RESULTS OF A SIMULATION FOR COMPARING TWO METHODS FOR ESTIMATING QUANTILES AND THEIR VARIANCES FOR DATA FROM A SAMPLE SURVEY

Sara C. Wheeless and Babubhai V. Shah, Research Triangle Institute
Sara C. Wheeless, P.O. Box 12194, Research Triangle Park, NC 27709

KEY WORDS: quantile, variance estimation, survey data

1. INTRODUCTION

The development of a comprehensive software package for survey data analysis is currently underway at the Research Triangle Institute under contract to the National Center for Health Statistics and the Public Health Service. As part of this effort, significant enhancements are being made to RTI's existing software system. These include the estimation of quantiles, such as the median, and their variances for data arising from complex sample surveys. RTI's existing procedures SESUDAAN, RTIFREQS, and SURREGR use Taylor series linearization for estimating the variance of statistics such as means, proportions, and regression coefficients that are obtained from complex sample surveys. The World Fertility Survey's CLUSTERS and Iowa State University's SUPERCARP and PCCARP are examples of other survey data analysis software packages that use Taylor series linearizations.

The linearized value of a quantile includes a term for the probability density function of the variable of interest. Francisco and Fuller (1986) presented a method for variance estimation based on the Woodruff (1952) confidence interval, and Rao and Wu (1987) have also done work with this estimator. This estimator that does not involve numerical estimation of the density. As part of its software development project, RTI conducted a Monte Carlo simulation that compared the variance obtained using this estimator with that when a histogram of the data was used to estimate the density.

In addition to assessing these two variance estimates, two methods for estimating the quantiles themselves were also compared as part of the simulation study. Both quantile estimation methods are based on a histogram estimator of the population distribution function. The first consists of a two point linear interpolation formula while the second uses a least squares quadratic fit to four points with the fitted equation constrained to be monotone nondecreasing. Histograms based on 20 bins and on 100 bins were considered for both methods.

2. METHODS

Consider a universe Ω of N identifiable units. A probability sample s of size n is a collection of n members of Ω. If Y_k denotes a survey outcome variable that is observable without error, then the finite population cumulative distribution function for the variate Y is

$$F_\Omega(x) = \sum_{k\epsilon\Omega} I\,(Y_k \le x) \div N \qquad (1)$$

where $I(Y_k \le x)$ is the one-zero indicator function for the event $(Y_k \le x)$. The quantile x_p associated with p in the interval (0,1) is

$$x_p = q_\Omega(p) = \inf_\Omega \{Y_k : F_\Omega(Y_k) \ge p,\ p\epsilon(0,1)\} \qquad (2)$$

An unbiased sample estimator for $F_\Omega(x)$ is the Horvitz-Thompson estimator, based on unbiased sample weights w_{sk}, defined such that $w_{sk} = 0$ if unit $k \notin s$,

$$F_s(x) = [\sum_{k\epsilon s} w_{sk}\, I(Y_k \le x)] \div [\sum_{k\epsilon s} w_{sk}].$$

A sample estimator for the p-th quantile is

$$\hat{x}_p = q_s(p) = \inf_s \{Y_k : F_s(Y_k) \ge p,\ p\epsilon(0,1)\}. \qquad (3)$$

The quantile corresponding to p can be estimated from the ordered x's by finding j such that $\hat{F}(x_j) \le p < \hat{F}(x_{j+1})$. Then $\hat{x}_p = x_j$.

2.1 Taylor Series Linearization for a Quantile

Fuller and Francisco (1986) give this following linear approximation for the estimated quantile $\hat{x}_p = q_s(p)$

$$= q_\Omega(p) - [f_\Omega(x_p)]^{-1}[F_s(x_p) - F_\Omega(x_p)] + o_p(n^{-1/2}) \qquad (4)$$

where $F_\Omega(x_p)$ is the population distribution function and $f_\Omega(x_p)$ is the derivative of $F_\Omega(x_p)$ evaluated at $x=x_p$.

