DACSEIS

IST-2000-26057

Workpackage 5

Resampling Methods for Variance Estimation

Deliverable 5.1

2004

List of contributors:

Anthony C. Davison and Sylvain Sardy, EPFL.

Main responsibility:

Sylvain Sardy, EPFL.

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

Preface

This report reviews resampling techniques used for variance estimation in sample surveys. The resampling techniques considered are based on linearization, jackknife, balanced repeated replication, and the bootstrap. Some of the variance estimation procedures take into account calibration and imputation mechanisms used to improve the estimators. The report draws practical conclusions based on theoretical considerations and empirical comparisons.

Sylvain Sardy Lausanne, May 2004

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Exact variance formulae for estimators based on sample survey data are available only for limited, albeit very important, classes of estimators, and so approximations are widely used in practice. Classical variance approximations are derived by variants of the delta method, whereby awkward estimators are approximated using Taylor series expansion, from which approximate variance formulae are obtained. When the sample is small or the estimator complex it is natural to be concerned about the accuracy of the Taylor expansion on which the validity of the resulting formula depends, and to seek other approaches. One important class of alternatives is based on resampling procedures, which involve repeated computations with the sample actually obtained; for this reason they are sometimes also called sample re-use methods. The most important of these is the bootstrap, introduced by Efron (1979), but this relative latecomer to the scene is predated by the jackknife (Quenouille, 1949a, and Tukey, 1958), which was originally introduced for bias estimation in time series analysis, and by half-sampling (McCarthy, 1969), which has its roots in work undertaken during the late 1950s at the US Bureau of the Census (Hall, 2003, Section 3.1, and Wolter, 1985; see also Hartigan, 1969). However the bootstrap is the most flexible and powerful of these procedures, and for that reason is widely used in applications. The purpose of this document is to review published work on the application of these ideas in the survey context, with an emphasis on issues of implementation and practical matters rather than on theory.

Chapter 2 briefly outlines the type of estimators considered and the calibration and imputation procedures that may be used to modify them. Chapter 3 describes linearization approaches, and is followed by outlines of the application of the jackknife, balanced repeated replication, and the bootstrap in the survey context. Numerical comparisons of the main resampling methods are reviewed in Chapter 7, and the report concludes with a brief discussion.

Chapter 2

Preliminaries

2.1 Parameters and estimators

In order to establish notation we consider initially the case of complete response for a stratified single stage unequal probability sampling scheme without replacement, with N units divided into H strata, from which n units are sampled. Let n_h be the number of units sampled from the N_h population units in stratum h, and let π_{hi} be the inclusion probability for unit i of this stratum. In household surveys this unit might consist of a cluster of individuals, in which case we suppose that the unit response of interest is accumulated over the cluster. Thus the total numbers of population and of sampled units are $N = \sum_{h=1}^{H} N_h$ and $n = \sum_{h=1}^{H} n_h$. Let x_{hi} and y_{hi} be variables that have been measured on the units, where y_{hi} is the scalar response of interest and x_{hi} is a $q \times 1$ vector of auxiliary variables, which may be continuous, categorical, or both.

Parameters of the finite population can be classified into two broad groups. The first and most important group comprises quantities that are smooth functions of the finite population responses, such as the total

$$\tau = \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi},$$

the ratio of two population totals for two different variables,

$$\psi = \tau_1/\tau_2, \tag{2.1}$$

or the change in the ratio between two different sampling occasions,

$$\delta(t_1, t_2) = \psi_{t_1} - \psi_{t_2}. \tag{2.2}$$

The other main group of parameters comprises non-smooth functions of the finite population responses, such as the median

$$y_{0.5} = \text{median}\left\{y_{hi} : i = 1, \ldots, N_h, h = 1, \ldots, H\right\},$$

other quantiles, and statistics based on them. An example is $F(ay_b)$, where F is the distribution function of household income for a population of households; taking $a = b =$

0.5 gives the proportion of households whose income is below one-half of the median, sometimes called the low-income proportion.

