Journal of Official Statistics
Vol. 9, No. 3, 1993, pp. 673–689

Standard Error and Confidence Interval Estimation for the Median

Philip J. McCarthy[1]

Abstract: This paper discusses and compares a wide variety of procedures (most of which are based on resampling) that may be used to obtain standard error estimates for the median of a sample drawn without replacement from a finite population. Confidence intervals are also considered. These procedures are evaluated by drawing repeated samples from a number of artificial populations, as well as drawing clustered samples from a population derived from U.S. Current Population Survey results for the month of September 1989.

Key words: Balanced repeated replications; bootstrap; jackknife; order statistics.

1. Introduction

This study is concerned with methods for estimating standard errors and confidence intervals for the median when simple random samples are drawn without replacement from finite populations.

The median has long been used as a robust estimate of location. Furthermore, the asymptotic variance of the sample median is known to be $1/4nf^2(\theta_{.5})$, where n is the sample size, f is the density of the population from which the sample was drawn and $\theta_{.5}$ is the population median (see Wilks 1962, p. 273). Since this result depends upon having a "large" sample of n i.i.d. observations for a variable with a density which is known at the population median, it is of little or no value for the situation that is the primary concern of this paper – possibly small finite populations with samples drawn without replacement. Thus there is no density, the sample sizes are small, and the observations are not independent.

During the 1930s and 1940s nonparametric methods were developed for obtaining confidence intervals for the median and other quantiles, intervals that do not depend upon knowing the density of the variable or upon having large samples. But they do depend upon having independent observations from a continuous c.d.f. $F(y)$. Wilks (1962, p. 333) did develop an exact nonparametric confidence interval procedure for the median

[1] 358 Ives Hall, New York State School of Industrial and Labor Relations, Cornell University, Ithaca, N.Y. 14853, U.S.A. and U.S. Bureau of Labor Statistics, Room 2126, 441 G St. NW, Washington, D.C. 20212, U.S.A.
Acknowledgements: A major portion of the research was done while McCarthy was the holder of an ASA/NSF/BLS Senior Research Fellowship at the U.S. Bureau of Labor Statistics, Washington, D.C. 20212, Feb – May 1989. The author appreciates the comments and suggestions of J.N.K. Rao, of Ali Hadi and Martin Wells at Cornell, and of Robert Casady and Richard Valliant of the Bureau of Labor Statistics. He is especially indebted to Cathryn Dippo of BLS who made the arrangements to obtain CPS data tapes from the U.S. Bureau of the Census, and to Troy Bishop of BLS who programmed the simulations based on the CPS data.

based on simple random samples drawn without replacement from a finite population, but it requires that the population elements have different values of the variable, a circumstance that will not ordinarily hold.

Recognizing these problems, Woodruff (1952) devised a method for obtaining confidence intervals for a quantile when samples are drawn without replacement from a finite population. The only requirement is that the probability of each item in the population coming into the sample be known. Furthermore, Kovar, Rao and Wu (1988) suggested a procedure for obtaining an estimate of the standard error of the median from these confidence intervals.

In the 1960s and 1970s, there was a great deal of attention devoted to the use of the delete-1 jackknife for variance estimation. However, Miller (1974) showed that this estimator is not consistent for nonsmooth quantities such as the sample quantiles. In order to obtain consistency of the variance estimate for sample quantiles, Shao and Wu (1986) demonstrate that the delete-d jackknife does provide this property. In addition, Shao (1989) argues that one can sample the $\binom{n}{d}$ recomputations of a point estimate required for the delete-d jackknife and still obtain consistency.

In 1979 Efron introduced the bootstrap, a procedure where repeated samples of size n are drawn with replacement from the original sample of n i.i.d. observations. This was modified in a variety of ways by Gross (1980), Bickel and Freedman (1981), McCarthy and Snowden (1985), Rao and Wu (1988), and Sitter (1990) to cover the case where samples are drawn without replacement from a finite population. Many highly theoretical papers have also appeared (e.g., Falk and Kaufmann 1991 and their cited references), but these are of no direct concern here since this paper is of a more applied nature.

Our goal in this paper is to consider some of the methods that have just been mentioned. Through simulation we will evaluate their behavior in producing standard error estimates and confidence intervals for the median when small to moderate sized samples are drawn without replacement from a variety of small, artificial populations. This paper is in the spirit of McKean and Schrader (1984), but the restriction to samples drawn without replacement from finite populations changes the emphasis.

2. Methods

2.1. Direct standard error estimation

The first two methods considered depend on McCarthy and Snowden's (1985) modification of the bootstrap, based upon the behavior of the sample mean. Briefly, this is:

1. If samples of n are drawn from a finite population of N, then

$$\hat{V}(\bar{y}) = \frac{(1-f)}{n} \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}.$$

2. If bootstrap samples of n^* are drawn from a sample with replacement, then

$V(\bar{y}^* | \text{original sample}) =$

$$\frac{1}{n^*} \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}.$$

3. If we equate $\hat{V}(y)$ and $V(\bar{y}^* |$ original sample), then

$$n^* = \frac{n-1}{1-f} \tag{2.1}$$

where $f = n/N$, the sampling fraction. The quantity n^* as obtained from (2.1) is not usually an integer. In the present study, we chose n and the population size so that n^* is an integer, and have not investigated the situation where n^* would have to be approximated by an integer.

