \text{if} \quad Y_{hij} \leq x$$

$$= 0 \quad \text{otherwise .}$$

Alternatively, $F_N(x)$ can be defined using the distribution function of Y in each stratum:

$$F_N(x) = M^{-1} \sum_{h=1}^{L} M_h F_{N_h}(x) \text{ ,} \qquad (2.2)$$

where, for $h = 1, \ldots, L$,

$$F_{N_h}(x) = M_h^{-1} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} I\{Y_{hij} \leq x\} \text{ ,}$$

and $M_h = \sum_{i=1}^{N_h} M_{hi}$.

Suppose that a sample of n clusters is selected from $U(N)$ and that sampling within each stratum is carried out independently. Let

$$n = \sum_{h=1}^{L} n_h \text{ , and } \sum_{i=1}^{N_h} \pi_{hi} = n_h \text{ ,}$$

where n is the total sample of n primary sampling units, $n(h) \geq 2$ is the number of clusters selected in the h-th stratum, and $\pi(hi) > 0$ is the probability of including cluster i of stratum h in the sample

[h = 1, ..., L; i = 1, ..., N(h)] . It will be assumed that all elemental units within selected clusters are included in the sample. Results for single-stage cluster sampling are easily extended to various forms of subsampling within clusters.

A general estimator of the cumulative distribution function for stratified two-stage cluster sampling is

$$F_n(x) = \hat{M}^{-1}\left[\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} I\{y_{hij} \le x\}\right] , \tag{2.3}$$

where

$$\hat{M} = \left[\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\right] ,$$

w(hij) is the sampling weight, m(hi) is the number of elemental units subsampled in cluster hi , and y(hij) is the value of the characteristic Y associated with the j-th elemental unit in the i-th sampled cluster of stratum h . For single-stage cluster sampling, m(hi) = M(hi) , and w(hij) = w(hi) for j = 1, ..., M(hi) . Expression (2.3) reduces to

$$F_n(x) = \hat{M}^{-1}\left[\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{M_{hi}} w_{hi} I\{y_{hij} \le x\}\right] , \tag{2.4}$$

where

$$\hat{M} = \left[\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi} M_{hi}\right] .$$

2.2 Quantiles and the Interquartile Range

The p-th quantile of Y in the finite population U(N) is defined as

$$q_N(p) = \inf\{x: F_N(x) \ge p\} \tag{2.5}$$

for $0 < p < 1$. A measure of dispersion is the interquartile range

$$R_N = q_N(0.75) - q_N(0.25) . \tag{2.6}$$

The usual estimator of $q_N(p)$ is the p-th sample quantile

$$\hat{q}_n(p) = \inf\{x: F_n(x) \ge p\} . \tag{2.7}$$

An estimator of R(N) is given by

$$\hat{R}_n = \hat{q}_n(0.75) - \hat{q}_n(0.25) . \tag{2.8}$$

3. Large Sample Properties of Estimators and Confidence Intervals

3.1 Quantiles

A framework for developing the asymptotic theory to support the large-sample procedures proposed by Woodruff (1952) is first established. Assumptions similar to those made by Fuller (1975, 1984) or by Krewski and Rao (1981) provide a basis for such a development. In both cases sequences of finite populations and samples which meet certain regularity conditions are defined. Also see Bickel and Freedman (1984).

Let $\{\xi(r): r=1, \ldots, \infty\}$ be a sequence of stratified finite populations, each having $L(r) \ge L(r-1)$ strata. Suppose the finite population in stratum h of $\xi(r)$ is a random sample of size $N(rh) \ge N(r-1,h)$ clusters selected from an infinite superpopulation. Associated with the j-th element in the i-th cluster of stratum h is a column vector of characteristics:

$$\underset{\sim}{Y}_{rhij} = (Y_{rhij1}, \ldots, Y_{rhijk})'$$

for h = 1, ..., L(r), i = 1, ..., N(rh), and j = 1, ..., M(rhi). It is assumed that the cluster totals have absolute $2 + \delta$ moments $(\delta > 0)$ which are uniformly bounded by $B(\delta) < \infty$. The cluster totals in the (rh)-th superpopulation have mean vector $\underset{\sim}{\mu}(rh) = [\mu(rh1), \ldots, \mu(rhk)]'$ and covariance matrix $\underset{\sim}{\Sigma}(rh)$, where the diagonal elements of $\underset{\sim}{\Sigma}(rh)$ are uniformly bounded below.

