Resampling-based Variance Estimation for Labour Force Surveys

Angelo J. Canty
Department of Mathematics and Statistics
Concordia University, Montreal, Canada
canty@discrete.concordia.ca

and

A. C. Davison
Department of Mathematics, Swiss Federal Institute of Technology,
1015 Lausanne, Switzerland
Anthony.Davison@epfl.ch

Abstract

Labour force surveys are conducted to estimate quantities such as the unemployment rate and the number of people in work. Interest is typically both in estimates at a given time and in changes between two successive time-points. Calibration of the sample to force agreement with known population margins results in random weights being assigned to each response, but the usual methods of variance estimation do not account for this. This paper describes how resampling methods — the jackknife, jackknife linearization, balanced repeated replication, and the bootstrap — can be used to do so. We also discuss implementation issues, and compare the methods by simulation based on data from the UK Labour Force Survey. The broad conclusions are these: bootstrap and jackknife linearization are less computer-intensive than the other resampling methods for such applications and give better standard errors; 'standard' methods can be badly biased downwards; and it is essential to take variability of the weights into account.

Keywords: Balanced repeated replication; Bootstrap; Calibration; Jackknife; Jackknife linearization; Panel survey; Post-stratification; Raking ratio.

[0]This work arises from contracts carried out for the UK Office for National Statistics and the Swiss Federal Statistical Office. We thank the Data Archive at the University of Essex, and David Elliott, Beat Hulliger and Rudi Peters for many useful discussions, and acknowledge financial support from the Swiss National Science Foundation.

1 Labour Force Surveys

Reliable estimation of the unemployment rate and the size of a workforce is one of the priorities of a government statistics office, so most countries have a labour force survey, typically conducted by stratified sampling of private addresses using geopolitical strata such as counties or cantons. Within each selected address respondents are selected with a certain probability. At one extreme all members are selected, as in the UK survey, whereas for the Swiss equivalent only one individual per household is selected, with the disadvantage that within-household variation cannot be estimated directly.

There is usually interest in estimates at the time the survey is taken, and also in the change since the previous survey. To help reduce variability in change estimates, most such surveys have a panel structure, i.e. each time the survey is conducted most individuals from the previous time are retained, with the rest replaced by stratified sampling. Individuals typically remain for five interviews, so about 20% of them are replaced each time, though there is attrition due to non-response.

Often the sampled units are reweighted so that certain known marginal totals, typically based on demographic variables such as age, sex, geographic location and race, are estimated without error; this is known as calibration or post-stratification. Although not known exactly, these margins can be estimated fairly precisely from census data, and are considered known for the purpose of calibration, which helps to correct for differential non-response. The usual method used is *iterative proportional fitting*, whereby the weights are iteratively adjusted for each margin in turn. Full convergence of the algorithm is very slow but as the effect of any change in weights after the first few iterations is very small (Oh and Scheuren, 1983), it is common in practice to stop after only a few iterations. As the resulting weights are dependent on the sample they are random and this should in principle be accounted for in standard error calculation. However, most current variance estimates assume that the weights are fixed by the sampling design.

The purpose of the work summarized in this paper was: to compare current methods of variance estimation for labour force surveys with resampling methods; to see whether variability of the weights is important in practice; to assess the feasibility of using resampling methods to account for it; and to investigate implementation issues. A point of particular concern was the number of iterations required to take reweighting into account when resampling. The literature on resampling methods in sample surveys suggests that the weights should be recalculated for each resample, but this would be time-consuming in practice, and it would be preferable if fewer iterations could be used.

In Section 2 we describe the estimates of interest and the current methods of calculating their variability. Section 3 reviews the resampling methods we considered and discusses their implementation. Section 4 describes our simulation study — based on data from the UK Labour Force Survey — and gives our main results. Section 5 contains a

brief discussion, and the appendix some technical details.

2 Estimates and Standard Errors

Standard texts on sample surveys, such as Cochran (1977), contain variance formulae applicable for a wide range of sampling designs. In practice, however, the formula used is generally that for a simple random sample, multiplied by an adjustment — or design effect — that takes into account the difference between simple random sampling and the actual design. In this section we describe some typical estimators, and the corresponding simplified variance formulae.

For static estimation the panel nature of the survey is irrelevant and we suppose that we have a stratified, and possibly clustered, sample. Let H be the number of strata and n_h be the number of sampled clusters in the hth stratum. Let Y and Z be two variables for which data has been collected, with y_{hjk} and z_{hjk} being the values observed for the kth respondent at the jth address in the hth stratum, and let w_{hjk} denote the adjusted weight associated with this individual. Then the three types of estimate of interest are the total, average, and ratio,

$$\hat{\theta}_{tot} = \sum_{h,j,k} w_{hjk} y_{hjk}, \quad \hat{\theta}_{av} = \frac{\sum_{h,j,k} w_{hjk} y_{hjk}}{\sum_{h,j,k} w_{hjk}}, \quad \hat{\theta}_{rat} = \frac{\sum_{h,j,k} w_{hjk} y_{hjk}}{\sum_{h,j,k} w_{hjk} z_{hjk}}.$$

We shall write $\hat{\theta}_{tot}(y)$ etc. when we need to distinguish totals for two different variables. An example is $\hat{\theta}_{rat} = \hat{\theta}_{tot}(y)/\hat{\theta}_{tot}(z)$.