Since $F_s(x_p)$ is an unbiased estimator of $F_\Omega(x_p)$, $Var(q_s(p)-q_\Omega(p))$

$$= [f_\Omega(x_p)]^{-2}\, Var[F_s(x_p)] + o_p(n^{-1/4}). \qquad (5)$$

The variance of $F_s(x_p)$ is estimated by substituting the linearized value for the kth sample unit for $F_s(\hat{x}_p)$, $T_k(F_s(\hat{x}_p))$ into a variance formula for a total estimator from the sample design. Hence, one could write the linearized value for x_p as

$$T_k(\hat{x}_p) = f_\Omega(x_p)^{-1} T_k(F_s(\hat{x}_p)) \qquad (6)$$

Using the formula for the linearized value for a ratio, $T_k\,(F_s\,(\hat{x}_p)) = \frac{w_k}{\sum_i w_i}\,[I(x_k \le \hat{x}_p) - p]. \qquad (7)$

2.2 Estimation of the Density

An estimate of the density function is needed in order to estimate the variance of the estimated quantile. Various methods such as kernel estimation, splines, or historams could be used. Francisco and Fuller (1986) show that the following estimator, which comes from Woodruff's (1952) confidence interval on $\hat{x}_p$, is consistent for $f_\Omega(\hat{x}_p)^{-1}$:

$$\hat{\partial}_p = [q_s(U_p) - q_s(L_p)] / (U_p - L_p) \tag{8}$$

where U_p and L_p denote the upper and lower $100(1-\alpha)$ confidence interval endpoints for $F_s(\hat{x}_p)$, with $\hat{x}_p$ viewed as a fixed value of x,

$$U_p = p + t_{\alpha/2}\, SE[F_s(\hat{x}_p)] \tag{9}$$

$$L_p = p - t_{\alpha/2}\, SE[F_s(\hat{x}_p)],$$

and where $SE[F_s(\hat{x}_p)]$ denotes the Taylor series standard error estimator for $F_s(x_p)$.

Then, $$\hat{\partial}_p = [q_s(U_p) - q_s(L_p)]/2t_{\alpha/2} \tag{10}$$

and $$\hat{Var}(\hat{x}_p) = \hat{\partial}_p^2\, Var[F_s(\hat{x}_p)]. \tag{11}$$

U_p and L_p are the upper and lower endpoints for the Woodruff method's $(1-\alpha)$ level confidence interval on $\hat{x}$. The $t_{\alpha/2}$ critical value is the standard normal value such that $Pr\{|Z|>t_{\alpha/2}\} = \alpha$.

The following sections describe two methods based on a histogram for estimating quantiles and the density function. A Monte Carlo simulation was performed to evaluate these quantile estimates and to compare the variance estimates given in (11) with that obtained using a histogram to estimate the density.

3. ESTIMATION OF THE DISTRIBUTION FUNCTION

Ideally quantiles would be estimated using the cumulative distribution function as described in Section 2. This requires sorting the data by the variable whose quantile is to be estimated. Sorting is not practical for estimating the quantiles for a large number of data items or for many domains since sample surveys typically consist of a large number of observations. In addition, algorithms for Taylor series variance estimation typically require that the data file be sorted by the sample design variables (for example, stratum, primary sampling unit, secondary sampling unit, etc.).

Alternatives for estimating the distribution function are kernel density methods, splines, and histogram estimators. The histogram estimator with equal width bins was used in this study because of its simplicity. Histograms, like other density estimators, are sensitive to the number of bins. Scott (1979) derives a formula for the optimal histogram bin width for density estimation.

4. ESTIMATION OF QUANTILES

Given the histogram estimate, $\hat{F}$ of the distribution function, two methods were considered for the estimation of quantiles. One method was a two point, linear interpolation formula and the other was a least squares fit of a quadratic to four points with the additional restriction of of enforcing a monotonically increasing function. Suppose there are m bins in the histogram, and denote the endpoints of the bins by $x_0', x_1', \ldots, x_m'$ where x_0' and x_m' are the maximum and minimum values of the data.

4.1 Linear Interpolation

For linear interpolation, the quantile for a given percentage point, p,was estimating by finding j such that $\hat{F}(x_j') \le p < \hat{F}(x_{j+1}')$. Then, the pth quantile was estimated by the linear interpolation formula

$$\hat{X}_p = x_j' + b(x_{j+1}' - x_j') \text{ where}$$

$$b = [p - \hat{F}(x_j')] / [\hat{F}(x_{j+1}') - \hat{F}(x_j')].$$

The estimate of the derivative used in equation (5) is the slope

$$\left(x_{j+1}' - x_j'\right) / \left(\hat{F}(x_j'+1) - \hat{F}(x_j')\right) .$$

