Bootstrap Methods for Finite Populations

James G. BOOTH, Ronald W. BUTLER, and Peter HALL*

We show that the familiar bootstrap plug-in rule of Efron has a natural analog in finite population settings. In our method a characteristic of the population is estimated by the average value of the characteristic over a class of empirical populations constructed from the sample. Our method extends that of Gross to situations in which the stratum sizes are not integer multiples of their respective sample sizes. Moreover, we show that our method can be used to generate second-order correct confidence intervals for smooth functions of population means, a property that has not been established for other resampling methods suggested in the literature. A second resampling method is proposed that also leads to second-order correct confidence intervals and is less computationally intensive than our bootstrap. But a simulation study reveals that the second method can be quite unstable in some situations, whereas our bootstrap performs very well.

KEY WORDS: Confidence interval; Edgeworth expansion; Empirical population; Plug-in rule; Resample; Second-order correct; Subsample; Survey data.

1. INTRODUCTION

In this article we show that the now-familiar nonparametric bootstrap method (Efron 1982) for approximating characteristics of a unknown distribution has a natural analog in classical finite-population sampling problems. In particular, we show that there exist second-order correct bootstrap estimates for the distribution of Studentized versions of stratified sample means, separate ratio estimates, and other well-known estimates of a finite population mean. Inversion of the bootstrap distribution function leads to second-order correct, percentile-t bootstrap confidence intervals that can be considerably more accurate than the standard intervals based on the normal approximation, especially in small-scale surveys. (See Hall 1988 for an in-depth discussion of the percentile-t method in the infinite population setting.)

In common with the infinite population case is the fact that exact computation of bootstrap estimates is rarely feasible. In Section 2 we describe a Monte Carlo technique for approximating bootstrap estimates and constructing bootstrap confidence intervals that involves resampling from empirical populations. In Section 3 we discuss two alternative approaches to estimation in finite-population settings that also require a form of resampling. The first of these methods, due to Bickel and Freedman (1984), is not second-order correct but in practice often gives answers close to our bootstrap, particularly when the sampling fractions are small. This method was also suggested independently by Chao and Lo (1985). Both our bootstrap method and the Bickel and Freedman (BF) method reduce to a resampling procedure suggested by Gross (1980) when the ratios of stratum sizes to sample sizes are all integers. The second alternative method involves estimating the distribution of a statistic by the distribution of its version based on subsamples selected at random and without replacement from the observed stratum samples, with the subsampling fractions as close as possible to the original sampling fractions. The resulting estimate is not second-order correct, but a modified estimate obtained by mixing the subsampling estimate with the standard normal distribution is second-order correct. This procedure is effective when the sampling fractions are large, which is often the case in small-scale surveys, but can perform badly when the sampling fractions are small. In Section 4 we summarize the results of a simulation study in which the three resampling techniques and the normal approximation are compared for various standard estimates. We discuss technical issues and give an outline of the proof of our assertions concerning second-order correctness in Section 5. The key theoretical tools required for this discussion are theorems 1 and 2 of Babu and Singh (1985) on Edgeworth expansions in finite-population sampling settings. We conclude in Section 6 with a brief discussion of our results. Other related work involving resampling methods in finite-population settings includes that of Kover, Rao, and Wu (1988), Krewski and Rao (1981), Lo (1988), McCarthy and Snowdon (1985), Rao and Wu (1988), and Sitter (1992). We conclude this section with a more detailed description of our method and introduce some notation.

In the infinite population setting, the bootstrap method may be described as a *plug-in rule,* whereby a functional $\theta = t(F)$ of an unknown distribution function F is estimated by $\hat{\theta} = t(\hat{F})$; that is, by plugging in the empirical distribution function $\hat{F}$ based on a random sample from F. In finite-population settings, the role of the unknown distribution function F is played by the set $\mathcal{P} = \{x_1, \ldots, x_N\}$ of numerical measurements on elements of a population of size N, where the measurement x_i corresponding to the ith element of the population may consist of k ($k \geq 1$) components. In an abuse of the usual sampling terminology, we shall refer to $\mathcal{P}$ as the population.

Suppose that $\mathcal{P}$ is divided into s ($s \geq 1$) mutually exclusive and exhaustive strata $\mathcal{P}_1, \ldots, \mathcal{P}_s$ of sizes $N_1, \ldots, N_s$. We write $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_s\}$ and use the notation $\mathcal{P}_r = \{x_{r,1}, \ldots, x_{r,N_r}\}$ to distinguish elements of the rth stratum. For each $r = 1, \ldots, s$ let $\mathcal{X}_r = \{X_{r,1}, \ldots, X_{r,n_r}\}$ be a sample of size $n_r \leq N_r$ selected at random and without replacement from $\mathcal{P}_r$, let $n. = n_1 + \cdots + n_s$, $p. = p_1 + \cdots + p_s$ and $q. = q_1 + \cdots + q_s$, where $p_r = n_r/N_r$ is the sampling fraction for the rth stratum and $q_r = 1 - p_r$. Write $\mathcal{X} = \{\mathcal{X}_1, \ldots,$

* James G. Booth is Associate Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Peter Hall is Professor, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 2601, Australia. Ronald W. Butler is Professor, Statistics Department, Colorado State University, Fort Collins CO 80523. The authors would like to thank two referees for their detailed comments that considerably improved the presentation of material in the article. The first author received support from National Science Foundation Grant DMS-93010836.

Journal of the American Statistical Association
December 1994, Vol. 89, No. 428, Theory and Methods

$\mathcal{X}_s\}$ for the complete sample. Let $E(\cdot|\mathcal{P})$ and $P(\cdot|\mathcal{P})$ denote the expectation and probability under this sampling scheme.