If we assume that the sample is a simple random sample and that the weights w_{hjk} are fixed, a standard error for $\hat{\theta}_{av}$ when Y is an indicator variable is

$$\mathrm{se}(\hat{\theta}_{av}) \doteq d(Y) \left\{ \hat{\theta}_{av}(1 - \hat{\theta}_{av})/n \right\}^{1/2}, \tag{1}$$

where $d(Y)$ is the design effect for Y and n the size of the sample on which $\hat{\theta}_{av}$ is based. The estimate of the total incidence of Y in the population, $\hat{\theta}_{tot}$, then has standard error N se$(\hat{\theta}_{av})$ where $N = \sum w_{hjk}$ is the population size. For a ratio estimator it is necessary to estimate the covariance between the two random variables involved and then use the formula

$$\mathrm{se}(\hat{\theta}_{rat}) \doteq \frac{1}{\hat{\theta}_{tot}(z)} \left[\mathrm{var} \left\{ \hat{\theta}_{tot}(y) \right\} - 2\hat{\theta}_{rat} \mathrm{cov} \left\{ \hat{\theta}_{tot}(y), \hat{\theta}_{tot}(z) \right\} + \hat{\theta}_{rat}^2 \mathrm{var} \left\{ \hat{\theta}_{tot}(z) \right\} \right]^{1/2},$$

where the variances are obtained from (1). When both Y and Z are indicator variables, the two variances in the above formula are calculated as described for a total. The covariance is estimated by the usual formula within each stratum,

$$\mathrm{cov} \left\{ \hat{\theta}_{tot}(y), \hat{\theta}_{tot}(z) \right\} \doteq \sum_{h=1}^{H} \frac{n_h}{(n_h - 1)} \sum_{j=1}^{n_h} (u_{hj} - \bar{u}_h)(v_{hj} - \bar{v}_h),$$

where $u_{hj} = \sum_k w_{hjk} y_{hjk}$, $v_{hj} = \sum_k w_{hjk} z_{hjk}$ and $\bar{u}_h$ and $\bar{v}_h$ are the sample stratum averages of the u_{hj} and v_{hj}.

For estimates of change $\hat{\theta}_2 - \hat{\theta}_1$ between two successive surveys, a typical approach is to combine standard errors for $\hat{\theta}_1$ and $\hat{\theta}_2$ using a formula such as

$$\begin{aligned} \mathrm{var}(\hat{\theta}_1 - \hat{\theta}_2) &\doteq \mathrm{var}(\hat{\theta}_1) + \mathrm{var}(\hat{\theta}_2) - 2kr \left\{ \mathrm{var}(\hat{\theta}_1) + \mathrm{var}(\hat{\theta}_2) \right\} /2 \\ &= (1 - rk) \left\{ \mathrm{var}(\hat{\theta}_1) + \mathrm{var}(\hat{\theta}_2) \right\}, \end{aligned}$$

where r is the correlation coefficient between the values of Y for an individual at the two times and k is the overlap proportion between addresses in the two samples. The value of r is usually not estimated from the sample but instead is taken to be a known constant. The true proportion of overlap should be used for k, rather than its theoretical value; the standard error is thereby slightly increased because of non-response.

3 Resampling Methods

Standard formulae such as those in the previous section do not allow for the randomness of the weights induced by recalibration. To account for this, we turn to resampling methods. Theoretical aspects of the literature on resampling for complex surveys are outlined in Chapter 6 of Shao and Tu (1995) and Chapter 4 of Thompson (1997), while Section 3.7 of Davison and Hinkley (1997) has a more practical discussion confined to bootstrap methods. Wolter (1985) gives an extensive description of jackknife and half-sampling methods; see also Chapter 11 of Särndal *et al.* (1992). There are also numerous recent papers on the subject. However we have found no discussion of estimates of change, nor of implementation issues for very large surveys. The sections below outline the main approaches, namely jackknifing, jackknife linearization, balanced repeated replications, and bootstrapping.

3.1 Jackknife

The jackknife idea is division of the sample into disjoint parts, recalculation of the statistic of interest based on the sample without each part in turn, and then combination of these recalculated statistics to estimate properties of the original statistic. With stratified cluster data each cluster is deleted in turn, and the variance calculations are done within strata. The jackknife bias and variance estimates are then

$$b_J = \sum_{h=1}^{H} (n_h - 1)(\bar{\hat{\theta}}_{-h} - \hat{\theta}), \qquad v_J = \sum_{h=1}^{H} \frac{(1 - f_h)(n_h - 1)}{n_h} \sum_{j=1}^{n_h} (\hat{\theta}_{-hj} - \bar{\hat{\theta}}_{-h})^2, \qquad (2)$$

where f_h is the proportion of clusters sampled in the hth stratum, $\hat{\theta}_{-hj}$ is the estimate recalculated without the jth cluster of stratum h, and $\bar{\hat{\theta}}_{-h}$ is the average of the estimates for that stratum. This requires a total of $\sum_h n_h$ recalculations of the statistic. Labour force surveys are usually large, and other jackknives have been proposed to reduce the number of recalculations involved in (2).

One alternative procedure is to group the n_h clusters in the hth stratum into $s_h \geq 2$ groups and to delete these in turn, requiring $\sum_h s_h$ recalculations. The only change in (2) is then replacement of n_h by s_h. This method produces less stable variance estimates than the full jackknife, and clearly s_h should be as large as possible to reduce the instability of the estimates. Although $s_h \equiv 2$ can be used, a larger value seems preferable in practice.