Let a simple random sample of n(rh) clusters $[n(rh) \ge 2, n(rh) \ge n(r-1,h)]$ be selected without replacement from the (rh)-th finite population. The vector of characteristics associated with the j-th element in the i-th selected cluster of stratum h is given by

$$\underset{\sim}{y}_{rhij} = (y_{rhij1}, y_{rhij2}, \ldots, y_{rhijk})'$$

for h = 1, ..., L(r) , i = 1, ..., n(rh) , and j = 1, ..., M(rhi) . For the r-th population, let

$$\bar{\underset{\sim}{y}}_{rn} = \sum_{h=1}^{L_r} W_{rh} n_{rh}^{-1} \sum_{i=1}^{n_{rh}}\sum_{j=1}^{M_{rhi}} \underset{\sim}{y}_{rhij} , \tag{3.1}$$

$$\bar{\underset{\sim}{Y}}_{rN} = \sum_{h=1}^{L_r} W_{rh} N_{rh}^{-1} \sum_{i=1}^{N_{rh}}\sum_{j=1}^{M_{rhi}} \underset{\sim}{Y}_{rhij} , \tag{3.2}$$

$$\underset{\sim}{\mu}_r = \sum_{h=1}^{L_r} W_{rh}\underset{\sim}{\mu}_{rh} , \text{ and } W_{rh} = N_{rh}N_r^{-1} . \tag{3.3}$$

Here $\bar{\underset{\sim}{y}}(rn)$ is the sample mean per cluster, $\bar{\underset{\sim}{Y}}(rN)$ is the finite population parameter, and $\underset{\sim}{\mu}(r)$ is the weighted superpopulation mean per cluster.

The asymptotic properties of $\bar{y}(rn)$ will be examined under the following regularity conditions.

C1. $f_{rh} \leq B_f < 1$, where $f_{rh} = N_{rh}^{-1} n_{rh}$.

C2. $\sup_{1 \leq h \leq L_r} n_r W_{rh}^2 n_{rh}^{-2} \to 0$ as $r \to \infty$.

C3. For all r ,

$$0 < B_{SL} < \left| n_r \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \Sigma_{rh} \right| < B_{SU} < \infty ,$$

where B_{SL} and B_{SU} are fixed numbers.

The following central limit theorem can be established.

Theorem 1. Let the sequence of finite populations and samples be as described. Under regularity conditions C1 - C3, as $r \to \infty$,

$$[\hat{\mathbf{V}}\{\bar{\mathbf{y}}_{rn} - \bar{\mathbf{Y}}_{rN}\}]^{-1/2} (\bar{\mathbf{y}}_{rn} - \bar{\mathbf{Y}}_{rN}) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}) ,$$

$$[\hat{\mathbf{V}}\{\bar{\mathbf{y}}_{rn} - \mu_r\}]^{-1/2} (\bar{\mathbf{y}}_{rn} - \mu_r) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}) ,$$

where

$$\hat{\mathbf{V}}\{\bar{\mathbf{y}}_{rn} - \bar{\mathbf{Y}}_{rN}\} = \sum_{h=1}^{L_r} W_{rh}^2 (1 - f_{rh}) n_{rh}^{-1} \hat{\Sigma}_{rh} ,$$

$$\hat{\mathbf{V}}\{\bar{\mathbf{y}}_{rn} - \mu_r\} = \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \hat{\Sigma}_{rh} ,$$

$$\hat{\Sigma}_{rh} = (n_{rh} - 1)^{-1} \sum_{i=1}^{n_{rh}} \mathbf{a}_{rhi.} \mathbf{a}'_{rhi.} ,$$

$$\bar{\mathbf{y}}_{rh..} = n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \mathbf{y}_{rhi.} ,$$

$$\mathbf{y}_{rhi.} = \sum_{j=1}^{M_{rhi}} \mathbf{y}_{rhij} ,$$

$$\mathbf{a}_{rhi.} = (\mathbf{y}_{rhi.} - \bar{\mathbf{y}}_{rh..}) . \qquad \square$$