Estimation of the finite population parameters is based on the data from the n sampled units and on their inclusion probabilities under the given sampling design. The most important estimator of a total is the Horvitz–Thompson estimator

$$\widehat{\tau} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \omega_{hi} y_{hi} = \omega^{\mathrm{T}} y, \tag{2.3}$$

say, where y is the $n \times 1$ vector of sampled responses and ω is the $n \times 1$ vector of their weights $\omega_{hi} = 1/\pi_{hi}$, the inverse inclusion probabilities. The exact variance of this estimator involves the joint inclusion probabilities for the different units and is readily obtained. It can be used to develop variance formulae for estimators of parameters such as (2.2) and (2.1) above; for example, but complications arise when the weights themselves are random, or when some of the responses are unavailable, as we now see.

2.2 Calibration

In many cases population totals are known for some of the auxiliary variables x, and this information can be used to increase precision of estimation, for example by making some allowance for unit non-response. Suppose that q_C marginals of the q auxiliary variables are known, with $q_C \leq q$, let c be the $q_C \times 1$ vector of known marginals, and let X_C denote the $n \times q$ matrix of auxiliary variables whose marginal total for the entire population is known to equal c. Using the estimation of a total to illustrate the idea of calibration, the quality of the Horvitz–Thompson estimator can be improved by calibrating the weights w_{hi} to be as close as possible to the original weights ω in some metric G, subject to the constraint that the weighted auxiliary variables match the marginals (DEVILLE and SÄRNDAL, 1992). One mathematical formulation of this is as the optimisation problem

$$\min_{w_{hi}} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \omega_{hi} G(w_{hi}/\omega_{hi})$$

subject to the constraint that

$$\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} x_{Chi} = c.$$

A common distance measure used in practice is the ℓ_2 or squared error metric, $G(x) = (x-1)^2/2$; in that case the calibrated weights equal

$$w = \omega + \Omega X_C (X_C^{\mathrm{T}} \Omega X_C)^{-1} (c - X_C^{\mathrm{T}} \omega), \tag{2.4}$$

where Ω denotes the diagonal matrix whose elements are the ω_{hi}. So the calibrated Horvitz–Thompson estimator of the total is

$$\begin{aligned} \widehat{\tau} &= w^{\mathrm{T}} y = \omega^{\mathrm{T}} y + (c - X_C^{\mathrm{T}}\omega)^{\mathrm{T}} (X_C^{\mathrm{T}} \Omega X_C)^{-1} X_C^{\mathrm{T}} \Omega y \\ &= \omega^{\mathrm{T}} y + (c - X_C^{\mathrm{T}}\omega)^{\mathrm{T}} \widehat{\gamma}, \end{aligned} \tag{2.5}$$

say, $\widehat{\gamma}$ being the regression estimator when y is regressed on X_C with weight matrix Ω. The second term of $\widehat{\tau}$ is an adjustment to the weights ω that accounts for the difference between $X_C^{\mathrm{T}}\omega$ and the target population value c of the margins.

A wide range of other measures of the distance between w and ω can be envisaged. For instance, the multiplicative measure $G(x) = x \log x - x + 1$ DEVILLE *et al.* (1993) ensures that the calibrated weights, like the original weights, are positive. The results of using various G metrics are asymptotically equivalent, however DEVILLE and SÄRNDAL (1992), and in practice DEVILLE *et al.* (1993) have observed that, although individual weights can vary a lot between methods, the choice of G has a relatively minor impact on point and variance estimates. D'ARRIGO and SKINNER (2003) confirm that although the assumptions on which the asymptotic equivalence of these methods is based often fail in practice, the effect on the overall conclusion of the calibration method is generally small. Since calculating weights for more complex measures such as the multiplicative measure entails more computational effort, our investigations focus mainly on the calibrated weights (2.4) with the ℓ_2 metric, which involves only a regression computation. Some of the numerical studies described below have used other forms of calibration, such as iterative proportional fitting, but the conclusions are very similar to those from using the ℓ_2 metric.