3.2 Jackknife linearization

The idea of jackknife linearization is to replace repeated recalculation of a statistic — in effect numerical differentiation — with analytic differentiation. The result is a formula which is simple to calculate, and which in large samples is a good approximation to the more burdensome full jackknife calculation. The technique is also known as the nonparametric delta method and the infinitesimal jackknife (Davison and Hinkley, 1997, Sections 2.7, 3.2). For an unstratified sample of size n, the nonparametric delta method variance approximation is $v_L = n^{-2} \sum l_j^2$. The empirical influence value l_j, the infinitesimal change in the statistic due to inclusion of the jth observation, is closely related to the influence function central to classical robust statistics (Hampel *et al.*, 1986). Replacement of n^{-2} with $\{n(n-1)\}^{-1}$ reduces the slight downward bias of v_L. For stratified cluster data the empirical influence value for the jth cluster in stratum h is l_{hj}, and the bias-adjusted variance formula is

$$v_L = \sum_{h=1}^{H} (1 - f_h) \frac{1}{n_h(n_h - 1)} \sum_{j=1}^{n_h} l_{hj}^2. \tag{3}$$

The key to this is calculation of the l_{hj}, which we outline in the appendix. Formula (3) presupposes sampling with replacement, but when the sampling fractions f_h are small it may also be used for sampling without replacement. Jackknife linearization produces textbook estimates of variance for statistics based on averages, but can also be applied to more complicated statistics, provided they are smooth.

This approach has the drawbacks that separate derivation of l_{hj} would be required for further statistics, and that it does not apply to statistics based on sample quantiles. A smaller difficulty is that our derivation of the l_{hj} assumes that the proportional fitting algorithm has iterated to convergence, whereas in practice the weights are based on rather few iterations. This seems unlikely to affect the results since the change in weights after the first few iterations is small.

3.3 Balanced repeated replication

The simplest form of balanced repeated replication, balanced half-sampling (McCarthy, 1969), applies when a sample consists of two observations from each of H strata. Then a half-sample is formed by taking one observation from each stratum, and recalculating the original statistic. There are 2^H resampled values of $\hat{\theta}$, and these can be combined to estimate the variance of the original estimate. This approach is infeasible unless H is small, but ideas from experimental design can be used to form a much smaller set of half-samples, balanced so as to produce exactly the same result as the full set of half-samples, at least for a linear statistic. This is done by taking H of the columns of a Hadamard matrix of order L, which can be chosen in the range $H+1, \ldots, H+4$. Appendix A of Wolter (1985) gives Hadamard matrices of orders up to 100.

Having obtained the original estimate, $\hat{\theta}$, and the L half-sample estimates, $\hat{\theta}_1^\dagger, \ldots, \hat{\theta}_L^\dagger$, the usual variance estimate for $\hat{\theta}$ is

$$v_{BRR} = L^{-1} \sum_{l=1}^{L} (\hat{\theta}_l^\dagger - \hat{\theta})^2. \tag{4}$$

When the hth stratum is of size $n_h > 2$, the simplest way to generalize balanced repeated replication is to split the stratum randomly into two groups of nearly-equal size, and then recalculate the $\hat{\theta}^\dagger$, each of which involves taking one group from each stratum. Suppose that the groups in stratum h have sizes m_h and $n_h - m_h$, and that there is a weight w_{hik} attached to the kth individual in the ith cluster in the hth stratum. Under the scheme with $n_h = 2$, a half-sample is constructed by doubling one of these weights, and setting the other to zero, separately for each stratum. These weights will not satisfy the marginal constraints, however, and so each half-sample must be recalibrated. Rao and Shao (1997) have suggested a better approach in the grouped setting, taking samples of size $m_h \leq n_h/2$ in each stratum and modifying the weights to

$$w_{hik}^+ = \left(1 + \sqrt{\frac{n_h - m_h}{m_h}}\right) w_{hik}, \quad w_{hik}^- = \left(1 - \sqrt{\frac{m_h}{n_h - m_h}}\right) w_{hik}, \tag{5}$$

which gives the usual variance estimator for a linear statistic; these reduce to $w_{hik}^+ = 2w_{hik}$, $w_{hik}^- = 0$ when $m_h = n_h/2$. This gentler perturbation of the sample reduces the instability of (4) that can arise when the weights vary greatly due to deleting half the sample. The weights (5) would be adjusted to satisfy the marginal constraints as before. One would suppose that this modification will also give improvements for nonlinear statistics. A drawback of this adjustment is that a set of balanced sub-samples can no longer be constructed directly from a Hadamard matrix. Instead it is necessary to use orthogonal multi-arrays (Sitter, 1993). This would require many more replications than the original method and appears infeasible for a very large survey.

3.4 Bootstrap

The bootstrap idea is to mimic how the original data were constructed, by estimating the original population from the sample, and then to construct bootstrap analogues of the original estimator by applying the algorithm that produced it to samples taken from the estimated population. Let F denote the population, $\hat{F}$ its estimate based on the sample, and $t(\cdot)$ the algorithm that when applied to the population gives the parameter and when applied to to the data gives the estimate, i.e. $\theta = t(F)$ and $\hat{\theta} = t(\hat{F})$. Then the bootstrap idea is to resample from $\hat{F}$ to get a bootstrap sample $\hat{F}^*$ and corresponding statistic $\hat{\theta}^* = t(\hat{F}^*)$. This process is repeated R times independently to get $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_R$, and these are used to estimate the bias and variance of $\hat{\theta}$ by

$$b_B = \bar{\hat{\theta}}^* - \hat{\theta} = R^{-1} \sum_{r=1}^{R} \hat{\theta}^*_r - \hat{\theta}, \qquad v_B = (R-1)^{-1} \sum_{r=1}^{R} (\hat{\theta}^*_r - \bar{\hat{\theta}}^*)^2. \tag{6}$$

The usual approach with a single homogeneous sample $y_1, \ldots, y_n$ taken independently from F is to take $\hat{F}$ to be the empirical distribution function (EDF) that puts masses n^{-1} on each of the original y_j, and to construct $\hat{F}^*$ by sampling $y^*_1, \ldots, y^*_n$ with equal probability and with replacement from $y_1, \ldots, y_n$; $\hat{F}^*$ is then the EDF of $y^*_1, \ldots, y^*_n$. For stratified data this is carried out independently in each stratum. For bias and variance estimation, R should be in the range 50–200, and values low in this range give the bootstrap a computational advantage over the jackknife when calculating confidence intervals based on normal approximation. Computation of more elaborate confidence intervals requires $R \doteq 1000$.

