Southampton Statistical Sciences Research Institute
University of Southampton

Robust sample survey inference via bootstrapping and bias correction: The case of the ratio estimator

R. L. Chambers, A. H. Dorfman

Abstract

The bootstrap approach to statistical inference is described in Efron (1982). The method has wide applicability and has seen considerable development in recent years. However, use of the bootstrap in sample survey inference has been somewhat limited. Rao and Wu (1988), describe an application of the bootstrap under the design-based approach to sample survey inference. Sitter (1992a, 1992b), has extended their results to more complex survey designs. More recently, Booth, Butler and Hall (1991) and Booth and Murison (1992) describe a rather different approach to constructing a design-based bootstrap. In this paper we describe how this approach to the bootstrap can be applied under model-based sample survey inference, focussing on an application where the popular ratio estimator is the estimator of choice.

S^{3}RI Methodology Working Paper M03/13

Robust sample survey inference via bootstrapping and bias correction: The case of the ratio estimator

R. L. Chambers, Australian National University, and A. H. Dorfman, Bureau of Labor Statistics

A. H. Dorfman, OSMR, BLS, 2 Mass. Ave. N.E., Washington D.C. 20212 USA

Key Words: Estimation, Confidence Intervals

1. Introduction and notation

The bootstrap approach to statistical inference is described in Efron (1982). The method has wide applicability and has seen considerable development in recent years. However, use of the bootstrap in sample survey inference has been somewhat limited. Rao and Wu (1988), describe an application of the bootstrap under the design-based approach to sample survey inference. Sitter (1992a, 1992b), has extended their results to more complex survey designs. More recently, Booth, Butler and Hall (1991) and Booth and Murison (1992) describe a rather different approach to constructing a design-based bootstrap. In this paper we describe how this approach to the bootstrap can be applied under model-based sample survey inference, focussing on an application where the popular ratio estimator is the estimator of choice.

Given a finite population of N elements, let Y denote a variable of interest, with population values Y_I, $I = 1, ..., N$, and let X denote an auxiliary variable, with corresponding population values X_I, $I = 1, ..., N$,. The variables Y and X are intrinsically positive, with Y approximately proportional to X. It will be assumed that the values of X are all known, but the values of Y are known only for a sample s of $n \leq N$ of the population elements. Furthermore, given the values of X the process that was used to decide which elements of the population to include in the sample s will be assumed to be independent of the values of Y.

Once the sample has been selected, the values $\{Y_J ; J \in s\}$ are known. The problem is how to use this information, together with the known population values of X to make an inference about the unknown population mean $\bar{Y}$ of Y.

2. Model-based inference

The model-based approach to the above problem is based on the assumption that the values of Y can be looked upon as realisations of random variables whose distribution, conditional on the known values of X, may be specified via an appropriate probability model. It follows that $\bar{Y}$ is also the realisation of a random variable. Estimating the value of $\bar{Y}$ is therefore equivalent to predicting a realisation of this random variable, and standard methods for optimal prediction can be applied.

For this situation, a widely used model for *Y* expresses the mean and variance of this random variable as proportional to *X*. Denoting this model by ξ, it follows that

$$E_{\xi}(Y_I \mid X_I) = \beta X_I \tag{1}$$

and

$$\mathrm{var}_{\xi}(Y_I \mid X_I) = \sigma^2 X_I \tag{2}$$

where β and σ^2 are unknown positive constants. The subscript ξ in (1) and (2) signifies that these expectation and variance expressions are defined with respect to the distribution of *Y* under ξ. The best linear unbiased estimator of β is

$$\hat{\beta} = \frac{\bar{y}_s}{\bar{x}_s} \tag{3}$$

while the best linear unbiased predictor of $\bar{Y}$ is the famous ratio estimator

$$\bar{Y}_R = \hat{\beta}\bar{X}. \tag{4}$$

Here $\bar{y}_s$ and $\bar{x}_s$ are the means of the sample values of *Y* and *X* respectively, and $\bar{X}$ denotes the population mean of *X*.

In addition to computation of a point estimate using (4), good statistical practice requires estimation of a confidence interval for the unknown value of the finite population parameter $\bar{Y}$. Under the model-based approach, such a confidence interval is usually constructed by first calculating an ξ-unbiased point estimate v of the ξ-variance of the estimation error $\bar{Y}_R - \bar{Y}$. A 100(1 – α) per cent confidence interval for $\bar{Y}$ is then

$$\bar{Y}_R \pm v^{1/2} t(1 - \frac{\alpha}{2}, n-1) \tag{5}$$

where $t(1 - \frac{\alpha}{2}, n-1)$ denotes the $(1 - \frac{\alpha}{2})$-quantile of a t_{n-1} distribution.

A number of variance estimators for (4) have been proposed in the literature. In this paper we focus on the heteroskedasticity robust estimator investigated by Royall and Eberhardt (1975) and Royall and Cumberland (1981). This is

$$v_D = \frac{\hat{\sigma}_D^2}{n}\left(1 - \frac{n}{N}\right)\frac{\overline{XX}_r}{\bar{x}_s}\left(1 - \frac{C_s^2}{n}\right)^{-1} \tag{6}$$

where

$$\hat{\sigma}_D^2 = (n-1)^{-1}\sum_{J \in s}(Y_J - \hat{\beta}X_J)^2$$

and

$$C_s^2 = \frac{(n-1)^{-1}\sum_{J\in s}(X_J - \bar{x}_s)^2}{\bar{x}_s^2}.$$

3. Bootstrap confidence intervals

A problem with using (5) to compute a confidence interval for $\bar{Y}$ is that it implicitly assumes that the sample size n is sufficiently large for a central limit result to apply to the distribution of the estimation error $\bar{Y}_R - \bar{Y}$. In practice, this is hardly ever the case. Consequently, even if the model ξ holds exactly, the coverage properties of this interval estimator can be suspect. That is, confidence intervals determined via (5) are unlikely to possess the nominal coverage properties implied by their central limit behaviour under ξ. An alternative approach to constructing such confidence intervals is required.