For stratified single-stage cluster sampling, the estimator of the distribution function which is given in (2.4) is a ratio of quantities of the form shown in (3.1). The estimator of the cumulative distribution function for $\xi(r)$ is

$$F_{rn}(x) = \hat{M}^{-1} \left[\sum_{h=1}^{L_r} W_{rh} n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \sum_{j=1}^{M_{rhi}} I\{y_{rhij} \leq x\} \right] ,$$

where

$$\hat{M} = \left[\sum_{h=1}^{L_r} W_{rh} n_{rh}^{-1} \sum_{i=1}^{n_{rh}} M_{rhi} \right] .$$

Additional regularity conditions are needed for some of the asymptotic results for the estimated cumulative distribution function. We begin by assuming that a common overall superpopulation distribution function exists for all $\xi(r)$ and is given by $F(x)$. That is, we assume for all r

$$F(x) = E_{\xi_r}\{F_{rN}(x)\} \tag{3.4}$$

$$= M_r^{-1} \sum_{h=1}^{L_r} \sum_{i=1}^{N_{rh}} \sum_{j=1}^{M_{rhi}} E_{\xi_{rh}}[I\{Y_{rhij} \leq x\}]$$

$$= M_r^{-1} \sum_{h=1}^{L_r} M_{rh} F_{rh}(x) ,$$

where the subscript $\xi(rh)$ on the expectation operator denotes expectation under the superpopulation model for the h-th stratum of the r-th finite population in the sequence, and the subscript rh identifies the superpopulation distribution function for stratum h of population $\xi(r)$.

Theorem 2. Let the sequence of populations and samples be as described. Let regularity conditions C1 - C3 hold, and let $F(x)$ satisfy (3.4). Then, for fixed x in the support of $F(x)$, as $r \to \infty$,

$$[\hat{V}\{F_{rn}(x)\}]^{-1/2} [F_{rn}(x) - F(x)] \xrightarrow{L} N(0, 1) ,$$

where

$$\hat{V}\{F_{rn}(x)\} = \sum_{h=1}^{L_r} (n_{rh} - 1)^{-1} n_{rh} \sum_{i=1}^{n_{rh}} a_{rhi.}^2 ,$$

$$z_{rhij} = 1 \quad \text{if} \quad y_{rhij} \leq x$$

$$= 0 \quad \text{otherwise} ,$$

$$d_{rhi.} = \hat{M}_r^{-1} N_{rh} n_{rh}^{-1} [z_{rhi.} - M_{rhi} F_{rn}(x)] ,$$

$$\bar{d}_{rh..} = n_{rh}^{-1} \sum_{j=1}^{n_{rh}} d_{rhi.} ,$$

$$\hat{M}_r = \sum_{h=1}^{L_r} \sum_{i=1}^{n_{rh}} N_{rh} n_{rh}^{-1} M_{rhi} ,$$

$$z_{rhi.} = \sum_{j=1}^{M_{rhi}} z_{rhij} ,$$

$$a_{rhi.} = d_{rhi.} - \bar{d}_{rhi..} \cdot \qquad \square$$

The estimated variance of Theorem 2 is a variance estimator for a combined ratio estimator of the mean per element. It is a Taylor series estimator of the variance of the approximate distribution. A confidence set procedure for quantiles is defined in Corollary 1.

Corollary 1. Let the assumptions of Theorem 2 hold for $x(\gamma)$. In addition, assume that for fixed x in the support of $F(x)$:

C4. the cumulative distribution function, $F(x)$, is continuous and has a continuous, positive derivative in a neighborhood of x .

Let $\Gamma(\gamma)$ be the set of x for which

$$F_{rn}(x) + z_{\alpha/2}[\hat{V}\{F_{rn}(x)\}]^{1/2} > \gamma$$

and

$$F_{rn}(x) - z_{\alpha/2}[\hat{V}\{F_{rn}(x)\}]^{1/2} < \gamma ,$$

where $z(\alpha/2)$ is defined by $\Phi[z(\alpha/2)] = 1 - \alpha/2$ and $\Phi(\cdot)$ is the distribution function of a standard normal random variable. Let $x(\gamma)$ be such that $F[x(\gamma)] = \gamma$. Then, as $r \to \infty$,

$$P\{x_\gamma \in \Gamma_\gamma\} \to \alpha . \qquad \square$$

3.2 Interquartile Range

While Theorem 2 and its corollary provide a method for constructing a confidence set for a given quantile, additional results are needed in order to justify the confidence set procedure proposed by Woodruff (1952) and to construct confidence sets for functions of quantiles such as the interquartile range.