4.2 Quadratic Fit to Four Points, Enforcing Monotonicity

A least squares fits of the equation $F(x) = ax^2 + bx + c$ was made to the four points surrounding p. First, j such that $\hat{F}(x_j') \le p < \hat{F}(x_{j+1}')$ was was found. If j=0, then the four lower bins of the histogram were used; if j=m-1 the four upper most bins of the histogram were used. Otherwise the four points used were

$$(x_{j-1}', \hat{F}(x_{j-1}')),\ (x_j', \hat{F}(x_j')),$$

$$(x_{j+1}', \hat{F}(x_{j+1}')), \text{ and } (x_{j+2}', \hat{F}(x_{j+2}')).$$

The fitted quadratic equation need not be monotonic nondecreasing, particularly if some of the $\hat{F}$'s are the same. It is monotonically nondecreasing on the interval, however, if the intercept, -b/2a, is outside the range [min(x'), max(x')] where min(x') and max(x') are the minimum and maximum values of the four x' points. In this case, the pth percentile was estimated by the root

$$\hat{x}_p = \frac{-b \pm \sqrt{b^2 - 4a(c-p)}}{2a}$$

that fell in the interval [min(x'), max(x')]. The estimate of the derivative in Equation (5) was $(2a\hat{x}_p+b)^{-1}$.

When the intercept was within the range [min(x'), max(x)], then the intercept was forced to fall at one of the endpoints. That is,

$-b/2a = \min(x')$ or $-b/2a = \max(x')$. Least squares solutions to $F(x) = ax^2 - 2a \min(x)x + c$ and $F(x) = ax^2 - 2a \max(x')x + c$ were found. The solution with the smallest residual sums of squares was used to estimate x_p. Then,

$$\hat{x}_p = \frac{2\min(x') \pm \sqrt{(2\min(x')^2 - 4a\ (c-p)}}{2a},$$

or the similar result obtained by substituting $\max(x')$ for $\min(x')$. The estimate of the derivative in Equation (5) was

$\{2a(\hat{X}_p - \min(x'))\}^{-1}$ or
$\{2a(\hat{X}_p - \max(x')))\}^{-1}$.

5. SIMULATION AND RESULTS

Two Monte Carlo simulations were performed to compare and evaluate estimates of quantiles and estimates of their variances. The first simulation was performed on a population of 10,000 random numbers from a normal distribution with zero mean and variance equal to unity. The second was performed on a population of 1,000 log normal random numbers with mean 4.65 and variance 1.99. Rao and Wu (1987) concluded that $\alpha = 0.05$ was a reasonable choice, so $t_\alpha = 1.96$ was used in equation (9).

5.1 Normal Population

From the population of 10,000 N(0,1) random numbers, 10,000 simple random samples of size 500 were selected. For each sample of 500, a histogram with equisized bins was used to estimate the distribution function. Using Scott's formula the optimal bin width for estimating a density function was approximately 0.44, or about 16 bins. For this simulation we used histograms with 20 bins and 100 bins.

Quantiles were estimated for p=0.10, 0.25, 0.50, 0.75, and 0.90 using the two point (linear) interpolation formula and the four point (least squares fit enforcing monotonicity) formula. Variance estimates were obtained for each quantile estimate using a histogram density estimate and the inverted confidence interval formula. The linearized values were substituted into the formula for the variance of a total from a simple random sample,

$$n^2 \sum_{i=1}^{n} (x_i-\bar{x})^2 /(n-1) \quad \text{with } n=500.$$

Table 1 presents the true quantile estimates for the population of size 10,000. Also given are the means of the quantile estimates obtained from the linear interpolation and the quadratic least square fit for the 20 and 100 bin histograms. The bias is small in all four cases. The biases are generally smaller for the 100 bin histogram, and the estimates are virtually identical regardless of whether the linear or quadratic fit is used. The 20 bin histogram with a quadratic fit performs almost as well.

Eight estimates of the variance were obtained from the combinations of the linear and quadratic formulas, the 20 and 100 bin histograms, and the histogram and confidence interval methods. Table 2 presents these variance estimates along with the computed variance of 10,000 quantile estimates. In the tables, V_H denotes variances based on the histogram density and V_W variances based on inverting Woodruff's confidence interval. Except for the tails of the distribution function (p = 0.10 and p = 0.90), the estimates based on Woodruff's symmetric confidence interval are roughly equal with respect to the first two significant digits; this is to be expected since their method depends on quantile estimation and the bias was found to be small regardless of whether the linear or quadratic method, or 20 or 100 bins, were used. Note also that the variance estimates are almost equal (in the first two significant digits) to each other for all except the 100 bin, linear interpolation formula.