Suppose we are given an original sample of n (odd integer) – ordered observations $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ with median $\hat{\theta}$– and then draw from this sample a bootstrap sample of size n^* as given by (2.1) with median $\hat{\theta}^*$. Then it follows from an adaptation of Efron's argument (1982, Sec. 10.3) that the probability that $\hat{\theta}^* = y_{(k)}$ is given by

$$p_{(k)} = \sum_{j=0}^{m^*-1} \frac{n^*!}{j!(n^*-j)!} \times \left(\frac{k-1}{n}\right)^j \left(1 - \frac{k-1}{n}\right)^{n^*-j} - \sum_{j=0}^{m^*-1} \frac{n^*!}{j!(n^*-j)!} \left(\frac{k}{n}\right)^j \times \left(1 - \frac{k}{n}\right)^{n^*-j} \quad (2.2)$$

where $m^* = (n^*+1)/2$. This result, with $n^* = n$, has also been given by Maritz and Jarrett (1978) and used by McKean and Schrader (1984). A bootstrap estimate of the standard error of the median is therefore given by

$$\hat{\sigma}_{MJ} = \left(\sum_{k=1}^{n} p_{(k)}[y_{(k)} - \hat{\theta}]^2\right)^{1/2}. \quad (2.3)$$

Another estimate for the standard error of the median was developed on a more or less *ad hoc* basis. It appeared interesting to see if one could make use of the average of absolute deviations about the sample median, that is

$$\sum_{k=1}^{n} p_{(k)}|y_{(k)} - \hat{\theta}|.$$

Some preliminary investigations on small populations suggested that this would provide an underestimate of the standard error of the median and that the normal distribution correction, $\sqrt{\pi/2}$, might improve the situation. Hence we have used

$$\hat{\sigma}_{MD} = \sqrt{\pi/2}\sum_{k=1}^{n} p_{(k)}|y_{(k)} - \hat{\theta}| \quad (2.4)$$

as an estimate of the standard error.

Note that it is not necessary to draw bootstrap samples since the preceding theory makes it possible to compute $\hat{\sigma}_{MJ}$ and $\hat{\sigma}_{MD}$ directly. The required probabilities, $p_{(k)}$, were obtained by using an APL program on an Apple Macintosh Plus.

Rao and Wu (1988) have developed a different approach to the bootstrap estimation of the standard error of the mean of a population. Given a sample drawn without replacement from a finite population, consider drawing bootstrap samples of size m (arbitrary) with replacement from the original sample. This will result in the bootstrap observations $y_1^*, y_2^*, \ldots, y_m^*$. Next, obtain the transformed values

$$\tilde{y}_i = \bar{y} + \sqrt{\frac{m(1-f)}{n-1}}\,(y_i^* - \bar{y}). \quad (2.5)$$

In the linear case, the bootstrap variance estimator based on the $\tilde{y}_i$ reduces to the customary variance estimator $\hat{V}(\bar{y}) = n^{-1}(1-f)(n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. This has the advantage over the McCarthy-Snowden approach that their n^* may not be an integer. In the present study, we took $m = n$ and then worked with medians based on the $\tilde{y}_i$. Thus if $\hat{\theta}$ is the original sample median and $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_B$ are the medians of B bootstrap samples, we have

$$\hat{\sigma}_{RW} = \left(\sum_{i=1}^{B} \frac{(\tilde{\theta}_i - \hat{\theta})^2}{B-1}\right)^{1/2}. \quad (2.6)$$

A corrected version of (2.5) for medians has been proposed (J.N.K. Rao, personal communication, 1991), but we learned of this too late to include it in the present simulations.

Shao (1989) discusses the general delete-d

jackknife variance estimator. Ordinarily one would like to enumerate all $\binom{n}{d}$ possibilities, but the number of these rapidly becomes extremely large. Shao suggests that one sample m cases without replacement from the total number, with m approximated by $n^{3/2}$. Thus, given a sample of n drawn without replacement from N and associated median $\hat{\theta}$, m delete-d samples are drawn from the n with associated medians $\hat{\theta}_1, \ldots, \hat{\theta}_m$. An estimate $\hat{\sigma}_{SHAO}$ of the standard error of $\hat{\theta}$ is then given by

$$\hat{\sigma}_{SHAO} = \left(\left(1 - \frac{n}{N}\right)\frac{n-d}{dm} \times \sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta})^2\right)^{1/2}. \qquad (2.7)$$

Sitter (1990) proposed a bootstrap method which he called the mirror-match method. The procedure is:

1. Choose $1 \leq n' < n$ and draw without replacement a sample of n' from the original sample of n to get $y^* = (y_1^*, y_2^*, \ldots, y_{n'}^*)$.
2. Repeat step 1, $k = [n(1-f^*)]/[n'(1-f)]$ times independently, replacing the resamples of size n' each time to get $y_1^*, y_2^*, \ldots, y_{n^*}^*$ where $f = n/N$, $f^* = n'/n$ and $n^* = kn'$. Let $\hat{\theta}^*$ be the median of the n^* observations. (If k is a non-integer a randomization between bracketing integers is used.)
3. Repeat steps 1 and 2 a large number of times, B, to get $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$. An estimate $\hat{\sigma}_{SITTER}$ of the standard error of $\hat{\theta}$ is then given by

$$\hat{\sigma}_{SITTER} = \left[\frac{\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta})^2}{(B-1)}\right]^{1/2}. \qquad (2.8)$$

Note that if $n' = 1$, the mirror-match method is the same as the McCarthy-Snowden method described earlier.

2.2. *Confidence intervals*

Woodruff (1952) proposed a method of determining confidence intervals for the population median based upon the empirical CDF (cumulative distribution function). Given a sample of n observations, the empirical CDF is defined to be

$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq z) \qquad (2.9)$$

where $I(y_i \leq z) = 1$ if y_i is $\leq z$ and 0 otherwise. If we then look at the fraction of observations in the sample that will be less than the population median, and obtain 95% (or any other number) normal probability limits for this fraction, the limits will be

$$.5 \pm 1.96\sqrt{\frac{N-n}{N-1}\,\frac{.5(.5)}{n}}.$$

These limits are now inverted with respect to $F_n(z)$ leading to confidence limits for the population median

$$\left[F_n^{-1}\left(.5 - 1.96\sqrt{\frac{N-n}{N-1}\,\frac{.5(.5)}{n}}\right)\right], \left[F_n^{-1}\left(.5 + 1.96\sqrt{\frac{N-n}{N-1}\,\frac{.5(.5)}{n}}\right)\right]. \qquad (2.10)$$