A crucial problem when applying the bootstrap to finite population sampling is to mimic the effect of sampling without replacement, which decreases the variance of the estimate. Most of the methods suggested to do this are both *ad hoc* and hard to apply in practice; they may involve randomization between fake sample sizes, shrinkage of results obtained using resampling with replacement, or with-replacement resampling with larger sample sizes in order to reduce the variance by the appropriate amount. Theoretical work of Presnell and Booth (1994) and simulations there and in Section 3.7 of Davison and Hinkley (1997) suggest that a good simple approach is as follows. Suppose that the sample $y_1, \ldots, y_n$ has been taken without replacement from a population of size N, and for simplicity suppose that the inverse of the sampling fraction $1/f = N/n$ is integer. Then the bootstrap population of size N is made by concatenating $1/f$ copies of $y_1, \ldots, y_n$, from which a bootstrap sample of size n is taken without replacement. This is applied separately to each stratum of a stratified population. If $1/f$ is small, as is usually the case for a labour force survey, the difference between sampling with and without replacement is negligible, and we may simply resample from $y_1, \ldots, y_n$ with replacement, as described above.

The construction of a resample just described will in general mean that the post-strata marginal totals are no longer satisfied, and so we should reweight the resample. The question is: to what margins should it be calibrated? The original statistic may be written as $T = t\{\hat{F}, m(F)\}$, where $\hat{F}$ represents the sample together with its original adjusted sampling weights, F the population from which it would have come in the absence of non-response, and $m(F)$ the population margin to which $\hat{F}$ is calibrated. Note that $\hat{F}$ is a sample from the *responding population*, F_π, say, rather than from the population of interest F. The bootstrap analogue of T is $T^* = t\{\hat{F}^*, m(\hat{F})\}$, where $\hat{F}^*$ is the bootstrap sample, and $m(\hat{F})$ is the margin for the sample. But calibration ensures that $m(\hat{F}) = m(F)$, i.e. the original sample is reweighted so that its calibrated margins match those of the population. This implies that the bootstrap data may be reweighted by applying to it the usual algorithm with the original population marginal totals.

It might be thought that taking a bootstrap sample from a sample $\hat{F}$ from the responding population F_π rather than from F itself would introduce a bias, but in fact it does not. To understand this, consider a sample $\hat{F}$ of values $(x_1, y_1), \ldots, (x_n, y_n)$ obtained by random sampling with fraction f from a population but then subject to non-response that depends on x. Suppose that x equals one of $1, \ldots, k$, and the probability of response is π_i for $x = i$. Recalibration of the sample to match known margins for x will result in weight w_i being given to each of the (x_j, y_j) for which $x_j = i$, and if the recalibration is performed correctly, $w_i \doteq (f\pi_i)^{-1}$, because one effect of the calibration is to adjust for non-response. The obvious way to generate a bootstrap observation is then to take (x^*, y^*) from the original pairs, but with probabilities equal to their relative weights, and then to apply the estimated non-response mechanism, deleting an observation (x^*, y^*) for which $x^* = i$ with probability $(fw_i)^{-1}$, which is the best estimate of π_i from the survey. This procedure is repeated until n resampled pairs are obtained; although the number of respondents n is random under the original scheme, it is an experimental ancillary statistic and so is held fixed under the bootstrap scheme. But this procedure amounts to sampling n times with replacement from $\hat{F}$, i.e. applying the bootstrap in the obvious way. In other words, the original sample is taken from F_π, and we wish to mimic the effect of sampling from this. But $\hat{F}$ is the best estimate of F_π, so we should resample from $\hat{F}$. This has the virtuous side-effect of removing any need to model non-response.

4 Simulation Study

4.1 Artificial survey

We use data from the British Labour Force Survey (LFS) as our artificial population and draw samples from it in a way that mimics as closely as possible the stratified cluster

sampling used in the LFS. One aspect of interest is estimates of change, so our artificial population is based on two consecutive LFS samples, in the autumn quarter of 1995 and the winter quarter of 1995–96, which we denote by $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively. There are about 60,000 addresses in $\mathcal{P}_1$, of which about 48,000 also appear in $\mathcal{P}_2$, along with 12,000 "new" addresses sampled for the first time; addresses correspond to clusters in the previous discussion. Addresses having their ith interview at a given occasion are said to be in wave i, so the LFS has five waves. Interviews for the LFS are carried out in interview areas, each of which is sampled in one particular week of the 13 weeks of a quarter.

We decided that our sample would consist of 1250 addresses divided equally between ten regions and the five waves of the survey, giving our artificial scheme a sampling fraction of about $\frac{1}{48}$. We stratify the sampling by region and wave, giving 50 strata, and get a sample $\mathcal{S}_1$ by taking 25 addresses without replacement from each stratum of $\mathcal{P}_1$. To get a sample $\mathcal{S}_2$ from $\mathcal{P}_2$, we account for the panel structure by replacing the 250 addresses in wave five of $\mathcal{S}_1$ with 250 addresses sampled from the the "new" addresses in wave 1 of $\mathcal{P}_2$, again sampling within regions. The remaining 1000 addresses of $\mathcal{S}_1$ should also appear in $\mathcal{P}_2$ so we use their quarter 2 data to complete $\mathcal{S}_2$. Ideally $\mathcal{S}_2$ would include all addresses in waves 1–4 of $\mathcal{S}_1$, but some 5% of the addresses are affected by circumstantial non-response, which we deal with by modifying the sampling to produce exactly 1250 addresses in both samples.