Such an alternative approach is provided by bootstrap simulation. To motivate this approach, we observe that the primary reason for constructing a confidence interval around a point estimate of $\bar{Y}$ is to provide a properly calibrated measure of the uncertainty associated with this estimate. In particular, the aim is to exhibit an interval which includes the estimate value and which 'covers' 100(1 $-\alpha$) per cent of the estimated sampling distribution of the associated point estimator. Under the model-based approach, this sampling distribution corresponds to the distribution of possible alternative point estimates that could arise given selection of the same sample s from populations 'like' those actually observed. Since the model ξ specifies what constitutes populations 'like' the actual population underlying the observed data, it follows that a 100(1 $-$ α) per cent confidence interval for $\bar{Y}$, based on the ratio estimator $\bar{Y}_R$, is the interval

$$\left(\hat{Q}(\frac{\alpha}{2}), \hat{Q}(\frac{1-\alpha}{2})\right) \tag{7}$$

where $\hat{Q}(p)$ denotes an estimate of the p^{th} - quantile of the distribution of $\bar{Y}_R$ under ξ. Clearly, (5) is a special case of (7), based on the assumption that the sampling distribution of $\bar{Y}_R$ under ξ is normal, with mean $\bar{Y}$ and with variance $\mathrm{var}_\xi(\bar{Y}_R - \bar{Y})$.

An approach to constructing a confidence interval for $\bar{Y}$ that reflects the actual finite sample and finite population characteristics of the distribution of $\bar{Y}_R$ is to simulate such a distribution from the sample data. That is, we use the sample data and our model ξ to generate a sequence of alternative realisations for *Y*. Under the assumption that the same sample s of units is selected in each realisation, we then generate a sequence of 'potential' values for $\bar{Y}_R$, and estimate the quantiles in (7).

The key to carrying out such a bootstrap simulation of the distribution of $\bar{Y}_R$ is to reformulate ξ so as to indicate clearly how the population values of Y can be simulated. In order to do so, we replace (1) and (2) by the slightly stronger assumption that for each unit I in the population, there exists a positive constant β such that the values

$$\varepsilon_I = \frac{Y_I - \beta X_I}{\sqrt{X_I}} \tag{8}$$

are independent and identically distributed realisations of a random variable ε with zero mean and variance σ^2. Note that each population value of Y then satisfies

$$Y_I = \beta X_I + \sqrt{X_I}\varepsilon_I \ . \tag{9}$$

We can use (8) and (9) to simulate a bootstrap replication of the population values of Y. A little algebra shows that the set of studentized sample residuals

$$R_J^{std} = \frac{Y_J - \hat{\beta} X_J}{\sqrt{X_J\left(1 - \frac{X_J}{n\bar{x}_s}\right)}} \tag{10}$$

have the same mean and variance as ε. If, in addition, one makes the assumption that these studentized residuals also have the same distribution as ε, then (9) can be used to simulate the population values of Y, with the unknown values ε_J in this expression replaced by sample residuals (10) that have been randomly selected, with replacement, from the complete set of these values. Let Y_I^*, I = 1, ..., N denote this bootstrap realisation of the population values of Y, with mean $\bar{Y}^*$. A bootstrap realisation of the value of the ratio estimator of this mean follows by applying (3) and (4) to this bootstrap population. We denote the value of this estimate by $\bar{Y}_R^*$. The difference $\bar{Y}_R^* - \bar{Y}$ is the estimation error of the ratio estimator, based on the sample s, for this bootstrap population.

Repeating the procedure outlined in the preceding paragraph corresponds to application of a percentile bootstrap (Hall 1992), and can be used to obtain a bootstrap distribution for the estimation error $\bar{Y}_R - \bar{Y}$ of the ratio estimator (4) given the sample s. By subtracting this distribution from the actual value of the ratio estimator for the observed sample data, we obtain an estimate of the sampling distribution of $\bar{Y}_R$ under (9) that conditions on the sample s actually selected and is located at the realised value of this estimator. The final bootstrap 100(1 – α) per cent confidence interval for $\bar{Y}$ is computed by evaluating (7) on this bootstrap sampling distribution. That is, as

$$\left(Q^*(\frac{\alpha}{2}), Q^*(\frac{1-\alpha}{2})\right) \tag{11}$$

where $Q^*(p)$ denotes the p^{th} quantile of the bootstrap distribution.