Now, suppose that $\theta = t(\mathcal{P})$ is a characteristic of the population $\mathcal{P}$ that is of interest. The natural analog of the infinite population bootstrap plug-in rule in this setting is to estimate θ by $\hat{\theta} = t(\hat{\mathcal{P}})$, where $\hat{\mathcal{P}}$ is an *empirical population* based on the sample $\mathcal{X}$. In the case where N_r/n_r is an integer, say m_r, for each $r = 1, \ldots, s$, an obvious candidate for $\hat{\mathcal{P}}$ is obtained by replicating $\mathcal{X}_r m_r$ times, giving say $\hat{\mathcal{P}}_r$, and then letting $\hat{\mathcal{P}} = \{\hat{\mathcal{P}}_1, \ldots, \hat{\mathcal{P}}_s\}$. This procedure was suggested by Gross (1980). More generally, however, we let m_r be the integer part of N_r/n_r and $k_r = N_r - n_r m_r$. Then for each $r = 1, \ldots, s$ form an empirical stratum $\hat{\mathcal{P}}_r^*$ by combining m_r replicates of $\mathcal{X}_r$ with a sample of size k_r, selected at random and without replacement from $\mathcal{X}_r$. We then call $\hat{\mathcal{P}}^* = \{\hat{\mathcal{P}}_1^*, \ldots, \hat{\mathcal{P}}_s^*\}$ an empirical population based on $\mathcal{X}$ and define the bootstrap estimate of θ by

$$\hat{\theta} = E\{t(\hat{\mathcal{P}}^*)|\mathcal{X}\}, \tag{1}$$

where $E\{\cdot|\mathcal{X}\}$ denotes expectation conditional on $\mathcal{X}$.

We focus in this article on estimating the distribution function of a Studentized estimate of a population mean μ. A confidence interval for μ can be constructed from the distribution function estimate by the usual inversion technique. Specifically, let $T = (n./q.)^{1/2}(\hat{\mu} - \mu)/\hat{\sigma}$, where $\hat{\mu}$ is an estimate such as a stratified sample mean or separate ratio estimate and $(q./n.)^{1/2}\hat{\sigma}$ is an estimate of the standard error of $\hat{\mu}$ under repeated sampling from $\mathcal{P}$. Then let $\theta = P(T \le y|\mathcal{P})$. It turns out that under certain regularity conditions, the bootstrap estimate $\hat{\theta}$ given by (1) approximates θ to second order in the sense that $\hat{\theta} = \theta + o_p(n.^{-1/2})$. In comparison, the normal approximation $\Phi(y)$ deviates from θ by a term of size $O(n.^{-1/2})$ under the same conditions. As with its infinite-population counterpart, our bootstrap method comes into its own with small- to moderate-sized samples. Thus we envisage our procedure being most useful in small-scale surveys. If $n.$ is large and the sampling fractions are not too close to 1, then the normal approximation will usually be adequate. In the latter situation, bootstrap methods and the normal approximation will typically yield very similar answers. The beauty of the bootstrap approach is that it adapts to the problem as required, leading to accurate inferences when the normal approximation works well and also when it does not.

Formula (1) can also be used to estimate the variance of $\hat{\mu}$ by letting $\theta = \text{var}(\hat{\mu}|\mathcal{P})$. As in the infinite population case, bootstrap variance estimates are biased, even for linear estimates; the bootstrap variance estimate typically have expectation $1 + O(n.^{-1})$ times the exact variance. For example, in the single-stratum case, when the population size is an integer multiple of the sample size ($N = nm$), the bootstrap estimate of the variance of the sample mean is $(n-1)/(n-1/m)$ times the usual unbiased estimate (McCarthy and Snowdon 1985, p. 2). Because of this, other resampling schemes may be preferred for variance estimation, particularly when some or all of the stratum sample sizes are very small. For example, the "rescaling" and "mirror-match" resampling methods proposed by Rao and Wu (1988) and Sitter (1992) reproduce the usual unbiased variance estimate in the linear case.

2. MONTE CARLO APPROXIMATION

As noted in Section 1, exact computation of the bootstrap estimate $\hat{\theta}$ in (1) is usually not practical. In this section we describe a simple Monte Carlo technique for approximating $\hat{\theta}$. We consider in particular the case where $\theta = P(T \le y|\mathcal{P})$ and T is a Studentized estimate of a population mean μ, and discuss the concentration of bootstrap confidence intervals for μ by the inversion method. Let $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_s)$ and $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_s)$, where $\bar{\mathbf{x}}_r$ and $\bar{\mathbf{X}}_r$ are the mean vectors corresponding to the rth stratum $\mathcal{P}_r$ and its associated sample $\mathcal{X}_r$. Suppose that $\hat{\mu} = \hat{\mu}(\bar{\mathbf{X}}; \bar{\mathbf{x}})$ is an estimate of $\mu = \mu(\bar{\mathbf{x}})$ and that $(q./n.)^{1/2}\hat{\sigma}$, where $\hat{\sigma} = \hat{\sigma}(\bar{\mathbf{X}}; \bar{\mathbf{x}})$ is an estimate of the standard error of $\hat{\mu}$ under repeated sampling from $\mathcal{P}$. Define $T = (n./q.)^{1/2}(\hat{\mu} - \mu)/\hat{\sigma}$. Notice that by allowing $\hat{\mu}$ to be a function of the population vector $\bar{\mathbf{x}}$ as well as the sample vector $\bar{\mathbf{X}}$, we include in our discussion estimates that use knowledge of the stratum means of one or more auxiliary variables. For example, suppose that the population consists of paired measurements, (x_{1i}, x_{2i}) $i = 1, \ldots, N$, and that $\mu = \bar{x}_1$ is the characteristic of interest. If the value of $\bar{x}_2$ is known exactly, then the ratio estimate of μ is given by $\hat{\mu} = (\bar{X}_1/\bar{X}_2)\bar{x}_2$. In this instance, the variance estimate obtained by linearization (Cochran 1977, sec. 6.4) is given by

$$(n/q)\hat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_{1i} - \hat{\mu})^2/\bar{x}_2^2.$$