Correlations between the variance estimates obtained from the two methods for the linear and quadratic formulas, and the 20 and 100 bin histograms were also calculated For the 20 bin histogram the correlations were all above 0.70; those obtained from the quadratic least squares fit are all above 0.90. The correlations obtained from the 100 bin histogram were not as high; the quadratic least squares fit gave values of about 0.4 in the tails of the distribution and about 0.7 elsewhere; the linear formula gave values in the range 0.2 to 0.4.

Table 3 presents coverage probabilities obtained when 95%, confidence intervals were computed using the estimated quantiles and variances. These coverage probabilities are the percentage of 10,000 confidence intervals that contain the true population quantile (given in Table 1). The confidence intervals obtained using the 100 bin histogram to estimate the density (with both linear and quadratic interpolation formulas) contained the true values less often than the 95 advertised for the confidence interval. These same confidence intervals with the symmetric confidence interval method also contained the true value generally less often than 95%, but were much closer than those using the histogram. The coverage probabilities from the 20 bin histogram were all close to 95% for both methods.

5.2 Lognormal Population

From the population of 1,000 lognormal random numbers, 10,000 samples of size 300 were selected without replacement. Histograms with 20 and 100 bins were used; Scott's formula gives 16 bins as the optimal number. The same statistics produced for the normal data were produced for this population as well. The finite population correction factor was used when calculating the variances. Tables 4, 5 and 6 present the summaries. The same observations made for the normal population are seen here as well. The bias in the quantile estimates (Table 4) is small, regardless of the number of bins used. With 20 bins (nearly optimal), the histogram and Woodruff interval give similar estimates, and the correlations between the variance estimates were high - above 0.8. For this skewed distribution, however, the coverage probabilities were not as close to 95% as they were for the normally distributed data.

6. ACKNOWLEDGEMENTS

This work was performed under Contract 282-86-0107 for the National Center for Health Statistics.

7. REFERENCES

Francisco, C. A. and W. A. Fuller (1986). "Estimation of the Distribution Function with a Complex Survey." Presented to the Section on Survey Research Methods at the 1986 Annual ASA meetings.

Fuller, W. A., D. Schnell, G. Sullivan, and W. J. Kennedy (1987). "Survey Variance Computations on the Personal Computer." Invited Paper 10.2 at the 46th Session of the ISI.

Rao, J. N. K. and C. F. J. Wu (1987). "Methods for Standard Errors and Confidence Intervals from Sample Survey Data: Some Recent Work." Invited Paper 10.1 at the 46th Session of the ISI.

Rao, J. N. K. (January 13, 1987). Personal Communication to B. V. Shah.

Scott, David W. (1979). "On optimal and data-based histograms," Biometrika, 66, 3, 605-610.

Shah, B. V. (1979). SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data. Research Triangle Park, NC: Research Triangle Institute.

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

Table 1. Comparison of Quantile Estimates
Normal Data

	Percentage	Population Value of the Quantile	Linear Formula Monte Carlo Estimate	Linear Formula Bias	Quadratic Least Squares Fit Monte Carlo Estimate	Quadratic Least Squares Fit Bias
20 Bins	10	-1.282	-1.289	0.007	-1.280	-0.002
	25	-0.691	-0.695	0.004	-0.692	0.001
	50	-0.006	-0.003	-0.003	-0.003	-0.003
	75	0.668	0.670	-0.002	0.665	0.003
	90	1.262	1.273	-0.011	1.266	-0.004
100 Bins	10	-1.282	-1.281	0.001	-1.281	-0.001
	25	-0.691	-0.689	-0.002	-0.689	-0.002
	50	-0.006	-0.002	-0.004	-0.002	-0.004
	75	0.668	0.666	0.002	0.666	0.002
	90	1.262	1.263	-0.001	1.263	-0.001