2.3 Imputation

In practice the application of survey methods rarely yields complete data sets, and observations are often missing due to an unknown non-response mechanism. There are two main types of non-response, unit nonresponse, in which all the data for a sampled unit are missing, and item nonresponse, in which some though not all of the data for the unit are missing. Unit nonresponse is often dealt with by calibration or another form of weight adjustment, but item nonresponse allows the possibility of imputation of the missing data, using an imputation model. This model is used to predict the missing responses for units with item non-response, and is generally constructed on the basis of the data for those respondents with complete data. It will rarely be the same as the mechanism that led to the non-response, and there may be bias if it is not. The imputation is said to be proper (RUBIN, 1996) if the point and variance estimators are approximately unbiased for their complete-data analogues. Imputation models and the imputed values they provide are of two types: deterministic, meaning that the model imputes a deterministic function of the observed data; and stochastic, meaning that the model imputes a random function of the observed data. 'Hot deck' imputation, whereby data from non-responding unit are replaced by values sampled from responding units somehow chosen to provide a close match to the missing ones - for example, from the same stratum - is one form of stochastic imputation.

A common deterministic approach to imputation of missing responses based on the corresponding vectors x_{hi} of auxiliary variables is to use a linear model or one of its generalizations, such as robust or logistic regression. All such models are M-estimators, and

the corresponding normal equations for an estimate of the parameters θ of the chosen imputation model across strata may be written as the vector equation

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h} x_{hi}\psi(y_{hi}, x_{hi}; \theta) = 0, \tag{2.6}$$

where ψ is the derivative of the implied loss function with respect to θ. The form of ψ depends upon this model; for logistic regression the y_{hi} are binary indicator variables and

$$\psi(y, x; \theta) = y - \frac{\exp(x^{\mathrm{T}}\theta)}{1 + \exp(x^{\mathrm{T}}\theta)},$$

whereas in a linear model settting, $\theta = (\beta, \sigma)$, where σ is a scale parameter, and

$$\psi(y, x; \theta) = \psi\left\{(y - x^{\mathrm{T}}\beta)/\sigma\right\};$$

setting $\psi(u) = u$ yields the ℓ_2 loss function. A further special case is ratio imputation using a scalar x, obtained by setting $\psi(y, x; \theta) = y - \theta x$. A more robust imputation model is obtained by using Huber's Proposal 2 (HUBER, 1981). This uses the robustified loss function

$$\begin{cases} u^2/2, & |u| \le \tau, \\ \tau \cdot |u| - \tau^2/2, & |u| > \tau, \end{cases}$$

whose derivative is

$$\psi(u) = \mathrm{sign}(u) \cdot \min(|u|, \tau);$$

the quantity τ controls the degree of robustness of the fit, with $\tau \to \infty$ recovering the fragile least squares estimator, and $\tau \to 0$ giving higher robustness. The scale parameter σ may be estimated using a simultaneous estimating equation, or separately; this is rarely crucial. Once the linear model M-estimate $\widehat{\beta}$ of β has been found, the missing response for an individual with explanatory variable x can be predicted by $x^{\mathrm{T}}\widehat{\beta}$. Similar comments apply to logistic regression, which would be used for predicting the value of a missing binary response.

For the numerical computations below we assume either a common linear model across strata,

$$Y_{hi} = x_{hi}^{\mathrm{T}}\beta + \epsilon_{hi}, \tag{2.7}$$

or a different linear model for each stratum,

$$Y_{hi} = x_{hi}^{\mathrm{T}}\beta_h + \epsilon_{hi}, \quad h = 1, \ldots, H. \tag{2.8}$$

We then fit this model to those individuals for which both x_{hi} and y_{hi} are available, and then use the fitted linear model to impute predicted responses for individuals for whom y_{hi} is missing. Let $z_{hi} = I(y_{hi} \neq \texttt{NA})$ be the indicator random variable corresponding to observed response, let $Z = \mathrm{diag}(z)$ be the $n \times n$ diagonal matrix of these indicators, and let X be the $n \times q$ matrix that contains the auxiliary variables corresponding to both respondents and nonrespondents. Also let $\widehat{y} = X\widehat{\beta}$ represent the $n \times 1$ vector of fitted values from the regression model used for imputation. Then (2.4) implies that the calibrated and imputed Horvitz–Thompson estimator may be written as

$$\begin{aligned}\widehat{\tau} &= w^{\mathrm{T}}\{Zy + (I-Z)\widehat{y}\} = \omega^{\mathrm{T}}Zy + (c - X_C^{\mathrm{T}}\omega)^{\mathrm{T}}(X_C^{\mathrm{T}}\Omega X_C)^{-1}X_C^{\mathrm{T}}\Omega Zy \\ &\quad + \omega^{\mathrm{T}}(I-Z)X\widehat{\beta} + (c - X_C^{\mathrm{T}}\omega)^{\mathrm{T}}(X_C^{\mathrm{T}}\Omega X_C)^{-1}X_C^{\mathrm{T}}\Omega(I-Z)X\widehat{\beta}. \qquad (2.9)\end{aligned}$$

As the preceding discussion implies, only minor modifications would be needed if instead it was desired to implement imputation models based on M-estimators.