Suppose that $\hat{\mathcal{P}}^*$ is an empirical population constructed in the manner described in Section 1. For $r = 1, \ldots, s$, let $\mathcal{X}_r^*$ be a sample of size n_r, selected at random and without replacement from the rth empirical stratum $\hat{\mathcal{P}}_r^*$. Then we call $\mathcal{X}^* = \{\mathcal{X}_1^*, \ldots, \mathcal{X}_s^*\}$ a *resample* from $\hat{\mathcal{P}}^*$. Let $\bar{\mathbf{x}}^*$ and $\bar{\mathbf{X}}^*$ denote the mean vectors corresponding to $\hat{\mathcal{P}}^*$ and $\mathcal{X}^*$ and define $T^* = (n./q.)^{1/2}(\hat{\mu}^* - \mu^*)/\hat{\sigma}^*$, where $\hat{\mu}^* = \hat{\mu}(\bar{\mathbf{X}}^*; \bar{\mathbf{x}}^*)$, $\mu^* = \mu(\bar{\mathbf{x}}^*)$, and $\hat{\sigma}^* = \hat{\sigma}(\bar{\mathbf{X}}^*; \bar{\mathbf{x}}^*)$. Observe that $\hat{\theta} = P(T^* \le y|\mathcal{X})$, where $P(\cdot|\mathcal{X})$ denotes probability conditional on $\mathcal{X}$.

Let $\hat{\mathcal{P}}_{(1)}^*, \ldots, \hat{\mathcal{P}}_{(B)}^*$ be B empirical populations constructed independently in the manner described in Section 1. For each $b = 1, \ldots, B$ let $\mathcal{X}_{(b,1)}^*, \ldots, \mathcal{X}_{(b,C)}^*$ be a collection of C independently selected resamples from $\hat{\mathcal{P}}_{(b)}^*$. An unbiased Monte Carlo approximation to $\hat{\theta}$ is then given by

$$\tilde{\theta} = (BC)^{-1}\sum_{b=1}^{B}\sum_{c=1}^{C} I\{T_{(b-1)C+c}^* \le y\}, \tag{2}$$

where $T_{(b-1)C+c}^*$ is the version of T computed using the empirical population $\hat{\mathcal{P}}_{(b)}^*$ and resample $\mathcal{X}_{(b,c)}^*$ and $I(\cdot)$ is the indicator set function.

We assume that the lack of continuity of the distributions of T in repeated sampling from $\mathcal{P}$ and T^* conditional on $\mathcal{X}$ is negligible. Then an "exact" α-level confidence interval for μ is given by

$$(\hat{\mu} - (q./n.)^{1/2}\hat{\sigma}y_{(1+\alpha)/2}, \hat{\mu} - (q./n.)^{1/2}\hat{\sigma}y_{(1-\alpha)/2}), \tag{3}$$

where y_β denotes the β quantile of the distribution of T satisfying $\beta = P(T \le y_\beta|\mathcal{P})$ for $0 < \beta < 1$. The bootstrap estimate of the interval in (3) is obtained by replacing exact quantiles y_β by their bootstrap approximations $\hat{y}_\beta$ satisfying $\beta = P(T^* \le \hat{y}_\beta | \mathcal{X})$. Finally, a Monte Carlo approximation to $\hat{y}_\beta$ is given by $\tilde{y}_\beta = T^*_{(R\beta)}$, where $R = B \times C$ and $T^*_{(a)}$ denotes the jth order statistic of $T^*_1, \ldots, T^*_R$, where j is the closest integer in the set $\{1, \ldots, B\}$ to the value a.

3. TWO ALTERNATIVE METHODS

Bickel and Freedman (1984) have proposed an alternative resampling procedure for situations in which the ratios of stratum sizes to samples sizes, N_r/n_r, $r = 1, \ldots, s$ are not all integers. The method differs from the one that we proposed in Section 1 by the way in which empirical populations are constructed. More specifically, Bickel and Freedman suggested letting the rth empirical stratum, say $\hat{\mathcal{P}}'_r$, consist of m_r replicates of $\mathcal{X}_r$ with probability π_r $(0 < \pi_r \le 1)$ and $m_r + 1$ replicates of $\mathcal{X}_r$ with probability $1 - \pi_r$. Let $\hat{\mathcal{P}}' = \{\hat{\mathcal{P}}'_1, \ldots, \hat{\mathcal{P}}'_s\}$ denote the resulting empirical population. A resample $\mathcal{X}'$ from $\hat{\mathcal{P}}'$ then consists of s samples of sizes $n_1, \ldots, n_s$, selected randomly and without replacement from the empirical strata $\hat{\mathcal{P}}'_1, \ldots, \hat{\mathcal{P}}'_s$. The values of π_r, $r = 1, \ldots, s$, are chosen so that the average resampling fraction in the rth stratum equals the true value p_r; that is, π_r satisfies $p_r = \pi_r/m_r + (1 - \pi_r)/(m_r + 1)$. Let $\bar{\mathbf{x}}'$ and $\bar{\mathbf{X}}'$ denote the empirical population and resample stratum mean vectors and define $T' = (n./q.)^{1/2}(\hat{\mu}' - \mu')/\hat{\sigma}'$, where $\mu' = \mu(\bar{\mathbf{x}}')$, $\hat{\mu}' = \hat{\mu}(\bar{\mathbf{X}}'; \bar{\mathbf{x}}')$, and $\hat{\sigma}' = \hat{\sigma}(\bar{\mathbf{X}}'; \bar{\mathbf{x}}')$. Then the BF estimate of $\theta = P(T \le y | \mathcal{P})$ is $\hat{\theta}' = P(T' \le y | \mathcal{X})$. Monte Carlo approximation of $\hat{\theta}'$ and of $\hat{y}'_\beta$, the corresponding β-quantile estimate, proceeds as in Section 2. Notice that if $m_r = 1$ for some value of r, then the corresponding empirical stratum will sometimes be the same as the stratum sample. Hence there is potentially no resampling variability in such strata, with the result that the estimate of θ may be extremely poor.