4. Calibrating the percentile bootstrap

Let E^* and v^* denote the expectation and variance respectively of the bootstrap sampling distribution. It is straightforward to show that

$$E^* = \bar{Y}_R + \bar{r}^{std}\,\bar{x}_s^{(0.5)}\left(\frac{\bar{X}^{(0.5)}}{\bar{x}_s^{(0.5)}} - \frac{\bar{X}}{\bar{x}_s}\right) \tag{12}$$

and

$$v^* = \frac{v(R^{std})}{n}\frac{\bar{X}^2}{\bar{x}_s}\left(1 - \frac{n}{N}\frac{\bar{x}_s}{\bar{X}}\right) \tag{13}$$

where $\bar{X}^{(m)}$ denotes the mean of the m^{th} powers of the population X-values, with $\bar{x}_s^{(m)}$ denoting the corresponding sample mean,

$$\bar{r}^{std} = n^{-1}\sum_{J \in s} R_J^{std}$$

where the R_J^{std} are the studentised residuals (10), and

$$v(R^{std}) = n^{-1}\sum_{J \in s}(R_J^{std} - \bar{r}^{std})^2 .$$

Since the ratio estimator is the optimal predictor of $\bar{Y}$ under ξ, it is reasonable to require that the bootstrap confidence interval for $\bar{Y}$ be centred around $\bar{Y}_R$, in the sense that the mean of the corresponding bootstrap sampling distribution be equal to $\bar{Y}_R$. This can be accomplished by using (12) to mean correct the distribution of the bootstrap errors $\bar{Y}_R^* - \bar{Y}^*$ before locating this distribution at $\bar{Y}_R$.

The other desirable feature one could require of the bootstrap sampling distribution of the ratio estimator under ξ is that its variance (13) be equal in expectation to the ξ-variance of the estimation error $\bar{Y}_R - \bar{Y}$. It can be shown that

$$E_\xi(v^*) \le \mathrm{var}_\xi(\bar{Y}_R - \bar{Y}).$$

Consequently, the bootstrap errors need to be rescaled in order to remove this bias.

Combining the mean correction needed to ensure that the bootstrap mean equals $\bar{Y}_R$ with the rescaling needed to ensure the bootstrap variance is unbiased for $var_\xi(\bar{Y}_R - \bar{Y})$ leads to the ξ-calibrated percentile bootstrap distribution for $\bar{Y}_R$:

$$\bar{Y}_R + \left\{ \frac{\bar{r}^{std}\,\bar{x}_s^{(0.5)}\left(\frac{\bar{X}}{\bar{x}_s} - \frac{\bar{X}^{(0.5)}}{\bar{x}_s^{(0.5)}}\right) - \left(\bar{Y}_R^* - \bar{Y}^*\right)}{\sqrt{1-\frac{C_s}{n}}} \right\}. \tag{14}$$

5. Heteroskedasticity robustness

In practice (1) and (2) will only approximate the true relationship between Y and X in the population. This raises the issue of how robust is inference based on (14) when this approximation fails.

To provide some insight in this regard, suppose that (1) continues to hold, but (2) is potentially incorrect. In particular, suppose that, instead of (2),

$$\text{var}_\xi(Y_I) = \sigma^2 \psi(X_I) \tag{15}$$

where $\psi(t) \neq t$ in general. Since (1) remains true, we still have $E_\xi(\bar{r}^{std}) = 0$ and so the mean of the bootstrap sampling distribution remains an unbiased estimator of the population mean of Y. However, now

$$\text{var}_\xi(R_J^{std}) = \sigma^2 \frac{\psi(X_J)}{X_J}\left[1 - \frac{\frac{X_J}{n\bar{x}_s}\left\{1 - \frac{X_J \bar{\psi}_s}{\psi(X_J)\bar{x}_s}\right\}}{1 - \frac{X_J}{n\bar{x}_s}}\right]$$

so that

$$E_\xi v(R^{std}) = \sigma^2\left(n^{-1}\sum_{J\in s}\frac{\psi(X_J)}{X_J}\right) + \text{ lower order terms}$$

and consequently the expected value of the bootstrap variance v^* is

$$E_\xi v^* = \frac{\sigma^2}{n}\left(n^{-1}\sum_{J\in s}\frac{\psi(X_J)}{X_J}\right)\left(\frac{\bar{X}^2}{\bar{x}_s}\right)\left\{1 - \frac{n}{N}\frac{\bar{x}_s}{\bar{X}}\right\} + \text{ lower order terms}. \tag{16}$$

In contrast, the actual prediction variance of $\bar{Y}_R$ under the model defined by (1) and (15) is

$$\text{var}_\xi(\bar{Y}_R - \bar{Y}) = \frac{\sigma^2}{n}\left(\frac{\bar{\psi}_s}{\bar{x}_s}\right)\left(\frac{\bar{X}^2}{\bar{x}_s}\right)\left\{1 - \frac{n}{N}\frac{\bar{x}_s}{\bar{X}}\right\} + \text{ lower order terms}. \tag{17}$$

Comparing (16) and (17) one can see that the variance of the bootstrap sampling distribution (14) is no longer unbiased for the prediction variance of $\bar{Y}_R$. Using (14) when heteroskedasticity is misspecified can be expected to result in confidence intervals with coverage probabilities different from their nominal levels.

In order to robustify the percentile bootstrap against this type of misspecification, we need to redefine the residuals that are used to generate the bootstrap distribution. In order to do so, we note that the basic idea behind the percentile bootstrap (14) is that there exists an invertible transformation (8) of the underlying random variables (dependent on known or estimated effects) that results in pivotal values that are essentially independently and identically distributed. Using the percentile bootstrap is equivalent to substituting the empirical distribution function generated by the sample pivotal values for their unknown distribution function. When (15) holds, 'pivotal' values generated via (8) are no longer identically distributed, and furthermore, since ψ is unknown, we are no longer in a position to generate a 'correct' pivotal. Since we do not know the true heteroskedasticity, it seems reasonable to finesse this problem by resampling from the raw residuals

$$R_J = Y_J - \hat{\beta} X_J$$

where $\hat{\beta}$ is still the ratio estimator (3). Clearly, these residuals still have zero mean, even if they are no longer identically distributed. Liu (1988) has shown that a bootstrap based on independent but not identically distributed data is still valid, provided the data all have essentially the same location. Below we use this idea to motivate an alternative to (14).