Each region of our artificial population is further divided into three strata which correspond to the month in which the interview took place, mimicking the division of the LFS into 13 interview areas. Below we consider the strata to be the 30 combinations of region $\times$ month. As in the LFS we sample entire clusters, except that we omit respondents aged less than 16 years.

The LFS itself is calibrated to certain marginal combinations of sex, age and geographical location. As our samples are much smaller we cannot use the same margins, so we calibrate to margins obtained by merging levels of the LFS margins: 23 areas formed by merging counties of Britain; a cross-classification of sex by age, in single years, for those between 16 and 24 and 25 or older; and a cross-classification of four large regions by age groups, 16–29, 30–44, 45–59, 60–75 and 75+. As in the LFS we calibrate by iterative proportional fitting, terminating after five iterations.

We report results for a total and a ratio, i.e. the total number unemployed and the unemployment rate. The unemployment rate is defined as the number unemployed divided by the number in the workforce, so its estimate is a ratio statistic. For each of these statistics we consider both static and change estimates. These are representative of the many other statistics we considered.

Table 1: *True population values of the unemployment statistics, with means and standard deviations of 10,000 estimates.*

Statistic	True value	Mean	Standard deviation
Total	5885	5881	579
Change in total	–207	–157	594
Rate (%)	8.29	8.28	0.81
Change in rate (%)	–0.25	–0.16	0.83

4.2 Implementation of resampling methods

We calculated jackknife linearization variance estimates using (3) and the formulae given in the appendix, ignoring finite population corrections.

For the jackknife we use the alternative method described in Section 3.1, with each of the 30 strata divided into $s_h = 10$ groups, so that each variance estimate requires 300 recalculations of the statistic.

As we are interested in change estimates, we subdivide each of our thirty strata into three parts for balanced repeated replication and the bootstrap, corresponding to those individuals in $\mathcal{S}_1$ only, those in both $\mathcal{S}_1$ and $\mathcal{S}_2$, and those in $\mathcal{S}_2$ only. With these 90 sub-strata we use a Hadamard matrix of order $L = 96$ for balanced repeated replication, and $R = 100$ bootstrap replicates.

Jackknife linearization involves no reweighting, but the other methods do. In order to examine its effect, we calculated three sets of estimates for each jackknife, bootstrap, and balanced repeated replication resample: the 'incorrect' estimate, which uses the original weights; a second estimate, which uses weights derived from these by a single iteration of the calibration algorithm; and a third estimate, which adjusts the original weights as fully as in the LFS, terminating after five iterations. In the case of balanced repeated replication the weights of those respondents in the half-sample are multiplied by two before calibrating the resample.

4.3 Results

Our first step was to find "target" values against which to compare the various methods. To do so we generated 10,000 samples and calculated the estimates of interest for each using five iterations of the calibration algorithm. Table 1 shows the true population values and the mean and standard deviation of these 10,000 estimates, whose distribution was very close to normal. Its final column gives the "true" standard deviation of the estimates, calculated from the 10,000 estimates, and we use these as the target values.

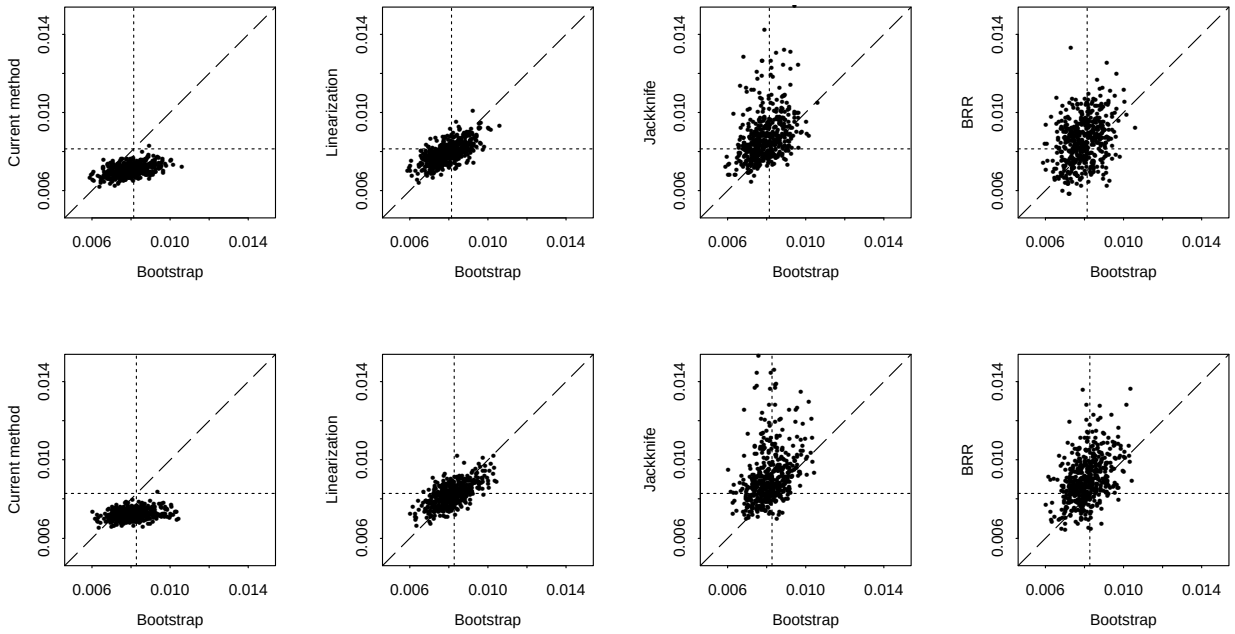
Table 2: *Summary statistics for standard errors of the total unemployed, the change in total unemployed, the overall unemployment rate and the change in the rate. RB is the relative bias, SD the standard deviation and MSE the mean squared error of the 500 estimates. BRR denotes balanced repeated replication.*