Computation of the variance of the calibrated and imputed Horvitz–Thompson estimator as if the imputed responses $\widehat{y}$ were true responses can lead to considerable underestimation of the true variance, so the variance estimation technique must reflect the variance inflation due to imputation. This is relatively easily accomplished using a resampling method, because an estimator such as (2.9) is simply treated like any other estimator. A simpler less computationally-demanding and therefore rather tempting approach in practice is to use a standard linearization formula as if the imputed values had been observed, but this can result in severe underestimation of the variance and is not recommended. A powerful variant is multiple imputation (Rubin, 1987), whereby standard formulae are computed for several datasets for which missing data have been stochastically imputed, and are then combined in such a way as to make proper allowance for the effect of imputation. This approach has been regarded as controversial for surveys by certain authors; see for example Fay (1996). Its implementation in the DACSEIS context is discussed by Rässler (2004). Further recent discussion of multiple imputation can be found in Zhang (2003), Nielsen (2003), and the subsequent commentaries.

With a proper imputation, the point and variance estimators are approximately unbiased. In practice, it is impossible to know whether the imputation approach is proper or not, because the missing data mechanism is unknown. As is commonly the case in statistics, the models whose use is outlined above are simply crude approximations to a much more complex reality, which is unlikely to be entirely captured by any model conceived by the data analyst. For this reason Rubin (1996) recommands including as many auxiliary variables as possible, the idea being to approximate this reality as well as feasible with the data at hand, even at the price of increasing the variance of the eventual estimates. It is natural to wonder whether some empirical approach analogous to bandwidth selection could be used to trade off the resulting variance inflation against the bias from using an ill-specified imputation model, but we shall not discuss this here.

2.4 Discussion

For simple sample surveys textbooks such as Cochran (1977) contain formulae giving estimates of the variances of a wide variety of estimators. The complexity of modern sample survey procedures, however, often does not allow the derivation of simple formulae, and a possible avenue is to turn to resampling methods such as the jackknife or the bootstrap. The adaptation of resampling-based variance estimation methods to the survey setting requires special care, because it must take into account the complex dependence structures induced by the probability sampling scheme as well as any calibration and imputation procedures. Otherwise, the variance may be severely under- or over-estimated.

Chapter 3

Linearization

3.1 Taylor linearization

Many estimators may be expressed as a differentiable function $\widehat{\theta} = g(\widehat{\tau})$ of a vector of linear estimators $\widehat{\tau}$. An example is the ratio estimator $\widehat{\theta} = g(\widehat{\tau}_{Y_1}, \widehat{\tau}_{Y_2})$ with $g(x_1, x_2) = x_1/x_2$, which is differentiable whenever the denominator is non-zero. Linear Taylor series expansion of g about the population mean τ of $\widehat{\tau}$ yields

$$\widehat{\theta} = g(\widehat{\tau}) \doteq g(\tau) + \nabla g(\tau)^{\mathrm{T}}(\widehat{\tau} - \tau),$$

whose variance is $\nabla g(\tau)^{\mathrm{T}} \mathrm{var}(\widehat{\tau}) \nabla g(\tau)$. The variance estimator is obtained by replacing unknowns in this formula with the estimators $\nabla g(\widehat{\tau})$ and the empirical covariance matrix for $\widehat{\tau}$. When applied to stratified samples, the resulting variance estimator is

$$v_L = \nabla g(\widehat{\tau})^{\mathrm{T}} \widehat{\mathrm{var}}(\widehat{\tau}) \nabla g(\widehat{\tau}) = \nabla g(\widehat{\tau})^{\mathrm{T}} \left(\sum_{h=1}^{H} \widehat{V}_h \right) \nabla g(\widehat{\tau}), \tag{3.1}$$

where $\widehat{V}_h$ is the contribution to the estimated covariance matrix for $\widehat{\tau}$ from the h-th stratum. If g is linear, then v_L is an unbiased estimator of the variance of $\widehat{\theta}$. Under mild conditions, Krewski and Rao (1981) show that v_L is a consistent estimator of the asymptotic variance of $\widehat{\theta}$.