The median is estimated as

$$\hat{\theta} = F_n^{-1}(.5) = inf\,(z : F_n(z) \geq .5). \qquad (2.11)$$

Since $F_n(z)$ is a step function, in performing the inversion indicated by (2.10), we used a linear approximation to $F_n(z)$. That is, for $k = 0, 1, 2, \ldots, (n-1)$ and $0 \leq \alpha <$

$1/n$ we took

$$F_n^{-1}\left(\frac{k}{n}+\alpha\right) = y_{(k)} + \alpha\,[y_{(k+1)} - y_{(k)}] \quad (2.12)$$

where $y_{(0)} = LL$, an appropriately selected *Lower Limit* for the population under consideration. Another procedure for obtaining confidence intervals for the median is Efron's bootstrap percentile method (1982, Sec. 10.5). This is based on the order statistics of the sample and the probabilities, $p_{(k)}$, given in (2.2). Thus the interval

$$[y_{(k)}, y_{(n-k+1)}] \quad (2.13)$$

should provide a confidence interval for the population median with confidence coefficient $(1-\alpha)$ where, *given a value of k*,

$$\alpha/2 = \sum_{i=1}^{k-1} p_{(i)} + .5p_{(k)}. \quad (2.14)$$

These intervals are similar in spirit to the ordinary nonparametric confidence intervals for the median based upon the order statistics of n independent and identically distributed observations from a continuous distribution. The justification for using $.5p_{(k)}$ is given by Efron (1982, Sec. 10.5).

Symmetric confidence intervals can always be produced by using the normal assumptions, namely

$$\left[\hat{\theta} - z_{\alpha/2}\hat{\sigma}, \hat{\theta} + z_{\alpha/2}\hat{\sigma}\right] \quad (2.15)$$

where $\hat{\sigma}$ can be any of $\hat{\sigma}_{MJ}$, $\hat{\sigma}_{MD}$, $\hat{\sigma}_{RW}$, $\hat{\sigma}_{SHAO}$, or $\hat{\sigma}_{SITTER}$.

It is possible to obtain bootstrap-t intervals by using a two stage bootstrap procedure as described by Rao and Wu (1988). The steps are as follows:

1. Draw a sample from the population and obtain an estimate $\hat{\theta}$. Using a bootstrap procedure also estimate the standard error $\hat{\sigma}$ of $\hat{\theta}$ by any one of $\hat{\sigma}_{MJ}$, $\hat{\sigma}_{MD}$, or $\hat{\sigma}_{RW}$, $\hat{\sigma}_{SHAO}$, or $\hat{\sigma}_{SITTER}$.
2. From the sample drawn in (1), draw a bootstrap sample. Let its median be $\hat{\theta}_1$. Now draw bootstrap samples from this sample and obtain $\hat{\sigma}_1$ by any one of the $\hat{\sigma}$'s as given in (1). Compute a t value as $t_1 = (\hat{\theta}_1 - \hat{\theta})/\hat{\sigma}_1$. Repeat this process, leading to $t_1, t_2, t_3, \ldots$ and an empirical distribution of the t's. Compute the values t_{LOWER} and t_{UPPER} such that, for example, 2.5 percent of the t's are less than or equal to t_{LOWER} and 97.5 percent are less than or equal to t_{UPPER}.
3. Confidence intervals are computed as

$$[\hat{\theta} - t_{LOWER}\hat{\sigma}, \hat{\theta} + t_{UPPER}\hat{\sigma}]. \quad (2.16)$$

2.3. *Standard error estimates based on confidence intervals*

A number of authors have suggested that standard error estimates can be obtained through the use of confidence intervals, e.g., McKean and Schrader (1984) and Kovar, Rao, and Wu (1988). In this work, we have considered two such estimates of the standard error:

1. Based upon the Woodruff confidence interval. Take $(UCL - LCL)/2z_{\alpha/2}$, where the confidence coefficient is $1-\alpha$. Kovar, Rao, and Wu present some empirical evidence showing that .05 is a reasonable choice for α. Thus we have

$$\hat{\sigma}_{WOOD} = \frac{(UCL - LCL)}{2z_{.025}}. \quad (2.17)$$

2. Based upon the order statistics of the sample. Take

$$\hat{\sigma}_{ORDER} = [y_{(n-k+1)} - y_{(k)}]/2z_{\alpha/2} \quad (2.18)$$

where $\alpha/2$ is given by (2.14) and k was chosen so as to make $\alpha/2$ as close to .025 as possible.

3. Simulation Study Based on Artificial Populations

3.1. Populations and sample sizes

Two different population sizes were employed in this simulation study: $N = 25$ and $N = 81$. These population sizes were chosen for the following reasons: (1) They are relatively small so that the simulations are not unduly time consuming and (2) they make possible a range of sample sizes so that n^* (as given by expression (2.1)) is an integer. For $N = 25$ the sample sizes were $n = 5$ with $n^* = 5$ (as given by 2.1) and $n = 15$ with $n^* = 35$. For $N = 81$ we used $n = 9$ with $n^* = 9$, $n = 27$ with $n^* = 39$ and $n = 45$ with $n^* = 99$. For each sample size four different populations were used:

- Integers 1-25 and 1-81,
- Normal Scores for 25 observations and for 81 observations,
- 25 observations and 81 observations drawn independently from a χ^2 distribution with 10 degrees of freedom,
- 25 observations and 81 observations drawn independently from a χ^2 distribution with 2 degrees of freedom.

Thus, we have two symmetric populations with differing shapes, one population slightly skewed to the right and one population heavily skewed to the right.