The bootstrap population values are then

$$Y_I^* = \hat{\beta} X_I + R_I^*$$

where R_I^* is selected via simple random sampling with replacement from the R_J. As before, we denote the mean of this bootstrap population by $\bar{Y}^*$ and the value of the ratio estimator defined by the sample s for this population as $\bar{Y}_R^*$. The bootstrap sampling distribution for the original sample ratio estimator $\bar{Y}_R$ is then defined by the values

$$\bar{Y}_R - (\bar{Y}_R^* - \bar{Y}^*)$$

It is straightforward to show that this distribution has mean E^* equal to $\bar{Y}_R$ and variance

$$v^* = \frac{v(R)}{n} \frac{\bar{X}^2}{\bar{x}_s^2} \left(1 - \frac{n}{N} \frac{\bar{x}_s}{\bar{X}} \left\{ 2 - \frac{\bar{x}_s}{\bar{X}} \right\} \right) \qquad (18)$$

where

$$v(R) = n^{-1} \sum_{J \in s} R_J^2 .$$

Under the model defined by (1) and (15), it can be shown that

$$E_\xi v^* = \frac{\sigma^2}{n} \left(n^{-1} \sum_{J \in s} \psi(X_J) \right) \frac{\bar{X}^2}{\bar{x}_s^2} + \text{ lower order terms} .$$

The leading term in this expectation is the same as the corresponding leading term of the actual prediction variance (17) of the ratio estimator in this case. That is, in large samples at least, the variance estimator defined by the bootstrap based on the raw residuals is heteroskedasticity robust.

As with the scaled percentile bootstrap (14), this unscaled percentile bootstrap can be calibrated so that it is unbiased for the prediction variance of the ratio estimator under (1) and (2). Noting that under this default model

$$E_{\xi}v^{*} = \frac{\sigma^2}{n}\frac{\bar{X}^2}{\bar{x}_s}\left(1-\frac{\bar{x}_s^{(2)}}{n\bar{x}_s}\right)\left(1-\frac{n}{N}\frac{\bar{x}_s}{\bar{X}}\left\{2-\frac{\bar{x}_s}{\bar{X}}\right\}\right)$$

it follows that a ξ-calibrated percentile bootstrap for the ratio estimator based on raw residuals is defined by the values

$$\bar{Y}_R - \sqrt{\frac{\left(1-\frac{n}{N}\frac{\bar{x}_s}{\bar{X}}\right)}{\left(1-\frac{\bar{x}_s^{(2)}}{n\bar{x}_s^2}\right)\left(1-\frac{n}{N}\frac{\bar{x}_s}{\bar{X}}\left\{2-\frac{\bar{x}_s}{\bar{X}}\right\}\right)}}(\bar{Y}_R^{*}-\bar{Y}^{*}). \qquad (19)$$

6. Bootstrapping a bias corrected estimator

What happens if one has both mean and variance misspecification in the assumed model? In particular, suppose

$$E_{\xi}(Y_J) = \mu(X_J)$$

and

$$\mathrm{var}_{\xi}(Y_J) = \sigma^2\psi(X_J)$$

with μ not proportional to the identity function. Since the average of the unscaled sample residuals is still zero, it follows that the bias of the bootstrap distribution generated by (19) is equal to the bias of the ratio estimator

$$E_{\xi}(\bar{Y}_R - \bar{Y}) = \left(\frac{\bar{\mu}_s}{\bar{x}_s} - \frac{\bar{\mu}}{\bar{X}}\right)\bar{X}.$$

Since

$$E^{*}(\bar{Y}_R - (\bar{Y}_R^{*} - \bar{Y}^{*})) = \bar{Y}_R$$

in any case, the effect of mean misspecification is therefore to shift the bootstrap distribution by an amount equal to the bias of the ratio estimator. Such a shift clearly affects the coverage properties of this distribution. We therefore consider a modification to the ratio estimator which reduces this bias.

To focus things, we assume that all that is known is about the underlying mean function μ is that it is a smooth non-linear function which is approximately proportional to the identity function over most of the range of X-values in the population. Two commonly occurring scenarios which correspond to this situation are (i) the presence of outliers in the population, or (ii) nonlinearity in μ.

Outliers in the population are often modelled by assuming that the population is in fact a mixture of outliers and non-outliers. That is, under ξ

$$Y_J = \Delta_J\left(\beta X_J + \sigma\sqrt{X_J}\varepsilon_J\right) + (1-\Delta_J)\left(\theta(X_J) + \sigma\sqrt{\gamma(X_J)}\eta_J\right) \tag{20}$$

where Δ_J is a zero-one random variable denoting outlier/non-outlier status with $pr(\Delta_J = 1) = \pi_J$ and ε_J, η_J are independent 'white noise' random variables (i.e. they both have zero mean and unit variance). Under (20) it is straightforward to show

$$\mu(X_J) = \beta X_J + (1-\pi_J)(\theta(X_J) - \beta X_J)$$

and

$$\psi(X_J) = \pi_J X_J + (1-\pi_J)\gamma(X_J) + \pi_J(1-\pi_J)\left(\frac{\beta X_J - \theta(X_J)}{\sigma}\right)^2.$$