Sometimes it is not possible to express the function g in closed form, for instance when the point estimator is defined as the solution of an estimating equation when complicated calibration or imputation are involved. Rao (1988), Binder (1996) and Deville (1999) discuss linearization in such implicit parameter cases.

More details of such estimators are given in deliverable D1.1.

3.2 Jackknife linearization

A somewhat more general approach to construction of linearization variance estimators for a compactly differentiable statistic $\widehat{\theta}$ is though the influence function components

(HAMPEL *et al.*, 1986), which may be interpreted as derivatives of $\widehat{\theta}$ with respect to changes in the weights placed on individual observations; for this reason the technique is sometimes known as the infinitesimal jackknife. The extra generality stems from use of a von Mises rather than Taylor series expansion of the statistic $\widehat{\theta}$, enabling theoretical variance formulae to be obtained for estimators such as the sample median and other quantiles. See for example BERGER and SKINNER (2003), who use kernel density procedures to estimate the density function that arises in these formulae.

In many cases the estimand θ can be written as a functional $t(F)$ of the underlying distribution function F which generated the data. A simple estimator of $t(F)$ is then $t(\widehat{F})$, where $\widehat{F}$ is the empirical distribution function of the data. For the mean, for instance, $t(F) = \int y dF(y)$ and $t(\widehat{F}) = \overline{Y}$ is its empirical analogue. In the case of simple random sampling with replacement, and assuming some differentiability properties for the functional $t(\cdot)$, the estimate $\widehat{\theta} = t(\widehat{F})$ can be expanded around $\theta = t(F)$ using

$$t(\widehat{F}) \doteq t(F) + n^{-1} \sum_{i=1}^{n} L_t(Y_i; F),$$

where

$$L_t(y; F) = \lim_{\epsilon \to 0} \frac{t\{(1-\epsilon)F + \epsilon\delta_y\} - t(F)}{\epsilon}$$

is the *influence function* for $t(\widehat{F})$, δ_y being the distribution function putting a point mass at y. This expansion can be used to establish that the estimator is asymptotically unbiased and Gaussian. Its variance $v_L(F) = n^{-1}\text{var}\{L_t(Y; F)\}$ can be estimated by

$$\widehat{v}_L = n^{-2} \sum_{i=1}^{n} l_i^2, \tag{3.2}$$

where $l_i = L_t(y_i; \widehat{F})$ are the *empirical influence values* for the statistical functional t evaluated at the observation y_i and the empirical distribution function $\widehat{F}$. Here l_i can be thought of as the derivative of t at $\widehat{F}$ in the direction of a distribution putting more mass on the i-th observation. Section 2.11 of DAVISON and HINKLEY (1997) gives more details of these expansions and examples of these calculations, and CAMPBELL (1980) outlines their extension to finite population situations.

For stratified random sampling without replacement (3.2) may be modified to

$$v_L = \sum_{h=1}^{H} (1 - f_h) \frac{1}{(n_h - 1) n_h} \sum_{i=1}^{n_h} l_{hi}^2, \tag{3.3}$$

where l_{hi} is the empirical influence value corresponding to the ith observation in stratum h. In simple situations (3.2) and (3.3) recover standard linearization formulae, but the use of influence functions gives a more general approach to variance estimation.

3.3 Horvitz–Thompson estimator

We now consider the Horvitz–Thompson estimator and give formulae for its empirical influence functions for stratified sampling in three situations of increasing complexity:

- the standard Horvitz–Thompson estimator given at (2.3), for which

$$l_{hi} = n_h \omega_{hi} y_{hi} - \omega_h^{\mathrm{T}} y_h;$$

- the calibrated Horvitz–Thompson estimator (2.5), for which

$$\begin{aligned} l_{hi} &= (n_h \omega_{hi} y_{hi} - \omega_h^{\mathrm{T}} y_h) + (X_{Ch}^{\mathrm{T}} \omega_h - n_h \omega_{hi} x_{Chi})^{\mathrm{T}} \widehat{\gamma} \\ &\quad + n_h \omega_{hi} (c - X_C^{\mathrm{T}} \omega)^{\mathrm{T}} (X_C^{\mathrm{T}} \Omega X_C)^{-1} x_{Chi} (y_{hi} - x_{Chi}^{\mathrm{T}} \widehat{\gamma}), \end{aligned}$$

  where ω_h and y_h are $n_h \times 1$ vectors of the weights and responses for the h-th stratum, X_{Ch} is the $n_h \times q_C$ matrix of calibration covariates for the h-th stratum, and $\widehat{\gamma} = (X_C^{\mathrm{T}} \Omega X_C)^{-1} X_C^{\mathrm{T}} \Omega y$; and