3.2. Standard error estimates

For each population and sample size, 1000 samples were drawn without replacement from the population. For each sample and each estimator, an estimate of the standard error of the median was obtained and these estimates were averaged over the 1000 samples. The averages were then compared with the true standard error (determined by selecting 5000 samples from the population). The results are given in Table 1. The following observations are in order concerning the manner in which these results were obtained:

1. APL programs were written to perform the operations, although some were done on an Apple Macintosh Plus while others were done on an IBM PC/AT. A new random seed was introduced for each run.
2. $\hat{\sigma}_{MJ}$ and $\hat{\sigma}_{MD}$ were computed on the same set of 1000 samples, but a different set was used for each combination of population and sample size.
3. For each combination of population and sample size, an independent set of 1000 samples was used for each of $\hat{\sigma}_{RW}$, $\hat{\sigma}_{WOOD}$, $\hat{\sigma}_{SHAO}$, $\hat{\sigma}_{SITTER}$, and $\hat{\sigma}_{ORDER}$.
4. If a result seemed out of line, e.g., for $\hat{\sigma}_{WOOD}$ (Normal Scores 25, $n = 25$), another set of 1000 samples was run. The two agreed in every instance.
5. It should be noted that for $n = 5$ and $n = 9$ $\hat{\sigma}_{MJ}$ and $\hat{\sigma}_{RW}$ should be comparable. The only difference is that (2.2) was used for $\hat{\sigma}_{MJ}$ instead of drawing bootstrap samples while bootstrap samples were actually selected for $\hat{\sigma}_{RW}$. One hundred bootstrap samples were used to compute $\hat{\sigma}_{RW}$. It will be noted from Table 1 that the two estimators provide almost identical results for these cases, and for many others as well.
6. For the Shao delete-d jackknife, the following sample sizes were used: $N = 25$, $n = 5$, $d = 2$, $m = 10$(all); $N = 25$, $n = 15$, $d = 8$, $m = 58$; $N = 81$, $n = 9$, $d = 4$, $m = 27$; $N = 81$, $n = 27$, $d = 14$, $m = 140$; and $N = 81$, $n = 45$, $d = 22$, $m = 302$.
7. For the Sitter mirror-match method, the following sample sizes were used:

Table 1. Comparison of standard error estimates (Entries are (estimated standard error)/(true standard error))

Population and sample size	MJ	MD	RW	SHAO	SITTER	Wood	Order
Integers 1–25							
$n = 5$	1.11	.94	1.13	.97		1.01	1.09
$n = 15$	1.08	.98	1.08	.96	1.09	1.02	1.08
Integers 1–81							
$n = 9$	1.12	.99	1.10	.99		1.02	.99
$n = 27$	1.06	.99	1.08	1.00	1.07	1.01	1.02
$n = 45$	1.03	.98	1.05	.98	1.06	.99	1.02
Normal scores 25							
$n = 5$	1.27	1.05	1.28	.98		1.72	1.31
$n = 15$	1.12	1.00	1.11	.98	1.11	1.04	1.10
Normal scores 81							
$n = 9$	1.17	1.02	1.14	1.05		1.18	1.11
$n = 27$	1.11	1.03	1.10	1.03	1.10	1.03	1.05
$n = 45$	1.08	1.02	1.09	1.01	1.07	1.03	1.05
Chi-square 10df.25							
$n = 5$	1.22	.98	1.23	.95		1.58	1.24
$n = 15$	1.28	1.05	1.27	1.12	1.27	1.23	1.23
Chi-square 10df.81							
$n = 9$	1.34	1.07	1.35	1.12		1.20	1.35
$n = 27$	1.20	1.03	1.37	1.18	1.19	1.11	1.13
$n = 45$	1.18	1.01	1.93	1.17	1.18	1.06	1.12
Chi-square 2df.25							
$n = 5$	1.19	.97	1.18	.97		.94	1.13
$n = 15$	1.14	.96	1.21	.99	1.13	.98	1.09
Chi-square 2df.81							
$n = 9$	1.26	1.00	1.23	1.07		1.03	1.17
$n = 27$	1.20	1.03	1.34	1.14	1.19	1.03	1.09
$n = 45$	1.15	1.01	1.87	1.09	1.14	.95	1.01
Average	1.17	1.01	1.26	1.04	1.13	1.11	1.12

Method of estimation: MJ, MD, RW, SHAO, SITTER, Wood, Order.

$N = 25, n = 15, n' = 5, k = 5$; $N = 81$, $n = 27$, $n' = 9$, $k = 3$; and $N = 81$, $n = 45$, $n' = 9$, $k = 9$. Note that $N = 25$, $n = 5$ and $N = 81$, $n = 9$ are not included. In both cases it would be desirable to have $n' = 1$ and the results would be the same as for $\hat{\sigma}_{MJ}$.

The interpretation of the contents of Table 1 seem fairly clear – $\hat{\sigma}_{MD}$ provides the standard error estimate with smallest bias while $\hat{\sigma}_{SHAO}$ is close behind. The average ratios at the bottom of the table provide a rough ordering of the other methods.

3.3. *Relative stability of the standard error estimates*

The simulations described in Sec. 3.2 in connection with Table 1 also provided

Table 2. Relative stabilities of the standard error estimates (Entries are (standard error of the standard error estimates)/(true standard error))

Population and sample size	Method of estimation MJ	MD	RW	SHAO	SITTER	Wood	Order
Integers 1–25							
$n = 5$	.369	.344	.381	.493		.196	.237
$n = 15$	.287	.269	.252	.213	.316	.234	.213
Integers 1–81							
$n = 9$	.347	.345	.387	.436		.233	.222
$n = 27$	.291	.301	.284	.251	.293	.224	.242
$n = 45$	.234	.240	.211	.175	.252	.186	.222
Normal scores 25							
$n = 5$	.452	.417	.488	.518		.270	.406
$n = 15$	.309	.288	.280	.234	.330	.243	.230
Normal scores 81							
$n = 9$	.384	.377	.402	.467		.329	.330
$n = 27$	.302	.311	.284	.255	.327	.229	.275
$n = 45$	.244	.244	.230	.194	.244	.211	.231
Chi-square 10df.25							
$n = 5$	.439	.393	.457	.503		.241	.450
$n = 15$	.402	.406	.373	.281	.424	.331	.298
Chi-square 10df.81							
$n = 9$	.609	.559	.652	.674		.466	.545
$n = 27$	.421	.396	.506	.447	.469	.420	.412
$n = 45$	.408	.393	.532	.327	.434	.304	.364
Chi-square 2df.25							
$n = 5$	.452	.423	.461	.555		.286	.337
$n = 15$	.293	.296	.317	.249	.334	.267	.319
Chi-square 2df.81							
$n = 9$	.604	.532	.606	.639		.425	.502
$n = 27$	.483	.436	.460	.435	.465	.420	.481
$n = 45$	.335	.335	.455	.291	.326	.277	.316
Average	.383	.365	.401	.382 (.279)*	.351	.290 (.279)*	.332