A second alternative procedure involves *subsampling* from the original sample $\mathcal{X}$. For $r = 1, \ldots, s$, let n°_r denote the integer part of $n_r p_r$, so that $n^\circ_r/n_r \approx p_r$, and let $\mathcal{X}^\circ_r$ denote a sample of size n°_r, selected at random and without replacement from $\mathcal{X}_r$. Write $\mathcal{X}^\circ = \{\mathcal{X}^\circ_1, \ldots, \mathcal{X}^\circ_s\}$ and let T° denote the version of T computed using $\mathcal{X}^\circ$ and $\mathcal{X}$ in place of $\mathcal{X}$ and $\mathcal{P}$. Then we call $P(T^\circ \le y | \mathcal{X})$ the subsampling method estimate of θ. As pointed out in Section 1, this subsampling estimate is not second-order correct. But this defect can be rectified by mixing with the standard normal distribution function to give a modified subsampling estimate,

$$\hat{\theta}^\circ = (n^\circ./n.)^{1/2} P(T^\circ \le y | \mathcal{X}) + \{1 - (n^\circ./n.)^{1/2}\}\Phi(y), \tag{4}$$

where $n^\circ. = n^\circ_1 + \cdots + n^\circ_s$. A modified subsampling estimate, say $\hat{y}^\circ_\beta$, for the quantile y_β is defined in the obvious way via the inverse of (4).

Monte Carlo approximation of $\hat{\theta}^\circ$ and $\hat{y}^\circ_\beta$ proceeds by first selecting R independent subsamples $\mathcal{X}^\circ_{(1)}, \ldots, \mathcal{X}^\circ_{(R)}$ from $\mathcal{X}$. Let T°_j denote the value of T° computed using $\mathcal{X}^\circ_{(j)}$ and $\mathcal{X}$, $j = 1, \ldots, R$. Then an unbiased Monte Carlo approximation to $P(T^\circ \le y | \mathcal{X})$ is given by $R^{-1} \sum I(T^\circ_j \le y)$, where the sum is over the range 1 to R. Substitution of this value into (4) results in an unbiased approximation to $\hat{\theta}^\circ$, say $\tilde{\theta}^\circ$. A Monte Carlo approximation to $\hat{y}^\circ_\beta$ is $\tilde{y}^\circ_\beta = T^\circ_{(j_\beta)}$, where

$$j_\beta = \min[j : j \in \{1, \ldots, R\} \text{ and } (n^\circ./n.)^{1/2} j/R + \{1 - (n^\circ./n.)^{1/2}\}\Phi\{T^\circ_{(j)}\} \ge \beta]$$

and $T^\circ_{(j)}$ denotes the jth order statistic of $T^\circ_1, \ldots, T^\circ_R$.

We conclude this section with a numerical illustration of the various methods involving the estimation of the total population of 196 large U.S. cities in 1930. The data, taken from Cochran (1977, p. 152), consist of the 1920 and 1930 populations (in 1,000s) of a random sample of 49 of the 196 cities. The total 1920 population of all 196 cities is also given, as 22,919 thousand. The 1920 and 1930 sample totals are 5,054 and 6,262, leading to a ratio estimate of 28,397 thousand for the 1930 total. The standard error of this estimate if 604, and hence a nominal 95% confidence intervals for the 1930 total based on the normal approximation is (27,213, 29,581) thousand. The corresponding bootstrap and subsampling intervals, (27,216, 29,702) and (27,165, 29,943) are considerably longer, particularly in the upper tail. Note that the BF method is identical to the bootstrap in this problem, because the population size is exactly four times the sample size. The discrepancy between the bootstrap and subsampling intervals in this example is discussed in Section 6.

4. A SIMULATION STUDY

The sampling distributions for T^*, T', and T° were simulated and the resulting quantiles used as previously described in the construction of confidence intervals and estimation of quantiles from the distribution to T. Table 1 summarizes coverage attainment for the three bootstrap intervals as well as for the usual normal approximation intervals. Table 2 addresses quantile estimation accuracy for the finite-population bootstrap method (BT) and the BF method, and Table 3 contains summary statistics for the half-widths of the various confidence intervals.

The collective information in the tables demonstrates that BT and BF perform best overall in terms of achieving nominal coverage accuracy and accuracy in estimation of quantiles for the distribution of T. Method BT appears to be slightly better for such quantile estimation. Subsampling is the next best method when the subsampling sizes are not too small. All three bootstrap methods were preferable to the normal approximations, however, even in situations where the finite populations were constructed as realizations of normal samples.

The study is based on simulated bivariate populations of correlated (x_1, x_2) pairs with two strata. We considered smaller populations with strata sizes $N_1 = 50$ and $N_2 = 30$, from which samples of sizes $n_1 = 20 = n_2$ were drawn, and larger populations with strata sizes $N_1 = 100$ and $N_2 = 50$ and sample sizes $n_1 = 15$ and $n_2 = 11$. These population sizes were considered in combination with two different distributions for generating the finite populations: correlated chi-squared with 15 degrees of freedom and correlated bi-