- the calibrated Horvitz–Thompson estimator (2.9) with imputation of missing responses. Let

$$\widehat{\gamma}_M = (X_C^{\mathrm{T}} \Omega X_C)^{-1} X_C^{\mathrm{T}} \Omega (I - Z) \widehat{y}$$

  correspond to $\widehat{\gamma}$, but for those individuals with missing responses, and let $l_i(\widehat{\beta})$ be the elements of the $q \times 1$ vector of influence functions for the imputation regression coefficients, corresponding to differentiation with respect to the ith case in stratum h. Then for $i = 1, \ldots, n_h$ and $h = 1, \ldots, H$,

$$\begin{aligned} l_{hi} &= (n_h \omega_{hi} z_{hi} y_{hi} - \omega_h^{\mathrm{T}} Z_h y_h) + (X_{Ch}^{\mathrm{T}} \omega_h - n_h \omega_{hi} x_{Chi})^{\mathrm{T}} \widehat{\gamma} \\ &\quad + n_h \omega_{hi} (c - X_C^{\mathrm{T}} \omega)^{\mathrm{T}} (X_C^{\mathrm{T}} \Omega X_C)^{-1} x_{Chi} (z_{hi} y_{hi} - x_{Chi}^{\mathrm{T}} \widehat{\gamma}) \\ &\quad + \{n_h \omega_{hi} (1 - z_{hi}) \widehat{y}_{hi} - \omega_h^{\mathrm{T}} (I_h - Z_h) \widehat{y}_h\} + \omega^{\mathrm{T}} (I - Z) X l_i(\widehat{\beta}) \\ &\quad + (X_{Ch}^{\mathrm{T}} \omega_h - n_h \omega_{hi} x_{Chi})^{\mathrm{T}} \widehat{\gamma}_M \\ &\quad + n_h \omega_{hi} (c - X_C^{\mathrm{T}} \omega)^{\mathrm{T}} (X_C^{\mathrm{T}} \Omega X_C)^{-1} x_{Chi} \{(1 - z_{hi}) \tilde{y}_{hi} - x_{Chi}^{\mathrm{T}} \widehat{\gamma}_M\} \\ &\quad + (c - X_C^{\mathrm{T}} \omega)^{\mathrm{T}} (X_C^{\mathrm{T}} \Omega X_C)^{-1} X_{Ch}^{\mathrm{T}} \Omega_h (I_h - Z_h) X_h l_i(\widehat{\beta}). \end{aligned} \tag{3.4}$$

  Imputation model (2.7) for least squares deterministic regression, for instance, yields

$$l_i(\widehat{\beta}) = n_h z_i (X^{\mathrm{T}} Z X)^{-1} x_i (y_i - x_i^{\mathrm{T}} (X^{\mathrm{T}} Z X)^{-1} X^{\mathrm{T}} Z y), \quad i = 1, \ldots, \sum_{h=1}^{H} n_h,$$

  where X is the regression matrix. For the imputation model (2.8), in which the regression coefficients can vary among the strata, the $l_i(\widehat{\beta})$ in (3.4) are taken to be

$$l_i(\widehat{\beta}_h) = n_h z_i (X_h^{\mathrm{T}} Z_h X_h)^{-1} x_i (y_i - x_i^{\mathrm{T}} (X_h^{\mathrm{T}} Z_h X_h)^{-1} X_h^{\mathrm{T}} Z_h y_h),$$

  where X_h, Z_h, and y_h are the covariate matrix, the indicator matrix for observed responses, and the response vector for stratum h.

These formulae are used below in our numerical comparisons, and in particular form the basis of the multiple imputation methods we apply in Section 7.4.