The expression for μ under (20) suggests the form of bias adjustment necessary for the ratio estimator under this model. Suppose that an outlier robust estimate $\tilde{\beta}$ of β can be computed, so $E_\xi\tilde{\beta} \approx \beta$ under (20). Then one could replace the standard ratio estimator $\bar{Y}_R$ by an estimator of the form

$$\bar{Y}_P = N^{-1}\left\{\sum_{J\in s} Y_J + \tilde{\beta}\sum_{I\notin s} X_I + \sum_{J\in s} m_J\varphi_P(Y_J - \tilde{\beta}X_J)\right\} \tag{21}$$

where φ_P is a bounded, skew-symmetric function which bounds the influence of sample outliers, and hence ensures that the variability of (21) remains low, and the m_J are suitably chosen weights which ensure that the bias of (21) also remains low.

Chambers (1986) recommends the m_J be chosen so that (21) reduces to the best linear unbiased estimator (ie the ratio estimator) under the "working model" (1) and (2) when $\varphi_P(t) = t$. That is

$$m_J = \frac{\sum_{I\notin s} X_I}{\sum_{K\in s} X_K} = \frac{N}{n}\frac{\bar{X}}{\bar{x}_s} - 1.$$

In practice, φ_P will not be chosen as the identity function, since this gives sample outliers undue influence on (21), and boosts the variability of this estimator. Again, following Chambers

(1986), a sensible trade-off between increasing variance and increasing bias in this estimator is obtained by choosing a 'Huber-type' φ_P, i.e.

$$\varphi_P(y-\tilde{\beta}x)=\min\left\{\left|y-\tilde{\beta}x\right|,h\tilde{\sigma}\sqrt{x}\right\}\operatorname{sgn}(y-\tilde{\beta}x)$$

where $\tilde{\sigma}$ is a robust estimate of the scale parameter σ in (22), for example the MAD estimate (ie 1.4826 times the median of the absolute deviations of the scaled residuals $(Y_J-\tilde{\beta}X_J)X_J^{-1/2}$ from their median), and h is a 'tuning constant' which curtails the influence of extreme outliers on (21), but allows an adjustment for the bias of the ratio estimator under moderate deviations from (1) and (2). This argument implies h should be chosen quite large, say $h = 6$.

Computation of $\tilde{\beta}$ can be carried out using a wide variety of outlier robust methods. Since outlier contaminated populations typically exhibit highly skewed marginal distributions for both Y and X, it is advisable to use a method which not only ensures robustness against outliers in Y, but is also not sensitive to high leverage points in the sample X-values. Since using the ratio estimator $\hat{\beta}$ of β is equivalent to estimating this parameter via ordinary least squares from the transformed model

$$E(Y_JX_J^{-1/2}\,|\,X_J)=\beta X_J^{-1/2}$$

and

$$\operatorname{var}(Y_JX_J^{-1/2}\,|\,X_J)=\sigma^2$$

the high leverage sample X-values are those corresponding to large values on the diagonal of the "hat" matrix defined by this transformed model. Since the diagonal entries of this matrix are easily seen to be proportional to the sample X-values, it follows that the leverage of a sample point on the ratio estimator is proportional to its X-value. An estimating equation for $\tilde{\beta}$ which is insensitive to sample outliers and high leverage sample points is therefore

$$\sum_{J\in s}\frac{1}{\sqrt{X_J}}\varphi_E\left(\frac{Y_J}{\sqrt{X_J}}-\tilde{\beta}\sqrt{X_J}\right)=0.$$

Note that the estimation influence function φ_E in this estimating equation will typically not be the same as the prediction influence function φ_P in (21). Ideally φ_E should be chosen so that all sample outliers are excluded from estimation of β. Influence functions that vanish outside a finite interval are appropriate choices in this regard. In the application reported in the next section, φ_E was chosen as the bisquare function, with tuning constant set to 4.685.

The bias adjustment implicit in (21) is motivated by the special situation where mean misspecification occurs because of the presence of outliers relative to the working model defined by (1) and (2). In general, however, mean misspecification can occur for a wide variety of reasons, and an appropriate parametric specification for this misspecification is not apparent. All that is known is that the underlying mean function μ is a smooth non-linear function which is approximately proportional to the identity function over most of the range of X-values in the population.

A method of nonparametrically adjusting the ratio estimator $\overline{Y}_R$ in this situation is described in Chambers, Dorfman and Wehrly (1993; referred to as CDW from now on). Under this approach, the bias of the ratio estimator is estimated nonparametrically by smoothing the raw sample residuals $R_J = Y_J - \hat{\beta}X_J$ against some suitably chosen function of the X-values. The estimated bias is then added to the ratio estimator as an adjustment term. The resulting estimator is of the form

$$\overline{Y}_{NP} = \overline{Y}_R + N^{-1}\sum_{J\in s} m_J R_J$$

where the weights m_J are nonparametric prediction weights defined by the method of smoothing used. For example, if the usual Nadaraya-Watson form of kernel regression smoothing is used to fit this bias then these weights are of the form

$$m_J = \sum_{I\notin s}\left\{\frac{K\left(b^{-1}(Z_J - Z_I)\right)}{\sum_{L\in s} K\left(b^{-1}(Z_L - Z_I)\right)}\right\}$$

where K is a kernel function, b is a bandwidth that needs to be chosen appropriately, and the Z-values are functions of the population X-values that are suited to smoothing the sample residuals R_J. In the context of estimating the finite population distribution function, CDW recommend setting Z equal to the rank of the corresponding X-value in the population.