Table 1. Empirical Coverage Percentages of Confidence Intervals

	Smaller population $N_1 = 50, N_2 = 30$ $n_1 = 20, n_2 = 20$				Larger population $N_1 = 100, N_2 = 50$ $n_1 = 15, n_2 = 11$			
Estimator	*BT*	*BF*	*SB*	*Z*	*BT*	*BF*	*SB*	*Z*
				Chi-squared populations				
Strat Est	93.0[a]	92.8	94.5	90.7	92.9	92.0	99.4	88.3
	87.7[b]	87.5	89.3	86.1	87.8	86.7	95.4	83.7
Sep Ratio	96.6	97.4	98.6	92.6	96.6	97.0	99.9	89.6
	92.6	92.8	95.0	87.6	92.0	93.1	98.7	84.3
Sep Tin	96.4	97.2	98.7	93.4	96.0	96.6	100.0	90.5
	92.2	92.2	95.1	88.8	91.3	92.3	99.0	85.4
Com Ratio	96.2	96.8	95.9	93.0	96.2	96.7	99.6	90.3
	91.4	92.0	91.8	88.0	91.2	92.5	97.5	85.1
Sep Regr	93.6	93.7	96.1	88.4	94.0	93.8	—[c]	86.3
	89.4	89.1	91.6	82.8	89.9	90.2	—[c]	80.6
Com Regr	93.6	94.1	94.5	88.2	93.8	93.6	100.0	86.0
	89.0	89.4	89.3	82.3	89.2	89.2	99.9	80.2
				Normal linear regression populations				
Strat Est	95.4	95.4	95.4	93.8	95.6	95.5	98.6	93.3
	91.1	90.4	90.3	88.6	91.1	91.6	95.3	88.3
Sep Ratio	95.0	95.2	95.2	94.0	95.6	95.2	98.9	93.7
	90.8	90.2	90.2	89.0	91.2	89.8	95.7	88.6
Sep Tin	95.1	95.2	95.2	94.0	95.6	95.2	98.9	93.7
	90.8	90.2	90.3	89.0	91.3	89.8	95.7	88.6
Com Ratio	95.1	95.2	95.2	94.0	95.6	95.0	99.0	93.7
	90.7	90.2	90.2	89.1	91.2	89.8	95.6	88.6
Sep Regr	96.2	95.7	95.4	93.3	96.6	95.9	—[c]	92.4
	91.2	90.5	90.2	88.0	92.0	91.0	—[c]	86.9
Com Regr	95.9	95.4	95.2	92.9	96.4	95.3	100.0	92.1
	90.9	90.8	90.6	87.5	91.8	90.0	100.0	86.5

[a] Nominal coverage of 95%.
[b] Nominal coverage of 90%.
[c] Subsample sizes are $n_1^\circ = 2 = n_2^\circ$, so the subsampled variance estimator is undefined.

variate normal vectors (a linear regression population situation). The latter populations were generated to have mean (15, 15), variance 1.0, and covariance .8 in the first stratum and the same mean, variances .5, and covariance .3 in the second stratum. The correlated chi-squared populations were generated by summing the component squares of 15 independently generated bivariate normal vectors with mean 0, variance 1.0, and covariance .8 in the first strata and mean 0, variance .5, and covariance .3 in the second stratum. The IMSL package with subroutine RNMVN performed the simulations.

We considered estimation of the population mean μ for the x_2 variable using Studentized estimates based on various ratio and regression estimators. In particular, $\hat{\mu}$ is taken to be a stratified estimator based on the x_2 data only (Cochran 1977, sec. 5.3), a separate ratio estimator (sec. 6.10), separate Tin estimator (secs. 6.15 and 6.16), combined ratio estimator (sec. 6.11), separate regression estimator (sec. 7.10), and combined regression estimator (sec. 7.10). Estimated standard deviations for each of the estimators are discussed in the appropriate sections of Cochran's book and used in Studentization for the T pivotals. All six pivotals assume forms that can be specified as smooth functions of vector means, so the theory of Section 5 is directly applicable.

Table 1 summarizes empirical coverage percentages of confidence intervals for μ constructed form the six T pivotals. A total of 2,000 samples were simulated from each of the four populations. The table entries record the frequency of coverage of 95% and 90% confidence intervals based on these samples. Cutoffs in the confidence intervals are approximated using appropriate quantiles from the distributions of T^*, T', and T°. These were computed by taking 3,000 resamples from each sample. For the BT method, we reconstructed $B = 200$ different empirical populations and took $C = 15$ resamples from each. The choice $B = 30$ and $C = 100$ gave virtually identical results. When implementing the BF method, reconstructed strata required randomizing between smaller and larger strata reconstructions; the smaller strata reconstructions were selected with probabilities (2/5, 1/3) for the smaller population and probabilities (3/10, 2/5) for the larger. The empirical coverages in Table 1 have an asymptotic standard error approximately equal to $100\sqrt{\alpha(1-\alpha)/2{,}000}$, or .67 when $\alpha = .9$. This calculation assumes an infinite number of resamples. However, as for the infinite population bootstrap, we expect the use of only a finite number of resamples to have very little effect on coverage accuracy, provided that the number used is sufficiently large (see Hall 1986 for discussion on this point).

Table 1 suggests that all three bootstrap methods show accurate nominal coverage, except for SB when used with the large population. In this situation the small sampling fraction leads to subsample sizes $n_1^\circ = 2 = n_2^\circ$, so that the