* These averages were computed by ignoring the $n = 5$ and $n = 9$ entries.

measures of the relative stability of the standard error estimates. These were computed as:

$$\frac{\text{Standard error of the standard error estimate}}{\text{True standard error}} \tag{3.1}$$

and the results are given in Table 2. As is apparent from the table, $\hat{\sigma}_{WOOD}$ has the best stability of the seven estimators. In terms of overall performance, the Shao delete-d jackknife does not do particularly well. However, it is interesting to observe that this performance is heavily influenced by the $n = 5$ and $n = 9$ cases where the medians are based on 3 and 5 observa-

Table 3. Confidence interval coverage (Entries are the two-tailed error rates; an asterisk indicates that the entry differs by more than two standard errors from the nominal rate.)

Population, sample size and nominal rate		Method of estimation: MJ	MD	RW	SHAO	SITTER	Wood
Integers 1-25							
$n=5$	5%	7.0*	12.6*	5.2	16.0*		4.0
	10%	9.5	17.6*	8.4	19.3*		10.5
$n=15$	5%	4.7	7.4*	1.8*	6.1	5.3	8.7*
	10%	9.4	13.9*	4.1*	11.9	9.3	9.4
Integers 1-81							
$n=9$	5%	6.7*	11.6*	6.2	11.1*		4.6
	10%	11.6	16.9*	8.5	17.4*		9.6
$n=27$	5%	5.7	8.9*	3.7	6.7*	5.8	4.7
	10%	10.6	13.6*	5.9*	12.2*	9.2	9.3
$n=45$	5%	5.3	6.7*	1.9*	5.8	5.3	5.6
	10%	10.5	13.1*	5.4*	9.9	9.7	8.7
Normal scores 25							
$n=5$	5%	3.2*	7.5*	4.9	15.2*		4.6
	10%	6.1*	12.0*	7.2*	19.5*		11.9
$n=15$	5%	2.8*	5.9	3.3*	6.6*	4.5	7.6*
	10%	7.1*	11.0	7.1*	12.7*	10.0	8.4
Normal scores 81							
$n=9$	5%	4.4	9.3*	4.8	8.3*		4.7
	10%	8.6	14.9*	9.1	14.2*		10.9
$n=27$	5%	5.0	7.7*	4.6	6.7*	5.0	5.8
	10%	9.2	11.7	7.2*	10.2	9.9	9.7
$n=45$	5%	5.2	7.0*	3.6	5.6	3.8	6.7*
	10%	10.9	12.6*	6.8*	11.5	7.6*	10.3
Chi-square 10df. 25							
$n=5$	5%	5.3	11.2*	3.5*	17.5*		5.9
	10%	8.2	18.2*	6.6*	23.5*		9.9
$n=15$	5%	3.4*	7.8*	.7*	4.7	1.3*	9.8*
	10%	5.9*	11.8	2.0*	5.2*	3.2*	10.5
Chi-square 10df. 81							
$n=9$	5%	2.0*	3.9	1.9*	5.8		4.1
	10%	3.5*	7.7*	3.0*	9.9		11.0
$n=27$	5%	3.3*	6.7*	1.7*	3.3*	2.4*	4.5
	10%	7.0*	11.5	5.4*	7.0*	5.2*	8.7
$n=45$	5%	3.7	5.3	4.2	4.9	2.1*	6.0
	10%	7.1*	8.8	5.6*	7.5*	4.7*	9.2*
Chi-square 2df. 25							
$n=5$	5%	4.5	9.7*	3.3*	13.8*		3.4*
	10%	8.7	15.4*	6.3*	17.6*		11.4

(contd)

Table 3. (Contd)

Population, sample size and nominal rate		Method of estimation MJ	MD	RW	SHAO	SITTER	Wood
$n = 15$	5%	3.2*	12.2*	3.2*	3.7	5.2	8.6*
	10%	6.9*	17.4*	7.2*	8.9	11.1	9.8
Chi-square 2df. 81							
$n = 9$	5%	3.5*	7.7*	3.5*	8.5*		3.9
	10%	6.7*	11.8	5.6*	12.8*		9.8
$n = 27$	5%	3.9	7.9*	1.9*	4.1	3.6	5.8
	10%	7.6*	13.5*	4.5*	9.2	6.9*	10.2
$n = 45$	5%	4.7	8.1*	3.0*	7.2*	3.1*	6.2
	10%	9.3	11.8	6.2*	14.0*	5.8*	8.5

tions, respectively. If the $n = 5$ and $n = 9$ cases are removed from the comparison, then $\hat{\sigma}_{WOOD}$ and $\hat{\sigma}_{SHAO}$ have almost identical performances with regard to stability.

3.4. *Confidence intervals*

Confidence intervals were obtained using the methods given in Section 2.2, together with the samples described in Section 3.2. Intervals based on the Woodruff procedure, (2.10), were produced with separate sets of 1000 samples drawn from the population. The two tailed error rates for $\hat{\sigma}_{MJ}$, $\hat{\sigma}_{MD}$, $\hat{\sigma}_{RW}$, $\hat{\sigma}_{SHAO}$, $\hat{\sigma}_{SITTER}$, and Woodruff are given in Table 3 with nominal rates of 5% and 10%. Confidence intervals based on the bootstrap-t procedure are not included because they are so computer intensive that they would not ordinarily be considered in practical applications. The one tailed error rates were also computed, but these were unequal for the two tails for many cases and are not given here. These values are summarized in one possible way in Table 4. It appears that confidence intervals based on the Woodruff procedure are superior to any of the others.