Let the sequence of populations, $\{\xi(r): r=1, \ldots, \infty\}$, and samples be as described in Section 3.1. Let

$$q(\gamma) = F^{-1}(\gamma)$$

be the quantile function. A set of k fixed, distinct quantiles (i.e., $0 < \gamma(i) < 1$, $0 < \gamma(j) < 1$, $\gamma(i) \neq \gamma(j)$, for $i \neq j$, $i = 1, \ldots, k$, $j = 1, \ldots, k$) is given by

$$[q(\gamma_1), q(\gamma_2), \ldots, q(\gamma_k)] = (x_1, x_2, \ldots, x_k) = \mathbf{x} .$$

The corresponding set of sample quantiles for the r-th sample in the sequence is denoted by

$$[\hat{q}_{rn}(\gamma_1), \hat{q}_{rn}(\gamma_2), \ldots, \hat{q}_{rn}(\gamma_k)] = (\hat{x}_{r1}, \ldots, \hat{x}_{rk}) = \hat{\mathbf{x}}_r .$$

Let $\hat{\underset{\sim}{\Omega}}(r)$ be the estimated covariance matrix of

$$[F_{rn}(\hat{x}_{r1}), F_{rn}(\hat{x}_{r2}), \ldots, F_{rn}(\hat{x}_{rk})] ,$$

where the notation means that the estimated variance is evaluated at the estimated quantile.

Four additional regularity conditions are used to establish the asymptotic normality of the estimated quantiles.

C5. For fixed x in the support of $F(x)$, $n_r V\{F_{rn}(x)\}$ is continuous in x .

C6. $V\{F_{rn}(x + \delta_{rn}) - F_{rn}(x)\} = O(n_r^{-1}\delta_{rn})$,

for all x and $x + \delta(rn)$ in the support of $F(x)$, where $\delta(rn) > 0$ and $\delta(rn) \to 0$ as $r \to \infty$.

C7. Let $\hat{\underset{\sim}{\Omega}}(r)$ be the $k \times k$ estimated covariance matrix of

$$[F_{rn}(x_1), F_{rn}(x_2), \ldots, F_{rn}(x_k)] ,$$

where $[x(1), \ldots, x(k)]$ are any k fixed distinct points in the support of $F(x)$. Let $\underset{\sim}{\Omega}$ be the corresponding true covariance matrix, where $\underset{\sim}{\Omega}$ is positive definite. Assume

$$n_r(\hat{\underset{\sim}{\Omega}}_r - \underset{\sim}{\Omega}) = O_p(n_r^{-1/2}) .$$

C8. The distribution function, F , has a continuous second derivative at the points of interest, $x(1), \ldots, x(k)$.

The p-th sample quantile, $0 < p < 1$, can be expressed asymptotically as a linear transformation of the empirical distribution function evaluated at $q(p)$. This expression is called the Bahadur representation in the literature on order statistics (Bahadur 1966).

The following Lemma is an extension of a weak version of Bahadur's result to single-stage cluster sampling. The method of proof parallels that used by Ghosh (1971) to establish the result for simple random sampling.

Lemma 3. Let $0 < p < 1$. Under assumptions C1, C2, C3, C4, and C6, the sample quantile can be represented as

$$\hat{q}_{rn}(p) = q(p) - [f(q(p))]^{-1}[F_{rn}(q(p)) - F(q(p))] + R^*_{rn} ,$$

where $R^*_{rn} = o_p(n_r^{-1/2})$. □

The asymptotic representation of the estimated quantile given in Lemma 3 is used in the following theorem to prove the asymptotic multivariate normality of the estimator.

Theorem 3. Let assumptions C1 through C8 hold. Let

$$\hat{\mathbf{D}}_r = \mathrm{diag}(\hat{d}_{r1}, \ldots, \hat{d}_{rk}) ,$$

where, for $1 \leq i \leq k$,

$$\hat{d}_{ri} = \frac{\hat{q}_{rn}(\gamma_{Ui}) - \hat{q}_{rn}(\gamma_{Li})}{2t_\alpha[\hat{V}\{F_{rn}(\hat{x}_{ri})\}]^{1/2}} ,$$

$$\hat{q}_{rn}(\gamma_{Ui}) = \hat{q}_{rn}(\gamma_i + t_\alpha[\hat{V}\{F_{rn}(\hat{x}_{ri})\}]) ,$$