Table 2. Summary Statistics for 2.5% and 97.5% Quantile Estimators

	Smaller population $N_1 = 50, N_2 = 30$ $n_1 = 20, n_2 = 20$				Larger population $N_1 = 100, N_2 = 50$ $n_1 = 15, n_2 = 11$			
	2½%		97½%		2½%		97½%	
Estimator	BT	BF	BT	BF	BT	BF	BT	BF
Chi-squared populations								
Strat Est	-3.19[a]		1.63		-3.64		1.58	
	-3.07[b]	-3.08	1.67	1.67	-3.46	-3.44	1.69	1.69
	.542[c]	.556	.103	.109	.956	.982	.180	.180
Sep Ratio	-2.18		2.18		-2.40		2.64	
	-2.23	-2.22	2.23	2.24	-2.47	-2.46	2.75	2.76
	.200	.196	.240	.240	.556	.547	.669	.653
Sep Tin	-2.25		1.96		-2.54		2.31	
	-2.29	-2.29	2.01	2.02	-2.61	-2.60	2.41	2.42
	.197	.195	.212	.219	.560	.552	.631	.612
Com Ratio	-2.19		2.12		-2.43		2.47	
	-2.22	-2.22	2.14	2.14	-2.48	-2.47	2.49	2.49
	.177	.172	.167	.164	.521	.508	.425	.419
Sep Regr	-3.23		1.92		-3.51		2.16	
	-3.11	-3.12	1.93	1.94	-3.50	-3.48	2.30	2.30
	.582	.594	.121	.123	.962	.985	.471	.451
Com Regr	-3.20		1.95		-3.53		2.16	
	-3.09	-3.10	1.96	1.96	-3.47	-3.44	2.24	2.24
	.526	.536	.120	.119	.906	.943	.341	.339
Normal linear regression populations								
Strat Est	-1.96		2.17		-1.98		2.26	
	-1.97	-1.97	2.17	2.17	-2.02	-2.01	2.26	2.26
	.102	.104	.141	.142	.165	.159	.203	.232
Sep Ratio	-2.05		2.05		-2.13		2.02	
	-2.06	-2.07	2.05	2.05	-2.14	-2.14	2.04	2.05
	.101	.106	.102	.106	.150	.160	.135	.139
Sep Tin	-2.06		2.04		-2.14		2.01	
	-2.07	-2.07	2.04	2.05	-2.14	-2.15	2.04	2.04
	.101	.106	.103	.107	.150	.159	.135	.140
Com Ratio	-2.06		2.04		-2.14		2.02	
	-2.06	-2.07	2.04	2.05	-2.14	-2.14	2.04	2.04
	.101	.106	.102	.106	.151	.160	.134	.138
Sep Regr	-2.06		2.14		-2.17		2.22	
	-2.09	-2.09	2.16	2.17	-2.24	-2.25	2.29	2.30
	.115	.118	.130	.134	.213	.219	.242	.250
Com Regr	-2.11		2.17		-2.19		2.22	
	-2.13	-2.13	2.17	2.18	-2.23	-2.23	2.27	2.28
	.113	.115	.120	.123	.176	.178	.209	.207

[a] "Exact" 2½ percentile for the stratified estimator based on 10^5 simulations.
[b] Average of 2,000 quantile estimators for the 2½ percentile based on the finite population bootstrap (BT).
[c] Root mean square of the 2,000 BT estimators for the 2½ percentile.

SB method is not accurate or reliable. The BT and BF methods were roughly equivalent in their coverage accuracies. Normal approximations (Z) show substantial undercovering even in the two populations constructed from bivariate normal simulations. In the next section we show that the second term in the Edgeworth expansion of the distribution of $P(T \leq y \mid \mathcal{P})$ is an even function in y. This implies that the coverage errors of two-sided equitailed confidence intervals obtained using the normal approximation are of the same order of magnitude as those enjoyed by percentile-t bootstrap confidence intervals. Despite this similarity, bootstrap confidence intervals often have better two-sided coverage properties than normal intervals in practical situations. A possible explanation of this phenomenon was provided by Hall (1990), who showed that the bootstrap is a large-deviation approximation in certain circumstances and hence can perform well beyond the levels predicted by Edgeworth expansions (also see Bhattacharya and Qumsiyeh 1989).

Table 2 contains summary statistics in the estimation of 2.5 and 97.5 percentiles of the T distribution using appropriate quantiles from the distributions of T^* and T'. The top number in each cell is the "exact" quantile for that T pivotal computed numerically from 10^5 simulations of T. Below that are the mean and root mean squared error about the "exact" value for the two bootstrap estimators. The table shows little preference between BT and BF in the skewed chi-squared generated populations. But in the normally generated populations, BT shows a consistently smaller error for all estimators. We have not shown the statistics for quan-

Table 3. Summary Statistics for the Half-Widths of 90% Confidence Intervals

	Smaller population $N_1 = 50, N_2 = 30$ $n_1 = 20, n_2 = 20$				Larger population $N_1 = 100, N_2 = 50$ $n_1 = 15, n_2 = 11$			
Estimator	BT	BF	SB	Z	BT	BF	SB	Z
				Chi-squared populations				
Strat Est	3.88[a]	3.89	3.85	3.31	6.81	6.80	10.5	5.26
	.956[b]	.978	.948	.714	3.01	2.98	5.92	1.73
Sep Ratio	3.11	3.10	3.11	2.85	5.16	5.17	20.2	4.15
	.461	.461	.458	.409	1.27	1.23	20.1	.815
Sep Tin	3.00	2.99	3.00	2.85	4.94	4.96	16.0	4.15
	.433	.431	.428	.409	1.28	1.23	10.2	.815
Com Ratio	3.02	3.01	3.02	2.81	4.92	4.92	8.73	4.12
	.404	.405	.402	.374	1.07	1.07	2.94	.796
Sep Regr	3.09	3.11	3.09	2.50	5.16	5.15	—[c]	3.70
	.558	.590	.566	.328	1.61	1.63	—[c]	.809
Com Regr	3.03	3.05	3.01	2.44	5.00	4.98	32.3	3.61
	.527	.557	.526	.317	1.58	1.59	15.7	.783
				Normal linear regression populations				
Strat Est	.196	.196	.202	.188	.311	.313	.391	.292
	.0216	.0220	.0231	.0201	.0476	.0474	.0793	.0425
Sep Ratio	.136	.137	.139	.132	.192	.191	.224	.182
	.0132	.0130	.0135	.0122	.0239	.0238	.0324	.0226
Sep Tin	.136	.137	.139	.132	.192	.191	.224	.182
	.0132	.0130	.0135	.0122	.0239	.0238	.0324	.0226
Com Ratio	.136	.137	.139	.132	.191	.191	.224	.182
	.0131	.0130	.0135	.0121	.0239	.0238	.0322	.0225
Sep Regr	.129	.129	.137	.121	.191	.192	—[c]	.170
	.0122	.0124	.0140	.0109	.0256	.0251	—[c]	.0218
Com Regr	.128	.128	.135	.118	.187	.187	.585	.166
	.0120	.0122	.0135	.0105	.0244	.0240	.140	.0205

[a] Mean half width from 2,000 BT confidence intervals with 90% coverage.
[b] Standard deviation of these 2,000 half-widths.
[c] Subsample sizes are $n_1^0 = 2 = n_2^0$, so the subsample variance estimator is undefined.

tile estimation with the SB method, because it was not competitive with those listed.