As explained in Section 2.3, confidence intervals can also be obtained from the order statistics of the sample. Thus the closed interval $[y_{(k)}, y_{(n-k+1)}]$ should have a confidence coefficient of $1 - \alpha$ where $\alpha/2$ is given by (2.14). The results of the simulations are given in Table 5, where the entries are the two-tailed percent of intervals that fail to cover. All entries are based on 1000, or more, samples drawn from the populations. It is apparent that the results are not particularly impressive, the overwhelming tendency being that the observed values are smaller than the theoretical values.

4. Simulations Based on Current Population Survey Earnings Data

It seemed desirable that at least some of the simulations described in Section 3 be repeated on real data. Accordingly, Cathryn Dippo of the Bureau of Labor Statistics made arrangements to obtain CPS data from the U.S. Bureau of the Census. These data were for September 1988. After "cleaning," the usable file of records consisted of monthly earnings data from:

10,841 individuals
in 6,936 households
in 2,826 segments.

Table 4. Number of error rates out of 20, which differed significantly from the nominal rates of 5% and 10%

Basis of intervals	5%	10%
MJ	9	9
MD	17	13
RW	12	17
SHAO	12	13
SITTER*	4	6
Wood	6	0

*Out of 12

The basic sampling plan was to select a simple random sample without replacement of $n = 50$ segments. Each sampled segment was enumerated completely. Because of varying segment sizes, different samples provided differing numbers of households and individuals. It was not possible to incorporate stratification into the sampling plan because of the lack of data.

The following notation will be used to describe the population and samples:

Segment i will contain M_i households.

The population will contain $\sum_{i=1}^{N} M_i = M$ households.

Household (i, j) will contain K_{ij} individuals.

Segment i will contain K_i individuals where $\sum_{j=1}^{M_i} K_{ij} = K_i$.

The population will contain $\sum_{i=1}^{N}\sum_{j=1}^{M_i} K_{ij} = \sum_{i=1}^{N} K_i = K$ individuals.

Table 5. Confidence intervals based on order statistics (Entries are two-tailed percentages of intervals that fail to cover)

Sample size and intervals	Integers	Normal scores		Chi-square 10 df	Chi-square 2 df
$n = 5$ $n^* = 5$					
$[y_{(1)}, y_{(5)}]$	3.2	2.5		3.9	3.3
Theory			**5.8**		
$n = 15$ $n^* = 35$					
$[y_{(5)}, y_{(11)}]$	.5	.8		1.6	1.6
Theory			**2.2**		
$[y_{(6)}, y_{(10)}]$	5.7	5.2		12.8	12.7
Theory			**13.5**		
$n = 9$ $n^* = 9$					
$[y_{(2)}, y_{(8)}]$	2.7	3.3		3.3	2.8
Theory			**3.2**		
$[y_{(3)}, y_{(7)}]$	16.1	14.9		15.0	14.7
Theory			**17.5**		
$n = 27$ $n^* = 39$					
$[y_{(10)}, y_{(18)}]$	4.3	4.1		5.6	7.0
Theory			**6.4**		
$[y_{(11)}, y_{(17)}]$	15.3	13.0		16.5	14.2
Theory			**16.9**		
$n = 45$ $n^* = 99$					
$[y_{(19)}, y_{(27)}]$	5.5	6.0		7.1	7.4
Theory			**8.1**		
$[y_{(20)}, y_{(26)}]$	14.2	15.5		17.5	16.8
Theory			**19.2**		

We see that y_{ijk} is the value of the variable (monthly earnings) for the k-th individual in the j-th household in the i-th segment, and that

$y_{ij} = \sum_{k=1}^{K_{ij}} y_{ijk}$ is the value of the variable (monthly earnings) for the j-th household in the i-th segment;

θ_I is the median monthly earnings for all individuals in the population;

θ_H is the median monthly earnings for all households in the population;

$\hat{\theta}_I$ is the median earnings for individuals in a sample of 50 segments which is used to estimate θ_I;

and

$\hat{\theta}_H$ is the median earnings for households in a sample of 50 segments which is used to estimate θ_H.

The goal of this investigation is to determine the performance of several different methods that can be used to estimate, from a sample, the Root Mean Squared Errors of $\hat{\theta}_I$ and $\hat{\theta}_H$ as estimators of θ_I and θ_H and to determine confidence intervals for these parameters.

Preparatory to the simulation study the values of θ_I and θ_H were computed directly. The values of $RMSE_{\hat{\theta}_I}$ and $RSME_{\hat{\theta}_H}$ were approximated by drawing 25,000 simple random samples of 50 segments each and computing

$$RMSE_{\hat{\theta}_I} = \left(\frac{\sum_{\alpha=1}^{25,000} (\hat{\theta}_{I,\alpha} - \theta_I)^2}{25,000} \right)^{1/2}$$

and

$$RMSE_{\hat{\theta}_H} = \left(\frac{\sum_{\alpha=1}^{25,000} (\hat{\theta}_{H,\alpha} - \theta_H)^2}{25,000} \right)^{1/2}.$$

For simplicity these will be denoted as $\sigma_{\hat{\theta}_I}$ and $\sigma_{\hat{\theta}_H}$.