$$\hat{q}_{rn}(\gamma_{Li}) = \hat{q}_{rn}(\gamma_i - t_\alpha[\hat{V}\{F_{rn}(\hat{x}_{ri})\}]) ,$$

and $t(\alpha)$ is the tabular value such that $P\{|Z| > t(\alpha)\} = \alpha$ for Z a $N(0, 1)$ random variable and $0 < \alpha < 1$. It is understood that the estimated quantile is the smallest observed x if $\gamma(Li)$ is negative and that the estimated quantile is the largest observed x if $\gamma(Ui)$ is greater than one. Then

$$[\hat{\mathbf{D}}_r \hat{\underset{\sim}{\Omega}}_r \hat{\mathbf{D}}_r]^{-1/2} (\hat{\mathbf{x}}_r - \mathbf{x})' \xrightarrow{L} N(\mathbf{0}, \mathbf{I})$$

as $r \to \infty$. □

The proof of Theorem 3 provides a justification for the confidence interval procedure of Woodruff (1952), because

$$[\hat{q}_{rn}(\gamma_{Li}), \hat{q}_{rn}(\gamma_{Ui})] ,$$

is the interval proposed by Woodruff.

The asymptotic distribution theory of Theorem 3 also provides procedures for estimating the standard error of the estimated interquartile range, in large-scale surveys.

4. Implementation of the Procedures

4.1 PC CARP Computer Algorithm

Iowa State University, in a joint undertaking with the International Statistical Programs Center of the U.S. Census Bureau, is currently developing a computer program which will analyze data from one-stage or two-stage stratified cluster samples. The program is designed for use with the IBM Personal Computer XT or AT. The program can be used to compute estimators and estimated variances for the overall population, for individual strata, and for subpopulations defined by classification variables. Taylor approximations for the variances of the approximate distributions of statistics are used. Schnell, et al (1986) gives a detailed description of program capabilities, available analyses, program structure, and the user interface.

The "Quantiles" option of PC CARP provides the following statistics for a user-specified variable and subpopulation:

(1) number of sampled elements in subpopulation;

(2) estimated subpopulation mean, its standard error, coefficient of variation, and the design effect for the estimated mean;

(3) estimated subpopulation variance and coefficient of variation for the subpopulation;

(4) three smallest and three largest values of observations, number of observations at these values, and a selected element identifier and sample weight for each value;

(5) estimated cumulative distribution function and standard error at 25 selected points;

(6) estimates of selected quantiles and standard errors;

(7) estimated interquartile range and its standard error.

4.2 Monte Carlo Simulations

A series of three Monte Carlo simulation studies were performed to evaluate the performance of the "Quantiles" option within PC CARP. Samples were selected from three different populations for which quantile estimation is of interest. All populations were skewed to the right.

The population selected for the first study consisted of 3,069 counties in the United States (excluding Alaska). The survey variable of interest was urban land area expressed as a percent of total land area within the county. Information on this variable was available from the 1982 National Resources Inventory, a survey conducted jointly by the Soil Conservation Service and the Statistical Laboratory at Iowa State University. A description of the inventory can be found in Goebel and Baker (1983) and Goebel and Schmude (1981). Table 1 gives summary statistics for this urban population. The percent of urban land within a county ranged from 0 to 88 percent, with 75 percent of the counties having less than 3.4 percent urban land.

Two sets of 500 simple random samples were drawn from the urban land population. The sample sizes were 50 and 100. Sample sizes were selected so that the performance of the estimators could be examined when n is moderate.

Table 1. Characteristics of the urban and soil loss populations.

Finite Population Characteristic	Urban Population	Soil Loss Population
SIZE		
Number Strata	1	35
Number Clusters	3,069	3,090
Number Elements Per Cluster	1	1-3
Total Number Elements	3,069	8,516
SUMMARY STATISTICS		
Minimum	0.0	0.00
Maximum	88.1	1,001.52
Mean	3.9	2.42
Standard Deviation	8.5	14.40
QUANTILES		
$q_N(0.25)$	0.6	0.02
$q_N(0.50)$	1.3	0.28
$q_N(0.75)$	3.4	1.68
Interquartile Range	2.8	1.66