Table 3 contains the means and standard deviations of half-widths for 90% confidence intervals generated for Table 1. All three bootstrap intervals show roughly the same summary statistics as long as the subsampling sizes for SB are not so small as to give unreliable results. Note that the Z intervals were shorter on average than their bootstrap counterparts, a fact reflected in their empirical coverages being too low.

Random sampling without replacement and Bernoulli sampling with the BF method were based on IMSL routines RNSRS and RNBIN using various types of SPARC and HP workstations.

5. THEORY

As in the infinite-population case, asymptotic methods—particularly the theory of Edgeworth expansions—provide a useful guide as to the performance of bootstrap procedures. To conceptualize the asymptotic arguments, it is necessary to impose a certain structure or asymptotic framework on the problem. We shall suppose that the rth stratum $\mathcal{P}_r$, $r = 1, \ldots, s$, is in fact a sample from a continuous superpopulation Π_r and that all the N_i's and n_j's diverge to infinity in such a way that each ratio N_i/N_j and n_i/n_j converges to a finite, nonzero limit and the sampling fractions are bounded below $1 - \varepsilon$, for some $\varepsilon > 0$. It is also necessary to assume that the first few moments of each superpopulation are finite. Under these assumptions, our claims concerning the second-order properties of the various estimators can be substantiated by extending theorems 1 and 2 of Babu and Singh (1985) to a multistrata setting. Babu and Singh considered a slightly weaker set of assumptions in which they imagined a sequence of finite populations of increasing size such that the sequence of distributions formed by assigning probability N^{-1} to each value x_i has a strongly nonlattice limit. Either of these sets of assumptions may seem quite artificial in a practical problem involving real data from a population of fixed size. But the asymptotic results will often still accurately predict the relative performance of competing statistical procedures.

The key structural assumption made by Babu and Singh (1985) is that $T = (n./q.)^{1/2}(\hat{\mu} - \mu)/\hat{\sigma}$ admits a short Taylor expansion, being a linear term of order 1, a quadratic term of order $n.^{-1/2}$, and a remainder of smaller order than $n.^{-1/2}$. This kind of expansion is available for many of the classical estimates of a finite population mean. For example, for the stratified sample mean $\hat{\mu} = \sum_{r=1}^{s} (N_r/N)\bar{X}_r$, we have

$$\hat{\sigma}^2 = \sum_{r=1}^{s} \left(\frac{N_r}{N}\right)^2 \frac{q_r/n_r}{q./n.} \frac{n_r}{n_r - 1} \hat{\sigma}_r^2,$$

where $\hat{\sigma}_r^2 = n_r^{-1} \sum_{i=1}^{n_r} (X_{ri} - \bar{X}_r)^2$ and $T = \bar{T} + o_p(n_\cdot^{-1/2})$ *with* $\bar{T} = \mathbf{l}\mathbf{Z}^T + (q./n.)^{1/2}\mathbf{Z}\mathbf{L}\mathbf{Z}^T$, where $\mathbf{l} = (l_1, \ldots, l_{2s})$ with $l_r = N_r/(N\sigma)$ and $l_{s+r} = 0$, $r = 1, \ldots, s$; $\mathbf{L} = (l_{ij})$, $i, j = 1, \ldots, 2s$ with

$$l_{ij} = -\frac{1}{2}\sigma^{-3}\frac{N_i}{N}\left(\frac{N_{j-s}}{N}\right)^2 \frac{q_{j-s}/n_{j-s}}{q./n.},$$

for $i = 1, \ldots, s, j = s+1, \ldots, 2s$, and zero for other values of (i, j); and $\mathbf{Z} = (Z_{11}, \ldots, Z_{1s}, Z_{21}, \ldots, Z_{2s})$ with $Z_{1r} = (n./q.)^{1/2}(\bar{X}_r - \bar{x}_r)$ and

$$Z_{2r} = (n./q.)^{1/2}\left\{n_r^{-1}\sum_{i=1}^{n_r}(X_{ri} - \bar{x}_r)^2 - \sigma_r^2\right\}$$

for $r = 1, \ldots, s$.

Extension of theorems 1 and 2 of Babu and Singh (1985) to the multistrata situation leads to the expansion for θ,

$$P(T \le y | \mathcal{P}) = \Phi(y) - \psi(y; \mathbf{p})\phi(y) + o(n_\cdot^{-1/2}), \quad (5)$$

with probability 1, where $\mathbf{p} = (p_1, \ldots, p_s)$ is the vector of sampling fractions and

$$\psi(y; \mathbf{p}) = E(\bar{T}) + \frac{1}{6}(y^2 - 1)E\{(\bar{T} - E\bar{T})^3\}.$$

Thus the coefficients of ψ depend on the sampling fraction vector $\mathbf{p}$. In the case of the stratified sample mean discussed earlier, it turns out that

$$E(\bar{T}) = -\frac{1}{2}(q./n.)^{1/2}\sum_{r=1}^{s}\left(\frac{N_r\sigma_r}{N\sigma}\right)^2\frac{q_r/n_r}{q./n.}\gamma_r$$

and

$$E\{(\bar{T} - E\bar{T})^3\} = 6E(\bar{T})\sum_{r=1}^{s}\left(\frac{N_r\sigma_r}{N\sigma}\right)^2 + (q./n.)^{1/2}$$
$$\times \sum_{r=1}^{s}\left(\frac{N_r\sigma_r}{N\sigma}\right)^2\left(\frac{q_r/n_r}{q./n.}\right)^2\frac{q_r - p_r}{q_r}\gamma_r,$$

where $\gamma_r = \sigma_r^{-3}E\{(X_{r1} - \bar{x}_r)^3\}$, $r = 1, \ldots, s$.