The emphasis in the CDW approach is controlling bias, not variance. Consequently, (20) is likely to be sensitive to outliers in the sample data. An obvious modification to this estimator when dealing with highly skewed data is to follow Chambers (1986) and to base the nonparametric bias adjustment on 'huberized' (rather than raw) residuals. That is, the suggested form of nonparametrically adjusted ratio estimator for use in such outlier prone situations is

$$\overline{Y}_{NP} = \overline{Y}_R + N^{-1}\sum_{J\in s} m_J \min\left\{|R_J|, h\tilde{\sigma}\right\}\mathrm{sgn}(R_J) \tag{22}$$

where $\tilde{\sigma}$ denotes a robust estimate (eg the MAD) of the scale of the residuals R_J. The arguments advanced in Chambers (1986) then indicate that the tuning constant h should be chosen quite large, say $h = 6$.

An alternative approach is to robustify the smoother applied to the R_J. For example, rather than using the Nadaraya-Watson smoother, which corresponds to local mean smoothing, we can carry out local M-smoothing. Thus, if a "huberized" M-smoother is used, then (22) is replaced by

$$\bar{Y}_{NP} = \bar{Y}_R + N^{-1}\sum_{I \notin s} \hat{B}_I \tag{23}$$

where

$$\sum_{J \in s}\left\{\frac{K\left(\frac{Z_J - Z_I}{b}\right)\min\left(\left|R_J - \hat{B}_I\right|, h\tilde{\sigma}_I\right)\operatorname{sgn}(R_J - \hat{B}_I)}{\sum_{L \in s} K\left(\frac{Z_L - Z_I}{b}\right)}\right\} = 0 .$$

Here $\tilde{\sigma}_I$ is a robust estimate of the scale of the raw ratio residuals R_J in a "neighbourhood" of Z_I and h is the tuning constant. Again, $h = 6$ seems appropriate.

A final option one could consider when adopting a nonparametric approach to bias "robustifying" the ratio estimator follows from A. H. Welsh (1993, personal communication), who suggested that, rather than adding on a nonparametric bias correction to the ratio estimate, one could “nonparameterize” the approach of Chambers (1986) by first replacing the ratio estimator by a corresponding outlier robust nonparametric predictor and then “bias-correct” this predictor by adding on an adjustment term to allow for possible sample (and hence population) outliers.

In general this estimator takes the form

$$\bar{Y}_W = N^{-1}\left\{\sum_{J \in s} Y_J + \sum_{I \notin s} \hat{f}(Z_I) + \sum_{J \in s} w_J \varphi_P(Y_J - \hat{f}(Z_J))\right\} \tag{24}$$

where $\hat{f}(Z_I)$ denotes the value at Z_I of an outlier robust smooth of the sample Y-values against the sample Z-values, and the weights w_J are those suggested by Chambers (1986) – i.e. such that $\bar{Y}_W$ reduces to the best linear unbiased predictor of $\bar{Y}$ under the “target” parametric model (1) and (2), so that $\hat{f}(Z_I)$ in fact equals (or at least closely approximates) the expected value of Y_I under this model. The function φ_P is generally defined as a “mildly” bounded influence function. In our case we take it to be the standard Huber influence function with tuning constant h chosen quite large, say $h = 6$.

Application of the bootstrap idea to (21), (22), (23) and (24) is straightforward. Since the scaled percentile bootstrap (14) is nonrobust under heteroskedasticity misspecification, only unscaled percentile bootstraps for these estimators will be considered. Also, since all these estimators are quite complex, variance calibration of their bootstrap sampling distributions will not be considered. Indeed, the difficulty of getting explicit variance estimators in these cases may be regarded as a strong motive for considering the bootstrap. We therefore focus on a simple mean corrected percentile bootstrap for each estimator.

In the case of the parametrically-based estimator (21), bootstrap residuals can be formed by sampling with replacement from the raw robust residuals $\tilde{R}_J = Y_J - \tilde{\beta} X_J$. Denoting the I^{th} such selection by $\tilde{R}_I^*$, the corresponding bootstrap population value of Y is obtained as $Y_I^* = \tilde{\beta} X_I + \tilde{R}_I^*$. Let $\overline{Y}_P^*$ denote the value of (21) for this bootstrap population, with $\overline{Y}^*$ the corresponding mean of *Y* in this bootstrap population. The bootstrap sampling distribution for (21) is then formed from the values

$$\overline{Y}_P - \left(\overline{Y}_P^* - \overline{Y}^* - av(\overline{Y}_P^* - \overline{Y}^*)\right) \tag{25}$$

where $av(\overline{Y}_P^* - \overline{Y}^*)$ denotes the average (over the *m* bootstrap populations) of the prediction errors $\overline{Y}_P^* - \overline{Y}^*$. Bootstrap confidence intervals for the population mean of *Y* follow from (11).

The bootstrap residuals R_J^* appropriate for bootstrapping (22) and (23) are defined by sampling with replacement from the raw ratio residuals $R_J = Y_J - \hat{\beta} X_J$, with the corresponding bootstrap population values of Y given by $Y_I^* = \hat{\beta} X_I + R_I^*$. The bootstrap sampling distribution for these estimators is then defined by

$$\overline{Y}_{NP} - \left(\overline{Y}_{NP}^* - \overline{Y}^* - av(\overline{Y}_{NP}^* - \overline{Y}^*)\right) \tag{26}$$

where $\overline{Y}_{NP}^*$ denotes the value of either (22) or (23) for a particular bootstrap population, and $av(\overline{Y}_{NP}^* - \overline{Y}^*)$ denotes the average bootstrap prediction error.