The following statistics were computed for each sample: 25-th, 50-th, and 75-th quantiles, respectively, the interquartile range and variance estimates for each estimator. The confidence interval procedure, based on Theorem 3, was used to determine 95 percent confidence intervals for the interquartile range. Three different procedures for calculating 95 percent confidence intervals for quantiles were used. The first method for computing confidence intervals was the large-sample test inversion procedure of Corollary 1. PC CARP uses a smoothed version of the procedure given by Corollary 1 in which the bounds are restricted to be monotone nondecreasing. The second procedure was a large-sample symmetric interval calculated as the estimated quantile plus or minus two times its estimated standard deviation. Finally, a confidence interval was formed by using order statistics as endpoints of the interval. The order statistics were selected such that the confidence interval was of minimum length and the confidence coefficient was approximately 95 percent. Since sample sizes precluded the calculation of confidence coefficients based on the hypergeometric distribution, binomial approximations were used.

To summarize the results of the 500 repetitions, the averages and the variances of the 500 observed values of the estimated quantiles for p-values of 0.25, 0.50, and 0.75 were computed. The average of the 500 estimates of the variance of the estimated quantiles also was computed. For p-values of 0.25, 0.50 and 0.75, the estimated quantiles have nearly no bias for $n \geq 50$. The effect of increasing the sample size from 50 to 100 is to decrease the estimated variance by a factor of about two. The quantile estimator for $p = 0.75$ has larger variance than that for $p = 0.25$, due to the positive skew present in this population.

Estimates of the probabilities that the confidence intervals contain the true value and the estimated expected length of the intervals were computed. With 500 replicates, the estimated coverage probabilities have standard errors of approximately 0.01. In almost all cases the estimated coverage probabilities are within 1.5 standard errors of the nominal level of 95 percent. The lengths of the intervals for the three procedures are quite comparable.

Soil erosion data from a quality evaluation study of the 1982 National Resources Inventory were used as the basis for the second experimental population. The 1982 National Resources Inventory employed a stratified area sampling scheme to collect soil erosion data. The design of the quality evaluation study has been described in detail by Francisco (1986).

Data from 3,090 primary sampling units included in the quality evaluation study were used to form a stratified population. Table 1 gives summary statistics for the soil loss population. The population had 35 geographical strata, which ranged in size from 59 to 235 primary sampling units. For purposes of the study, primary sampling units were assumed to be clusters of one acre plots on which soil erosion data were collected. Based on the data collected at each plot, an estimate of the erosion rate (tons/acre/year) was made for each plot. Cluster sizes ranged from one to three plots, with an average of 2.8 plots per primary sampling unit.

Five hundred stratified random samples of size 100 clusters were selected from the population. The size of the simple random sample within each stratum was approximately proportional to the size of the stratum and ranged from 2 to 7.

Results of the simulation study for this population are comparable to the results for the urban population.

The population used for the third study had ten strata with strata sizes [N(h) , h=1, ..., 10] as shown in Table 2. The observations in the strata were generated as simple random samples from 10 lognormal superpopulation distributions. Table 2 lists the parameters for each stratum. The cumulative distribution function is a mixed lognormal distribution with weights given by the stratum weights.

A series of 1,000 finite populations of size 500 were selected from the superpopulation. One stratified random sample of size 100 was selected from each population. The sample was allocated equally among the 10 strata. This means that the sampling rate varied from one-in-three to one-in-seven.

This design permits inferences either for the infinite superpopulation or for the finite population. If inference is from the sample to the finite population, then finite population correction factors are used in variance calculations. If inference is from the sample to the superpopulation, then finite population correction factors are not used in variance calculations.

Quantile averages for p-values of 0.25 , 0.50, and 0.75 are within one standard error of the respective superpopulation values. As with the other two populations, the quantile estimator for p-values of 0.25, 0.50, and 0.75 displays near zero bias for this population. The Monte Carlo estimate of the variance of the estimated quantile is an acceptable estimator of the variance for this population, also displaying little bias. Both the observed and the estimated variance of the 0.75-quantile are larger than that of the 0.25-quantile due to the positive skew in the population. Coverage probabilities for the test inversion and the symmetric confidence interval procedures were comparable, and the obtained confidence coefficients are near the nominal level of 95 percent. See Table 3.

In summary, for the three populations investigated in this study, use of either the symmetric or the test inversion confidence interval procedures leads to confidence intervals with actual confidence coefficients acceptably close to the 95 percent nominal level.