Statistic	Method	Bias	RB (%)	SD	MSE
Total	current	–35.2	–6.1	25.4	1880
	bootstrap	–14.6	–2.5	55.0	3230
	jackknife	45.3	7.8	96.8	11400
	jackknife linearization	–13.6	–2.4	38.4	1650
	BRR	28.0	4.8	79.6	7100
Change in total	current	–42.7	–7.2	21.6	2290
	bootstrap	–8.2	–1.4	55.5	3140
	jackknife	64.3	10.8	107.0	15600
	jackknife linearization	–4.7	–0.8	41.8	1770
	BRR	55.5	9.3	85.5	10400
Rate	current	–0.105	–12.9	0.0291	0.0119
	bootstrap	–0.019	–2.3	0.0789	0.0070
	jackknife	0.066	8.1	0.1370	0.0232
	jackknife linearization	–0.017	–2.1	0.0543	0.0033
	BRR	0.041	5.1	0.1160	0.0151
Change in rate	current	–0.105	–12.6	0.0252	0.0116
	bootstrap	–0.012	–1.5	0.0781	0.0062
	jackknife	0.089	10.8	0.1490	0.0302
	jackknife linearization	–0.007	–0.8	0.0583	0.0034
	BRR	0.077	9.3	0.1180	0.0197

We then took 500 of these 10,000 samples and for each calculated standard errors using the methods described in Sections 2 and 3. For the former we used a design effect of 1.05 and a correlation of 0.6 between quarters; these values were checked by a separate simulation. The implementation of the resampling methods is described in Section 4.2.

Table 2 summarizes the 500 standard errors calculated using five-step reweighting of each resample. The standard errors for the unemployment rate are plotted in Figure 1. The corresponding figures for the total unemployed and its change show the same pattern, namely that the non-resampling standard errors are stable but badly biased downwards. The bootstrap and jackknife linearization standard errors are comparable and both per-

Figure 1: *Standard errors for the unemployment rate (top row) and the change in unemployment rate (bottom row); the dashed line is $y = x$, the dotted lines are the 'true' sampling standard errors.*



form well. The versions of balanced repeated replication and the jackknife used here give very unstable standard errors, though they could be improved at the expense of more computing. As this extra computing would make the methods infeasible in practice, we did not examine the improvements.

4.4 Effect of reweighting

As stated earlier, one of our goals was to see the effect of reweighting the resamples, and in particular to check if fewer than five iterations of the calibration algorithm would suffice. Standard errors for the different methods of variance estimation showed the same pattern for all the static estimates. Jackknife standard errors show a large reduction in the bias of the standard error and a smaller reduction in its variance when just one iteration is used, but the situation worsens when five steps are used. For change estimates, there is little or no improvement with one-step reweighting and again a worsening when five-step reweighting is used. Overall the best jackknife strategy appears to be to use one iteration.

Reweighting has only a small effect for the bootstrap or balanced repeated replications, and the effect is almost entirely captured in just one iteration. This is also true for the change estimates.

Figure 2 shows standard errors for five-step reweighting against the other two levels

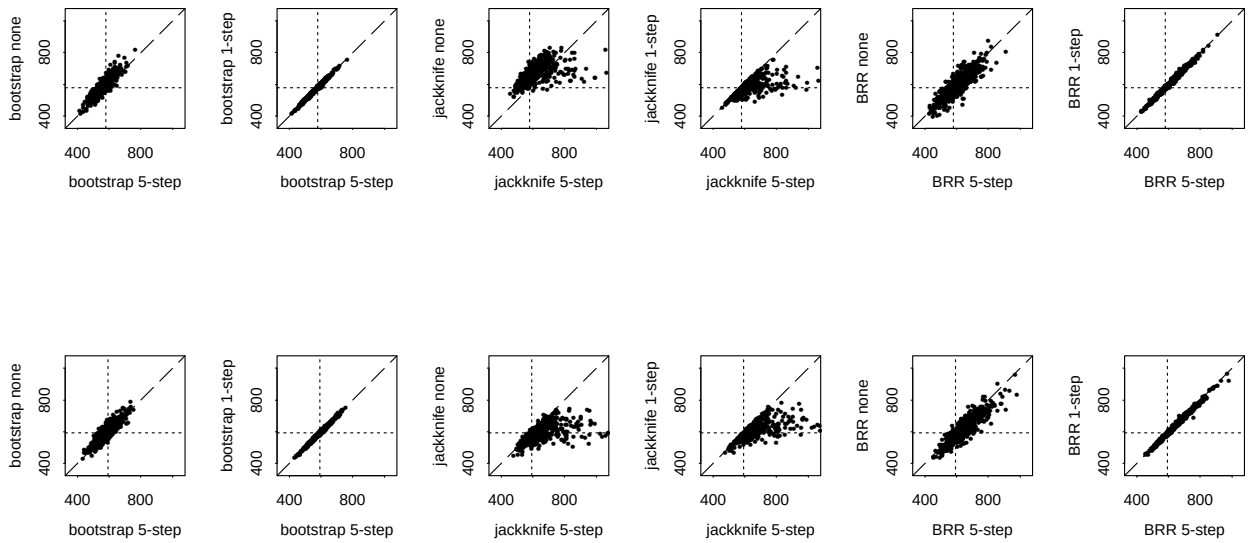
Table 3: *Summary statistics for resampling standard errors of the total unemployment and its change with different levels of reweighting based on 500 samples. RB is the relative bias, SD the standard deviation and MSE the mean squared error of the 500 estimates. BRR denotes balanced repeated replication.*