The bootstrap estimate (6) of $P(T \le y | \mathcal{P})$ is $\hat{\theta} = E\{P(T^* \le y | \hat{\mathcal{P}}^*) | \mathcal{X}\}$. We now observe that the expansion (5) also holds for the conditional probability $P(T^* \le y | \hat{\mathcal{P}}^*)$ with the population moments in the coefficients of ψ replaced by the corresponding moments of the empirical population $\hat{\mathcal{P}}^*$. In addition, observe that the empirical population moments differ from sample moments within each stratum by terms no greater than $O_p(n_\cdot^{-1/2})$. Hence the fact that the resample sizes and resampling fractions are the same as the original ones and the sample moments deviate from the population moments by terms of order $O_p(n_\cdot^{-1/2})$ explains why the bootstrap estimate is second-order correct. In contrast, the fact that the resampling fractions do not equal the original ones in the BF approach explains why this method does not yield second-order correct estimates. The latter result was noted by Babu and Singh (1985).

An empirical version of (5) also holds for the subsampling estimate $P(T^\circ \le y | \mathcal{X})$. But in this case, although the resampling fractions are asymptotically correct, the effective resample size is $n_\cdot^\circ$ rather than $n.$. The modified estimate $\hat{\theta}^\circ$ given in (4) correctly adjusts the subsampling estimate so as to account for the $n_\cdot^{-1/2}$ term in (5).

6. DISCUSSION

In conclusion, we would like to give a summary of arguments in favor of our bootstrap method. First, as we have stressed throughout the article, the bootstrap method leads to second-order–correct percentile-*t*–type confidence intervals, a property not shared by the BF method. The reason that the BF method is not second-order correct is that the resampling fractions do not match the original ones. Consider the case of a single stratum. In the BF approach the sampling fraction is either m_1^{-1} or $(m_1 + 1)^{-1}$, where $(m_1 + 1)^{-1} \le p_1 < m_1^{-1}$. Hence the smaller the sampling fraction, the closer the BF method to the *true* bootstrap.

Our use of the term "bootstrap" can be justified from another point of view. If we let the stratum sizes diverge to infinity with the sample sizes held fixed, then the estimate (1) converges to $E\{t(\hat{F}) | \mathcal{X}\}$, the analogous infinite population bootstrap estimate. This limit result also holds for the BF method but not for the subsampling estimate or its modified form. In fact, the subsampling procedure completely degenerates under this limit. For example, if $N_1 = 1{,}000$ and $n_1 = 35$, then $n_1^\circ = 1$! This poor asymptotic behavior of the subsampling method explains, to some extent, the discrepancy between the subsampling and bootstrap confidence intervals in the example at the end of Section 3. The effective resample size under the subsampling procedure is only 12 in this example, leading to highly variable estimates of the standard error.

Given that the finite population bootstrap estimate (1) converges to its infinite population counterpart, it might seem surprising that the error in the bootstrap is only $o(n_\cdot^{-1/2})$ rather than $O(n_\cdot^{-1})$, as it is in the infinite-population setting. By asking that the expansion of T be extendable to a cubic term of order $n_\cdot^{-1}$, plus a remainder of smaller order than $n_\cdot^{-1}$, it is possible to extend Babu and Singh's Edgeworth expansion by showing that the remainder is $O(n_\cdot^{-1})$ rather than simply $o(n_\cdot^{-1/2})$.

Finally, we remark that there is no reason why techniques other than the percentile-*t* method for constructing confidence intervals that have been developed for the infinite-population setting cannot be used in finite-population problems. Some possibilities for future investigation include the use of the bias correction methods of Efron (1987) and the iterated bootstrap (see, for example, Beran 1987 and Hall and Martin 1988).

[Received December 1991. Revised October 1993.]

REFERENCES

Babu, G. J., and Singh, K. (1985), "Edgeworth Expansions for Sampling Without Replacement From Finite Populations," *Journal of Multivariate Analysis,* 17, 261–278.

Beran, R. (1987), "Prepivoting to Reduce Level Error of Confidence Sets," *Biometrika,* 74, 457–468.

Bhattacharya, R. N., and Ghosh, J. K. (1978), "On the Validity of Formal Edgeworth Expansions," *The Annals of Statistics,* 6, 434–451.

Bhattacharya, R. N., and Qumsiyeh, M. (1989), "Second-Order L^p Comparisons Between the Bootstrap and Empirical Edgeworth Expansion Methodologies," *The Annals of Statistics,* 6, 434–451.

Bickel, P. J., and Freedman, D. A. (1984), "Asymptotic Normality and the Bootstrap in Stratified Random Sampling," *The Annals of Statistics,* 12, 470–482.

Chao, M. T., and Lo, S. H. (1985), "A Bootstrap Method for Finite Populations," *Sankhya,* Ser. A, 47, 399–405.

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* Philadelphia: Society for Industrial and Applied Mathematics.

——— (1987), "Better Bootstrap Confidence Intervals" (with discussion), *Journal of the American Statistical Association,* 82, 171–185.

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

Hall, P. (1986), "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval," *The Annals of Statistics,* 14, 1453–1462.

——— (1988), "Theoretical Comparison of Bootstrap Confidence Intervals," *The Annals of Statistics,* 16, 927–985.

——— (1990), "On the Relative Performance of Bootstrap and Edgeworth Approximations of a Distribution Function," *Journal of Multivariate Analysis,* 35, 108–129.

Hall, P., and Martin, M. A. (1988), "On Bootstrap Resampling and Iteration," *Biometrika,* 75, 661–671.

Kovar, J. G., Rao, J. N. K., and Wu, C. F. J. (1988), "Bootstrap and Other Methods to Measure Errors in Survey Estimates," *Canadian Journal of Statistics,* 165, 25–45.

Krewski, S., and Rao, J. N. K. (1981), "Inference from Stratified Samples: Properties of Linearization, Jackknife and Balanced Repeated Replication Methods," *The Annals of Statistics,* 2, 1010–1019.

Lo, A. Y. (1988), "A Bayesian Bootstrap for a Finite Population," *The Annals of Statistics,* 16, 1684–1695.

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 Publication, Washington, DC: U.S. Government Printing Office, 85–1369.

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.

Sitter, R. R. (1992), "A Resampling Procedure for Complex Survey Data," *Journal of the American Statistical Association,* 87, 755–764.