Table 2. Comparison of Variance Estimates
Normal Data

Percentage	Linear Interpolation V_H	Linear Interpolation V_W	Linear Interpolation Monte Carlo Estimate	Quadratic Least Squares Fit V_H	Quadratic Least Squares Fit V_W	Quadratic Least Squares Fit Monte Carlo Estimate
20 Bins						
10	0.005494	0.005385	0.004886	0.005698	0.005733	0.004913
25	0.003732	0.003754	0.003326	0.003807	0.003703	0.003286
50	0.003283	0.003279	0.002936	0.003357	0.003254	0.002989
75	0.004059	0.004011	0.003634	0.003841	0.003882	0.003529
90	0.005768	0.005834	0.005072	0.005421	0.005759	0.004968
100 Bins						
10	0.006193	0.005648	0.004813	0.005403	0.005645	0.004813
25	0.004110	0.003772	0.003488	0.003762	0.003771	0.003499
50	0.003488	0.003215	0.003056	0.003210	0.003213	0.003055
75	0.004784	0.004191	0.004059	0.004266	0.004193	0.004080
90	0.007522	0.005874	0.005434	0.006021	0.005877	0.005426

Table 3. Coverage Probabilities for 95% Confidence Intervals Normal Data

	20 bins				100 bins			
	Linear		Quadratic		Linear		Quadratic	
Percentage	V_H	V_W	V_H	V_W	V_H	V_W	V_H	V_W
10	94.36	95.78	95.42	95.76	90.63	95.29	93.95	95.23
25	94.89	95.50	96.25	95.74	91.58	94.47	93.77	94.62
50	95.50	95.78	96.01	95.48	92.43	94.88	94.04	94.81
75	94.34	94.81	95.29	94.99	90.69	93.36	92.02	93.42
90	94.18	95.38	95.43	95.56	88.28	94.72	92.44	94.62

Table 4. Comparison of Quantile Estimates Lognormal Data

		Linear Formula		Quadratic Least Squares Fit	
Percentage	Population Value of the Quantile	Monte Carlo Estimate	Bias	Monte Carlo Estimate	Bias
20 Bins					
10	3.025	3.008	-0.017	3.021	-0.004
25	3.656	3.609	-0.047	3.608	-0.048
50	4.426	4.442	0.016	4.437	0.011
75	5.443	5.398	-0.045	5.398	-0.043
90	6.541	6.406	-0.135	6.395	-0.011
100 Bins					
10	3.025	3.021	-0.004	3.022	-0.003
25	3.656	3.613	-0.043	3.614	-0.001
50	4.426	4.444	0.018	4.444	0.018
75	5.443	5.386	-0.057	5.386	-0.057
90	6.541	6.390	-0.151	6.392	-0.149

Table 5. Comparison of Variance Estimates
Lognormal Data

Percentage	Linear Interpolation V_H	Linear Interpolation V_W	Linear Interpolation Monte Carlo Estimate	Quadratic Least Squares Fit V_H	Quadratic Least Squares Fit V_W	Quadratic Least Squares Fit Monte Carlo Estimate
20 Bins						
10	0.005363	0.005439	0.004689	0.005627	0.005733	0.005020
25	0.004990	0.004965	0.004645	0.053150	0.004839	0.004584
50	0.007813	0.007791	0.007326	0.007550	0.007912	0.007811
75	0.009524	0.009529	0.008629	0.009288	0.009046	0.008064
90	0.028641	0.026616	0.025620	0.022929	0.026193	0.024564
100 Bins						
10	0.004697	0.004881	0.003822	0.004455	0.004872	0.003842
25	0.005372	0.004893	0.004682	0.004890	0.004891	0.004718
50	0.009740	0.007633	0.008108	0.007914	0.007627	0.008005
75	0.012278	0.009924	0.009796	0.010206	0.009936	0.009839
90	0.030066	0.028619	0.028752	0.042568	0.028662	0.029038

Table 6. Coverage Probabilities for 95% Confidence Intervals
Lognormal Data

Percentage	20 bins				100 bins			
	Linear		Quadratic		Linear		Quadratic	
	V_H	V_W	V_H	V_W	V_H	V_W	V_H	V_W
10	94.25	94.73	95.72	95.38	93.99	96.37	96.04	96.32
25	90.18	90.38	91.39	88.20	86.73	88.56	87.49	88.74
50	94.18	94.55	93.80	94.01	83.72	94.15	90.82	93.77
75	89.00	91.19	92.51	91.88	81.87	86.12	86.39	86.32
90	77.64	83.56	77.53	81.91	64.99	79.59	71.38	79.55