Statistic	Method	Reweighting	Bias	RB (%)	SD	MSE
Total	bootstrap	none	7.2	1.3	61.4	3820
		one-step	–15.4	–2.7	54.6	3220
		five-step	–14.6	–2.5	55.0	3230
	jackknife	none	91.5	15.8	58.3	11800
		one-step	1.4	0.3	50.1	2500
		five-step	45.3	7.8	96.8	11400
	BRR	none	16.0	2.8	81.5	6880
		one-step	23.6	4.1	78.4	6700
		five-step	28.0	4.8	79.6	7100
Change in total	bootstrap	none	1.78	0.3	58.8	3460
		one-step	–10.2	–1.7	55.0	3120
		five-step	–8.2	–1.4	55.5	2140
	jackknife	none	0.3	0.0	51.2	2610
		one-step	11.3	1.9	53.6	2990
		five-step	64.3	10.8	107.0	15600
	BRR	none	22.6	3.8	80.8	7030
		one-step	48.9	8.2	83.6	9370
		five-step	55.5	9.3	85.5	10400

of reweighting. The plots for the jackknife show a clear bias for the no-reweighting static estimate but not for the change estimate. A curious feature is that with five-step reweighting the distribution of standard errors is very skewed with some extremely large outliers. We have as yet no explanation for this, which occurred for all statistics of interest. Table 3 gives summary statistics for the standard errors with different levels of reweighting for the unemployment statistics. The presence of the large outliers make the jackknife standard errors with five-step reweighting upwardly biased and very variable and thus they have very large mean squared errors.

For the bootstrap and balanced repeated replication there is a very strong relationship between the estimates with one- and five-step reweighting, whereas there is a weaker relationship with the no-reweighting estimates.

Figure 2: *Standard errors for the total unemployed in quarter 1 (top row) and the change in total unemployment (bottom row) with different levels of reweighting; the dashes show the line $y = x$ and the dotted lines are the 'true' target sampling standard deviations.*



5 Discussion

Our results give clear answers to the original questions. It is important to take calibration into account, and resampling methods can do this. In practice the bootstrap and jackknife linearization give more reliable standard errors than do the jackknife or balanced repeated replications. These latter could be made more accurate, but the added computational burden seems likely to make them infeasible in practice, and in any event they would be unlikely to improve on the bootstrap or jackknife linearization. Although jackknife linearization slightly outperforms the bootstrap and is less computationally intensive, the analytical work it involves may make it less attractive in practice when a new quantity is of interest. If the bootstrap is used, weights should be recalculated, but a single iteration of the recalibration algorithm suffices.

As previously mentioned, the results presented are typical of those obtained for many more statistics. We also performed a separate simulation study, based on data from the Swiss labour force survey, for which similar results were obtained.

Appendix: Jackknife Linearization

This appendix outlines the derivation of empirical influence values for calibrated survey estimates. The results parallel those of Yung and Rao (1996), though the derivation is different. We give enough detail to show how the l_{hj} are obtained for estimates of change from such surveys; as far as we know these calculations are not available elsewhere.

We use y_{hjk} to denote the value of y for the kth individual in the jth of the n_h clusters in stratum h, for $h = 1, \ldots, H$, with similar subscripting for other quantities, and use ω_{hjk} to denote the initial weights which would be used in the absence of calibration. The estimated population total would then be $\hat{\theta} = \sum_{h,j,k} \omega_{hjk} y_{hjk} = \sum_{h,j} y'_{hj}$, where here and below the prime denotes an appropriately weighted sum over the clusters, i.e. $y'_{hj} = \sum_k \omega_{hjk} y_{hjk}$. Note that $\hat{\theta}$ is linear in the data. For stratified cluster sampling, the jackknife is performed by deleting clusters, and a standard calculation for linear statistics (Davison and Hinkley, 1997, Problem 2.10) gives that the empirical influence value for the jth cluster in stratum h is $l_{hj} = n_h y'_{hj} - \sum_i y'_{hi}$. Combined with (3) and ignoring the finite population correction, this yields

$$v_L = \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)} \sum_{j=1}^{n_h} \left(n_h y'_{hj} - \sum_i y'_{hi} \right)^2 , \tag{7}$$

in agreement with expression (2.2) of Yung and Rao (1996).

Now suppose that the sample consists of n observations of form $(\omega_{hjk}, x_{hjk}, y_{hjk})$. Here x_{hjk} is a $p \times 1$ vector of covariates with known $p \times 1$ population total c, and it is intended to estimate the population total for y. Let Ω denote the $n \times n$ diagonal matrix with elements ω_{hjk}, X the $n \times p$ matrix with rows x_{hjk}^T, y the $n \times 1$ vector with elements y_{hjk}, and 1 a vector of ones, conformable wherever it appears. Then calibrated weights obtained by iterative proportional fitting and a form of generalized regression are equivalent, provided the former is iterated to convergence (Stukel *et al.*, 1996). The generalized regression estimator of the population total for variable y may be written as

$$\begin{aligned} \hat{\theta} &= \sum_{hjk} \omega_{hjk} y_{hjk} + \left(c - \sum_{hjk} \omega_{hjk} x_{hj} \right)^T (X^T \Omega X)^{-1} X^T \Omega y & (8) \\ &= 1^T \Omega y + (c^T - 1^T \Omega X)(X^T \Omega X)^{-1} X^T \Omega y & (9) \\ &= 1^T W y, \end{aligned}$$

where W is the $n \times n$ diagonal matrix with elements

$$w_{hjk} = \omega_{hjk} \left\{ 1 + (c^T - 1^T \Omega X)(X^T \Omega X)^{-1} x_{hjk}^T \right\} ;$$

note that setting $y = X$ in (9) gives the known population total c for x. Thus $\hat{\theta}$ may be interpreted as a reweighting of the y_{hjk}, with random weights w_{hjk} that produce the

known total for the variables in x. We need to account for the variability of the w_{hjk} when calculating the analogue of (7), where it is assumed that the weights are fixed, and this may be done by applying the chain rule when differentiating $\hat{\theta}$.