ACKNOWLEDGEMENTS

This research was partly supported through Joint Statistical Agreement JSA 86-2 with the U.S. Bureau of the Census. The portion of PC CARP used in this study was written by H. J. Park and Sharon Loubert. We thank H. J. Park, Stephen Miller, and Sharon Loubert for assistance with the Monte Carlo studies.

Table 2. Superpopulation parameters, stratum sizes, and sample sizes for the lognormal population.

Stratum Number	Superpopulation Mean	Superpopulation Standard Deviation	Finite Population Size	Sample Size
1	4.69	1.44	40	10
2	8.00	3.33	40	10
3	8.85	3.68	50	10
4	24.05	15.83	50	10
5	13.80	7.36	60	10
6	6.55	2.73	60	10
7	5.18	1.59	70	10
8	6.55	2.73	50	10
9	24.05	15.83	50	10
10	61.56	58.29	30	10
Population Values	14.23	21.36	500	

Table 3. Monte Carlo coverage probabilities of 95 percent confidence intervals in 1,000 stratified random samples of size 100 from the lognormal population.

Parameter	Coverage Probability		Average Length
	Test Inversion Procedure	Symmetric Procedure	
Finite Population			
$q_N(0.25)$	0.955	0.941	1.14
$q_N(0.50)$	0.964	0.947	1.95
$q_N(0.75)$	0.958	0.942	5.56
Interquartile Range		0.943	5.77
Superpopulation			
q(0.25)	0.963	0.950	1.26
q(0.50	0.966	0.946	2.17
q(0.75)	0.953	0.944	6.22
Interquartile Range		0.950	6.46

REFERENCES

Bahadur, R. R. (1966). A note on quantiles in large samples. Ann. Math. Statist., **37**, 577-580.

Binder, D. A. (1982). Non-parametric Bayesian models for samples from finite populations. J. R. Statist. Soc., Ser. B, **44**, 388-393.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in statistical sampling. Ann. Statist., **12**, 470-482.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from data. Biometrika, [in press].

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. J. R. Statist. Soc., Ser. B, **31**, 195-233.

Francisco, C. A. (1986). A Quality Evaluation Study of the 1982 National Resources Inventory. Statistical Laboratory, Iowa State University, Ames, Iowa.

Fuller, W. A. (1975). Regression analysis for sample survey. Sankhya, Ser. C., **37**, 117-132.

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. Survey Meth., **10**, 97-118.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. Ann. Math. Statist., **42**, 1957-1961.

Goebel, J. J. and Baker, H. D. (1983). The 1982 National Resources Inventory Sample Design and Estimation Procedures. Statistical Laboratory, Iowa State University, Ames, Iowa.

Goebel, J. J. and Schmude, K. O. (1981). Planning the SCS National Resources Inventory. Arid Land Resources Inventories: Developing Cost Efficient Methods, Forest Service General Technical Report WO-28, U.S. Department of Agriculture, 148-153.

Gross, S. T. (1980). Median estimation in sample surveys. Proc. Sec. Survey Res. Methods, Amer. Statist. Assoc., Washington, D. C., 181-184.

Konijn, H. S. (1973). Statistical Theory of Sample Survey Design and Analysis. North-Holland Publishing Company, Inc., New York.

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. Ann. Statist., **9**, 1010-1019.

Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the vari-ance of the sample median. J. Amer. Statist. Assoc., **73**, 194-196.

McCarthy, P. J. (1965). Stratified sampling and distribution-free confidence intervals for a median. J. Amer. Statist. Assoc., **60**, 772-783.

Meyer, J. S. (1972). Confidence intervals for quantiles in stratified random sampling. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

Schnell, D., Sullivan, G., Kennedy, W. J. and Fuller, W. A. (1986). PC CARP: Variance estimation for complex surveys. Paper presented at Computer Science and Statistics: 18th Symposium on the Interface, Fort Collins, Co., March 19-21, 1986.

Sedransk, J. and Meyer, J. (1978). Confidence intervals for the quantiles of a finite population: Simple random and stratified simple random sampling. J. R. Statist. Soc., Ser. B, **40**, 239-252.

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

Thompson, W. R. (1936). On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. Ann. Math. Statist., **7**, 122-128.

Wilks, S. S. (1962). Mathematical Statistics. John Wiley & Sons, Inc., New York.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. J. Amer. Statist. Assoc., **47**, 635-646.