Finally, the bootstrap residuals R_I^* for the Welsh estimator (24) are obtained by sampling with replacement from the nonparametric residuals $R_J = Y_J - \hat{f}(Z_J)$, with the bootstrap population values of *Y* given by $Y_I^* = \hat{f}(Z_I) + R_I^*$. The bootstrap sampling distribution for (24) is obtained using (26), but with $\overline{Y}_W^*$ replacing $\overline{Y}_{NP}^*$.

7. A numerical study

This section presents results from a numerical study of the parametric and bootstrap confidence interval methods defined in earlier sections. The target population, denoted Beef in what follows, consists of 430 beef cattle farms, with Y corresponding to Income from Sale of Cattle, and X denoting the Number of Cattle on hand. Beef was used by CDW as a clear example of model misspecification under (1) and (2), though they also point out that there is a strong linear relationship between log(*Y*) and log(log(*X*)) for these farms. For our purpose, we ignore this alternative specification, and proceed to analyse Beef as if (1) and (2) were valid.

A total of 500 independent simple random samples of size 60 were selected from Beef. For each sample, the ratio estimator $\overline{Y}_R$, see (4), was calculated, together with the parametric variance estimate v_D, see (6). These estimates were then used to compute confidence intervals for $\bar{Y}$ based on the normal approximation (5). In addition, scaled (14) and unscaled (19) percentile bootstrap distributions for $\bar{Y}_R$ were computed from each sample, based on 500 bootstrap resamples in each case. Confidence intervals for $\bar{Y}$ were obtained from these distributions via (11). Finally, bias corrected versions $\bar{Y}_P$, $\bar{Y}_{NP}$ and $\bar{Y}_W$, see (21), (22), (23) and (24), of the ratio estimator were calculated for each sample, and bootstrap distributions for their errors computed, via (25) and (26).

All estimators used in the study are described in Table 1. All versions using nonparametric smoothing are characterised by a variable bandwidth b_I which depends on the smoothing variable value Z_I of the nonsample unit being predicted, and is calculated as

$$b_I = \min\left(b_{I5}, \frac{c \times (\text{sample range of } Z)}{4n^{1/5}}\right) \tag{27}$$

where the value of c is specified by the user and b_{I5} denotes the smallest bandwidth such that there are at least 5 sample units with Z-values inside the smoothing "window" located at Z_I. In all cases the kernel smoother used an Epanechnikov kernel. Smoothing was local linear smoothing, and, with some alternatives considered in the study, these smooths were robustified. Details are set out in Table 1.

The average errors (AVE), root mean squared errors (RMSE) and median absolute deviation errors (MADE) of the various estimators of $\bar{Y}$ considered in the study are set out in Table 2. It is clear that the ratio estimator (either calculated directly or averaged over a bootstrap distribution) performs badly for Beef, and is substantially outperformed by the nonparametric alternatives NP(1X-Hu/6) and W(1X-Hu/2-Hu/6) based on smoothing against the population X-values. However, it is also clear that both the robust parametric estimator P/Hu6 and the nonparametric estimator based on smoothing against the population X-ranks, NP(1R-Hu/6), perform badly for this population, recording large values for AVE. This is not entirely surprising as far as P/Hu6 is concerned, since this estimator is based on the assumption that the majority of the survey population follow the simple linear target model (1) and (2), which is not the case for the Beef population. In the case of NP(1R-Hu/6), however, this result indicates that the use of a nonparametric approach to bias adjustment is not a simple "cure-all", but something that needs to be used with judgement. In particular, there appears to be a potential for a substantial bias induced by smoothing against the "wrong" Z-value. Appropriate diagnostics for picking the "right" Z-value for nonparametric bias adjustment are currently being researched by the authors.

Confidence interval coverage performances are set out in Table 3. The conventional method (5), as well as the unscaled percentile bootstrap, can be seen to perform weakly, recording some undercoverage at every nominal coverage level examined in the study. The scaled percentile bootstrap performs poorly, in large part due to its sensitivity to misspecification of the underlying heteroskedasticity. The large biases associated with P/Hu6 and NP(1R-Hu/6) are reflected in very poor coverage performance for their bootstrap CI's. Surprisingly, the Welsh estimator-based bootstrap CI's also record very poor coverage. This is explained by the fact that this estimator has little variability, so its bootstrap CI coverage is extremely sensitive to its bias, which, though relatively small, is still appreciable (see the AVE value in Table 2). The only method to record a reasonable coverage performance was the bootstrap CI's based on the nonparametric estimator NP(1X-Hu/6) defined by smoothing against the population *X*-values.

Further insight into the coverage performances of the various methods can be obtained from Table 4 which sets out the average lengths of the nominal 90% and 95% CI's recorded by each method over the 500 samples. Here we see that the small variability inherent in Welsh estimator leads to bootstrap CI's that are much shorter than those generated by the other methods. In terms of length, the CI's generated by the unscaled bootstrap are much larger, and seem to be basically the same length as the normal theory CI's generated by v_D, with the scaled bootstrap generating much tighter confidence intervals. Overall, when one takes both coverage performance and average length of confidence intervals into account, the best performing CI's are clearly those generated by bootstrapping the nonparametrically adjusted estimator NP(1X-Hu/6).