A number of different approaches were used to estimate the values of $\sigma_{\hat{\theta}_I}$ and $\sigma_{\hat{\theta}_H}$ from a sample, and to estimate confidence intervals for θ_I and θ_H. These were variations on some of the methods described in Section 2. The approaches used were as follows:

1. *The McCarthy-Snowden method for determining the bootstrap sample size as given by expression (2.1).* In the present instance this means that a simple random sample of $n = 50$ segments is drawn without replacement from the population of segments, with medians $\hat{\theta}_I$ and $\hat{\theta}_H$. From this sample, a random sample (the bootstrap sample) of $n^* = 50$ segments is drawn with replacement. The medians are denoted by $\hat{\theta}^*_{I,1}$ and $\hat{\theta}^*_{H,1}$. This process is repeated 200 times and the standard errors are estimated as

$$\hat{\sigma}^*_I = \left(\frac{\sum_{\alpha=1}^{200} (\hat{\theta}^*_{I,\alpha} - \hat{\theta}^*_I)^2}{199} \right)^{1/2}$$

and

$$\hat{\sigma}^*_H = \left(\frac{\sum_{\alpha=1}^{200} (\hat{\theta}^*_{H,\alpha} - \hat{\theta}^*_H)^2}{199} \right)^{1/2}. \qquad (4.1)$$

Confidence intervals for θ_I and θ_H are estimated as

$$\hat{\theta}_I - z_{\alpha/2}\hat{\sigma}^*_I,\ \hat{\theta}_I + z_{\alpha/2}\hat{\sigma}^*_I$$

and

$$\hat{\theta}_H - z_{\alpha/2}\hat{\sigma}^*_H,\ \hat{\theta}_H + z_{\alpha/2}\hat{\sigma}^*_H. \qquad (4.2)$$

It was then determined whether or not θ_I and θ_H fell within these intervals.

A total of 1000 independent samples were drawn from the population of segments and the results were summarized as

a. The average values of the bootstrap standard error estimates, $\overline{\hat{\sigma}^*_I}$ and $\overline{\hat{\sigma}^*_H}$.

b. The standard errors of the bootstrap standard error estimates, s.e. $(\hat{\sigma}_I^*)$ and s.e. $(\hat{\sigma}_H^*)$.

c. The fraction of times that θ_I was less than $\hat{\theta}_I - z_{\alpha/2}\hat{\sigma}_I^*$ and the fraction of times that θ_I was greater than $\hat{\theta}_I + z_{\alpha/2}\hat{\sigma}_I^*$. Similar fractions were obtained for θ_H.

2. *The Rao-Wu transformation approach as defined by expression (2.5).* It had been planned to use this expression with $n = 50$ segments, $m = 50$ and $\bar{y}$ replaced by either

$$\hat{\hat{\bar{Y}}}_I = \left(\sum_{i=1}^{n}\sum_{j=1}^{K_i}\sum_{k=1}^{K_{ij}} y_{ijk}\right) \Big/ \sum_{i=1}^{n} K_i$$

or

$$\hat{\hat{\bar{Y}}}_H = \left(\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}\right) \Big/ \sum_{i=1}^{n} M_i .$$

Unfortunately, the squared root factor in (2.5) becomes approximately equal to one with this choice of n and m so that the Rao-Wu approach reduces to the McCarthy-Snowden approach.

3. *The Woodruff method.* This procedure, with simple random sampling, for determining confidence intervals for a population median is defined by expressions (2.9), (2.10), and (2.11). In the present instance, which is based on cluster sampling of individuals and households, it is necessary to replace the squared root expression in (2.10) by one that is appropriate for this type of sampling. Briefly, the argument is as follows:

a. A simple random sample of $n = 50$ segments is drawn without replacement from the population of segments. The median for individuals is $\hat{\theta}_I$.

b. Let a_{Ii} be the number of individuals in segment i that have values of the variable less than or equal to $\hat{\theta}_I$. Then the squared root expression in (2.10) is replaced by

$$\left[(1-f)n^{-1}\left(\sum_{i=1}^{n} K_i/n\right)^{-2}(n-1)^{-1} \times \left(\sum_{i=1}^{n}(a_{Ii} - .5K_i)^2\right)\right]^{1/2}. \qquad (4.3)$$

The rationale for using this expression is explained in Cochran (1977, p. 66).

c. Exactly the same steps are followed for households with $\hat{\theta}_H$, a_{Hi}, and M_i.

The preceding steps lead to confidence intervals for θ_I and θ_H, the error rates for these confidence intervals, and by (2.17) to standard error estimates. These estimates were averaged over 1000 samples of $n = 50$ drawn independently from the population of segments.

4. *Balanced Repeated Replications.* Strictly speaking, the method of repeated replications, McCarthy (1966), is defined when one is using a stratified simple random sample with two units chosen from each stratum. A half sample is determined by selecting one of the two units from each stratum. This method was not used in Section 3 with the artificial populations. It is used here because of its wide applicability when complex sample designs are necessary. In the present instance, where stratified sampling is not used, we have approximated this situation by randomly pairing the 50 segments into 25 pairs. A set of 28 orthogonal half samples was then obtained by using the last 25 columns in the 28×28 Hadamard matrix given by Wolter (1985, p. 324).

If $\hat{\theta}_{Ii}$ is the median earnings for individuals in the i-th half sample and $\hat{\theta}_{Ii,C}$ is the median for the complementary i-th half sample, then the four estimates of the standard error of $\hat{\theta}_I$ are given by

$$\hat{\sigma}_{BRR-H} = \left(\sum_{i=1}^{28}(\hat{\theta}_{Ii} - \hat{\theta}_I)^2/28\right)^{1/2}$$

$$\hat{\sigma}_{BRR-C} = \left.\sum_{i=1}^{28}(\hat{\theta}_{Ii,C} - \hat{\theta}_I)^2/28\right)^{1/2}$$

$$\hat{\sigma}_{BRR\text{-Full}} = ((\hat{\sigma}^2_{BRR-H} + \hat{\sigma}^2_{BRR-C})/2)^{1/2}$$

$$\hat{\sigma}_{BRR\text{-Difference}} = \left(\sum_{i=1}^{28}(\hat{\theta}_{Ii} - \hat{\theta}_{Ii,C})^2/4 \times 28\right)^{1/2}. \qquad (4.4)$$

Similar expressions can be defined for household earnings. Confidence intervals are obtained by the ordinary normal assumptions as exemplified by (2.15).

A total of 1000 samples were drawn independently from the population of segments, each containing 50 segments drawn without replacement. The results were summarized as:

a. The average values of the four BRR standard error estimates.

b. The standard errors of the BRR standard error estimates.

c. The fraction of times that the confidence intervals failed to cover the population medians.

These results were obtained for both individuals and households.