Expression (8) shows that $\hat{\theta}$ is a function of the linear expressions $\sum \omega_{hjk} y_{hjk}$ and $c - \sum \omega_{hjk} x_{hjk}$, and the weighted least squares estimate $(X^T \Omega X)^{-1} X^T \Omega y$, for which empirical influence values for deletion of clusters are respectively

$$n_h y'_{hj} - \sum_i y'_{hi}, \quad -\left(n_h x'_{hj} - \sum_i x'_{hi} \right), \quad n_h (X^T \Omega X)^{-1} \sum_k x_{hjk} \omega_{hjk} e_{hjk}, \tag{10}$$

where $e_{hjk} = y_{hjk} - \hat{y}_{hjk}$ is a residual for the weighted regression. The first two expressions in (10) are obtained from the discussion just before (7), and the third by specializing Problem 7.1 of Davison and Hinkley (1997) to weighted linear regression. On applying the chain rule to $\hat{\theta}$, we eventually find that the empirical influence values for $\hat{\theta}$ are

$$l_{hj} = n_h \sum_k w_{hjk} e_{hjk} - \sum_i \sum_k w_{hik} e_{hik} = n_h e'_{hj} - \sum_i e'_{hi}. \tag{11}$$

Substitution of (11) into (3) gives a formula that agrees with (4.6) of Yung and Rao (1996), and amounts to replacing y'_{hj} in (7) by e'_{hj}.

For stratified cluster sampling applied to a ratio of two calibrated estimators, $\hat{\theta} = 1^T W y / 1^T W z$, the chain rule gives

$$l_{hj} = \frac{l^y_{hj} - \hat{\theta} l^z_{hj}}{1^T W z}, \tag{12}$$

where l^y_{hj} and l^z_{hj} are the empirical influence values (11) for the numerator and denominator weighted averages corresponding to the jth element of the hth stratum; the variance is obtained by substituting (12) into (3).

We now turn to the estimates of change from one sample to the next. The change in rate of unemployment may be written as

$$\hat{\theta} = t(\hat{F}_2, \hat{F}_3) - t(\hat{F}_1, \hat{F}_2), \tag{13}$$

where $\hat{F}_1$ represents those addresses present only in the first sample, $\hat{F}_2$ those present in both, $\hat{F}_3$ those present only in the second, and $t(\cdot, \cdot)$ is the ratio of two calibrated estimates. Thus $t(\hat{F}_1, \hat{F}_2) = 1^T W y / 1^T W z$, where W, y and z are formed using $(w_{hjk}, y_{hjk}, x_{hjk})$ and $(w_{hjk}, z_{hjk}, x_{hjk})$ for $\hat{F}_1$ and $\hat{F}_2$; note that the the weights W differ between these samples. Application of the chain rule to (13) gives empirical influence values l^1_{hj} for the ratio $-t(\hat{F}_1, \hat{F}_2)$ for an address present only in the first sample, l^3_{hj} for the ratio $t(\hat{F}_2, \hat{F}_3)$ for an address present only in the second sample, and l^2_{hj} for the difference of ratios (13) for an address present in both. In each case the empirical influence values are obtained using

(12). The variance estimate is

$$\sum_{h=1}^{H}\left\{\frac{1}{n_{h1}(n_{h1}-1)}\sum_{j=1}^{n_{h1}}(l_{hj}^{1})^2+\frac{1}{n_{h2}(n_{h2}-1)}\sum_{j=1}^{n_{h2}}(l_{hj}^{2})^2+\frac{1}{n_{h3}(n_{h3}-1)}\sum_{j=1}^{n_{h3}}(l_{hj}^{3})^2\right\},$$

where the hth stratum contains n_{h1} households present in the first sample only, n_{h2} households present in both samples, and n_{h3} households present in the second sample only, and the corresponding empirical influence values are given an obvious notation.

References

Cochran, W. G. (1977) *Sampling Techniques.* Third edition. New York: Wiley.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics: The Approach Based on Influence Functions.* New York: Wiley.

McCarthy, P. J. (1969) Pseudo-replication: half samples. *Review of the International Statistical Institute* **37**, 239–264.

Oh, H. L. and Scheuren, F. J. (1983) Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys.* Vol 2, eds W. G. Madow, I. Olkin and D. B. Rubin, volume 2, pp. 143–184. New York: Acaedemic Press.

Presnell, B. and Booth, J. G. (1994) Resampling methods for sample surveys. Technical Report 470, Department of Statistics, University of Florida, Gainesville.

Rao, J. N. K. and Shao, J. (1997) Modified balanced repeated replication for complex survey data. Preprint, Department of Mathematics and Statistics, Carleton University.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling.* New York: Springer.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap.* New York: Springer.

Sitter, R. R. (1993) Balanced repeated replications based on orthogonal multi-arrays. *Biometrika* **80**, 211–221.

Stukel, D. M., Hidiroglou, M. A. and Särndal, C.-E. (1996) Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology* **22**, 117–125.

Thompson, M. E. (1997) *Theory of Sample Surveys.* London: Chapman & Hall.

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

Yung, W. and Rao, J. N. K. (1996) Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology* **22**, 23–31.

List of Tables

List of Figures