8. Conclusion

The goal of our research has been sound confidence intervals for estimates of means based on simple random sampling. Standard normal theory intervals can fail to attain coverage near the nominal; we are thus led to investigate bootstrap confidence intervals. Our investigation has focussed on a simple, model-based, unscaled, percentile bootstrap, and has proceeded from the commonly accepted regression through the origin model with residual variances proportional to the independent variable through modifications of this model allowing for indeterminate heteroscedasticity and outliers, to a purely non-parametric regression model allowing for outliers.

The evidence of an extended simulation study on the Beef population is that the achievement of this research has been to an extent orthogonal to its goal. We see greater efficiency using the successive model refinements and estimators, but, with the exception of the CDW estimator with adjustments based on population *X*-values (as opposed to ranks), the corresponding bootstrap intervals have not yielded improved coverage; for one of the more efficient estimators, the Welsh estimator, coverage was among the worst.

Thus the attainment of sound confidence intervals using the bootstrap requires more work. More sophisticated methods of bootstrapping seem to be in order, for example use of alternative methods for generating the bootstrap errors (as suggested in Liu, 1988), and use of the *t*-percentile bootstrap. The barrier to the use of the latter, for many of the estimators considered, has been the lack of corresponding variance estimators. An important step will be the development of such variance estimators. Work on these possibilities is underway.

References

Booth, J. G., Butler, R. W. and Hall, P. G. (1991). Bootstrap methods for finite populations. *CMA Technical Report SR31*, The Australian National University.

Booth, J. G. and Murison, R. (1992). Bootstrap confidence intervals in finite populations. *CMA Technical Report SR25*, The Australian National University.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association* **81**, 1063-1069.

Chambers, R. L., Dorfman, A. H. and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* **88**, 260-269.

Efron, B. (1982). *The Jacknife, the Bootstrap and Other Resampling Plans*. SIAM: Philadelphia.

Hall, P. G. (1992). *The Bootstrap and Edgeworth Expansion*. Springer: New York.

Liu, R. Y. (1988). Bootstrap procedures under some non-I.I.D. models. *Annals of Statistics* **16**, 1696-1708.

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.

Royall, R. M. and Eberhardt, K. R. (1975). Variance estimates for the ratio estimator. *Sankhya* **C37**, 43-52.

Royall, R. M. and Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association* **76**, 66-82.

Sitter, R. R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association* **87**, 755-765.

Sitter, R. R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics* **20**, 135-154.

Table 1: Estimation procedures investigated in the study

Notation	Description
R	Standard ratio estimator (4)
R/Scaled	Empirical mean of calibrated percentile bootstrap distribution (14) based on scaled residuals
R/Unscaled	Empirical mean of calibrated percentile bootstrap distribution (19) based on unscaled residuals
P(Hu/6)	Parametrically adjusted alternative (21) to the ratio estimator based on "Huberized" residuals, h = 6
NP(1R-Hu/6)	Nonparametrically adjusted ratio estimator (22) based on local linear smoothing (c = 3) of "Huberized" residuals (h = 6) against population X-ranks
NP(1X-Hu/6)	Nonparametrically adjusted ratio estimator (23) based on robust (Huber influence function, h = 6, c = 3) local linear smoothing of raw residuals against population X-values
W(1X-Hu/2-Hu/6)	Welsh estimator (24) with prediction term defined by a robust (Huber influence function, h = 2, c = 3) local linear smooth against X, and adjustment term based on "Huberized" (h = 6) residuals from this smooth

Table 2 Average error (AVE), root mean squared error (RMSE) and median absolute deviation error (MADE) for estimators of $\overline{Y}$ for Beef (N = 430, n = 60, $\overline{Y}$ = 130441)

Estimator	AVE	RMSE	MADE
R	5768	30802	21269
R/Scaled	5830	30898	21085
R/Unscaled	5728	30802	21089
P(Hu/6)	15058	26647	16974
NP(1R-Hu/6)	12918	28026	17545
NP(1X-Hu/6)	-2471	19813	12533
W(1X-Hu/2-Hu/6)	3597	22456	11622

Table 3 Unconditional coverage performances of confidence interval estimators for Beef. The figures in the table show the actual coverage percentages achieved over the 500 samples at different levels of nominal coverage

	Nominal coverage (%)			
	80	90	95	98
Normal theory CI's based on v_D	74.8	84.4	90.8	95.0
R/Scaled bootstrap (14)	46.2	62.4	74.0	82.4
R/Unscaled bootstrap (19)	72.8	83.2	89.4	94.4
P(Hu6)/Mean corrected bootstrap (25)	62.6	73.2	80.8	87.8
NP(1R-Hu/6)/Mean corrected bootstrap (26)	53.6	68.0	78.8	86.0
NP(1X-Hu/6)/Mean corrected bootstrap (26)	76.4	88.2	94.0	96.8
W(1X-Hu/2-Hu/6)/Mean corrected bootstrap (26)	55.2	64.8	71.8	79.8

Table 4 Average lengths of 90 and 95 per cent confidence intervals

	90%CI	95%CI
Normal theory CI's based on v_D	89624	107308
R/Scaled bootstrap (14)	53330	65038
R/Unscaled bootstrap (19)	87244	104066
P(Hu6)/Mean corrected bootstrap (25)	63729	77029
NP(1R-Hu/6)/Mean corrected bootstrap (26)	53802	66648
NP(1X-Hu/6)/Mean corrected bootstrap (26)	67374	83833
W(1X-Hu/2-Hu/6)/Mean corrected bootstrap (26)	37565	45680