The data from the simulations that have just been described for the Bootstrap, Woodruff and BRR procedures are summarized in Tables 6, 7 and 8. Table 6 gives a comparison of the standard error estimates; Table 7 provides a comparison of the stabilities of these standard error estimates; and Table 8 contains the confidence interval summaries. As is apparent from Table 6, there are no great differences among the biases of the standard error estimates for the median earnings for individuals and households from the Bootstrap, Woodruff and BRR methods. From Table 7 we see that the standard error estimates are somewhat less stable for the BRR methods, except for the Difference procedure. Table 8 contains the coverage errors of the confidence intervals for the population median. Most of these are larger than the nominal rates with none excessively out of line. The most consistent over coverages are for the BRR methods, the .05 level and for households.

Table 6. Comparison of standard error estimates for current population survey earnings data (Entries are (estimated standard error)/(the true standard error))

Method of estimation	Individuals	Households
Bootstrap (McCarthy-Snowden)	1.02	1.02
Woodruff		
$\alpha = .01$	.99	1.02
$\alpha = .05$	.99	1.00
$\alpha = .10$	.99	.99
BRR		
Half	1.03	1.02
Complement	1.03	1.01
Full	1.03	1.02
Difference	.97	.96

Table 7. Relative stabilities of standard error estimates for current population survey earnings data (Entries are (standard error of standard error estimates)/(true standard error))

Method of estimation	Individuals	Households
Bootstrap		
(McCarthy-Snowden)	.231	.236
Woodruff		
$\alpha = .01$	.205	.204
$\alpha = .05$	.224	.219
$\alpha = .10$	.235	.237
BRR		
Half	.243	.259
Complement	.252	.254
Full	.241	.249
Difference	.233	.236

5. Summary

This paper has collected together a wide variety of procedures that can be used to produce standard error estimates and confidence intervals for the median of a finite population when simple random samples are drawn without replacement. All these procedures are evaluated by drawing repeated samples from a number of small artificial finite populations. In addition, a number of the procedures are evaluated by drawing repeated samples from a relatively large clustered population derived from a Current Population Survey sample.

For each population, sample size, and estimation procedure, we determined an estimate of the standard error of the median, the stability of the standard error estimate and confidence intervals for several different confidence coefficients. Through an oversight, the standardized length of the intervals was not determined. For the artificial populations, the principal results are

1. The standard error estimate, defined in expression (2.4) and based on absolute deviations, has the smallest bias.
2. The standard error estimate based on the Woodruff confidence intervals has the best stability.

Table 8. Confidence interval coverage for current population survey earnings data (Entries are the two-tailed error rates; an asterisk indicates the entry differs by more than two standard errors from the nominal rate.)

Nominal Rate	Individuals			Households		
Method of estimation	.01	.05	.10	.01	.05	.10
Bootstrap						
(McCarthy-Snowden)		.058	.092		.067*	.112
Woodruff	.011	.039	.077	.018*	.059	.124*
BRR						
Half		.062	.101		.081*	.114
Complement		.063	.100		.077*	.119
Full		.062	.097		.073*	.116
Difference		.073*	.113		.085*	.131*

3. The confidence intervals produced by the Woodruff procedure generally seem to be closest to the numerical values.

A summary for the simulations based on the population derived from a Current Population Survey is given at the end of Section 4.

It should be observed that many of the procedures discussed in this paper can be adapted to more complex survey procedures. Some of these adaptations are described by Kovar, Rao, and Wu (1988) and Francisco and Fuller (1991).

6. References

Bickel, P.J. and Freedman, D.A. (1981). Some Asymptotic Theory for the Bootstrap. Annals of Statistics, 9, 1196–1217.

Cochran, W.G. (1977). Sampling Techniques. 3rd edition, New York: John Wiley and Sons.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: Society for Industrial and Applied Mathematics, Monograph No. 38.

Falk, M. and Kaufmann, E. (1991). Coverage Probabilities of Bootstrap-Confidence Intervals for Quantiles. Annals of Statistics, 19, 485–495.

Francisco, C.A. and Fuller, W.A. (1991). Quantile Estimation with a Complex Survey Design. Annals of Statistics, 19, 454–469.

Gross, S.T. (1980). Median Estimation in Sample Surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 181–184.

Hall, P. and Sheather, S.J. (1988). On the Distribution of a Studentized Quantile. Journal of the Royal Statistical Society B, 50, 381–391.

Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. (1988). Bootstrap and Other Methods to Measure Error in Survey Estimation. The Canadian Journal of Statistics, 16, Supplement, 25–45.

Maritz, J.S. and Jarrett, R.G. (1978). A Note on Estimating the Variance of the Sample Median. Journal of the American Statistical Association, 73, 194–196.

McCarthy, P.J. (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. National Center for Health Statistics, Series 2, No. 14.

McCarthy, P.J. and Snowden, C.B. (1985). The Bootstrap and Finite Population Sampling. In Vital and Health Statistics (Ser. 2, No. 95), Public Health Service Publications 85–1369, Washington D.C.: U.S. Government Printing Office.

McKean, J.W. and Schrader, R.M. (1984). A Comparison of Methods for Studentizing the Sample Median. Communications in Statistics B, 13, 751–773.

Miller, R.G. (1974). The Jackknife. A Review. Biometrika, 61, 1–15.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling Inference with Complex Survey Data. Journal of the American Statistical Association, 83, 231–241.

Shao, J. (1989). The Efficiency and Consistency of Approximations to the Jackknife Variance Estimators. Journal of the American Statistical Association, 84, 114–119.

Shao, J. and Wu, C.F.J. (1986). Some General Theory for the Jackknife. Technical Report 797, University of Wisconsin-Madison, Dept. of Statistics.

Sitter, R. (1990). A Resampling Procedure for Complex Survey Data. Technical Report 149, Dept. of Mathematics and Statistics, Carleton University, Ottawa, Canada.

Wilks, S.S. (1962). Mathematical Statistics. New York: John Wiley and Sons.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Woodruff, R.A. (1952). Confidence Intervals for Medians and Other Position Measures. Journal of the American Statistical Association, 47, 635–646.

Received February 1992
Revised February 1993