Chapter 4

Jackknife

4.1 Basic ideas

The jackknife, originally introduced as a method of bias estimation by Quenouille (1949a,b) and subsequently proposed for variance estimation by Tukey (1958), involves the systematic deletion of groups of units at a time, the recomputation of the statistic with each group deleted in turn, and then the combination of all these recalculated statistics. More explicitly, the simplest, 'delete-one', jackknife involves dropping each observation of a simple random sample in turn, recomputing the statistic to obtain values $t_{-1}, \ldots, t_{-n}$ in addition to its original value, t, and then estimating the bias of the estimator by

$$\frac{n-1}{n} \sum_{i=1}^{n} (t_{-i} - t) = (n-1)(\bar{t} - t),$$

say, and its variance by

$$\frac{n-1}{n} \sum_{i=1}^{n} (t_{-i} - \bar{t})^2.$$

In order for these estimators to be consistent, the statistic t must be sufficiently smooth as a function of the observations, for example possessing a Taylor series or similar expansion. Certain types of non-smooth estimators, such as the median and other estimators based on quantiles, do not possess a suitable expansion, and so delete-one jackknife estimators of their variances are inconsistent (Efron, 1982). Although this is a major disadvantage from the theoretical viewpoint, a more important practical consideration is the performance of an estimator for the types and sizes of sample met in applications, and it turns out that despite its inconsistency the jackknife variance estimator for sample quantiles based on sample surveys can be competitive with other estimators in terms of mean squared error (Rao *et al*., 1992). Shao and Wu (1989) have shown that the inconsistency can be repaired by deleting groups of d observations, where $d \to \infty$ as $n \to \infty$.

Rao and Wu (1985) showed that linearization and the jackknife are first order asymptotically equivalent, and that in some cases they produce exactly the same variance estimates. Some simulations also tend to show that jackknife is less biased but more variable than

the linearization variance estimator. BERGER and SKINNER (2004) develop theory for jackknife variance estimation for a wide range of estimators under unequal probability sampling; in further unpublished work Y. G. Berger and J. N. K. Rao have extended this to imputation.

One view of the delete-one jackknife is as a numerical means of obtaining the derivatives of t with respect to the observations. Dropping one observation amounts to an $O(n^{-1})$ perturbation of the sample, however, and as sampling variation is typically of $O(n^{-1/2})$, one might suspect that the jackknife perturbations are in some sense too local to give an accurate idea of the variance of the estimator. It turns out that jackknife variance estimators typically slightly underestimate the true variance, and that this underestimation can be attributed to the localness of the Taylor expansion. MILLER (1974) is an excellent review of early work on the jackknife, and a recent more extended account is given by SHAO and TU (1995). More general subsampling estimators, useful particularly for time series and other forms of correlated data, have been proposed by POLITIS *et al.* (1999).

4.2 Modified jackknives

There are several complications in the survey sampling context. A first is that the without-replacement sampling designs typically used in practice result in correlated observations, while a second is that the data are typically stratified. With stratified survey data, the delete-one jackknife involves deleting each primary sampling unit in turn. One way of viewing this is that deletion of the i-th unit in the h-th stratum changes the weight for the jth unit in the kth stratum to

$$\omega_{kj}^{(hi)} = \begin{cases} \omega_{kj}, & \text{if } k \neq h, \\ \omega_{hj} n_h/(n_h - 1), & \text{if } k = h, j \neq i, \\ 0, & k = h, j = i. \end{cases}$$

In the cases of interest here the ratio $n_h/(n_h - 1) \doteq 1$, and so the use of these updated weights will give the statistics $\widehat{\theta}_{h,-i}$, that is $\widehat{\theta}$ calculated without the i-th unit of stratum h, for $i = 1, \ldots, n_h$ and $h = 1 \ldots, H$. Then the delete-one jackknife estimate of variance is given by

$$v_J = \sum_{h=1}^{H} \frac{(1 - f_h)(n_h - 1)}{n_h} \sum_{i=1}^{n_h} (\widehat{\theta}_{h,-i} - \overline{\widehat{\theta}}_h)^2, \qquad \overline{\widehat{\theta}}_h = n_h^{-1} \sum_{i=1}^{n_h} \widehat{\theta}_{h,-i}, \tag{4.1}$$

where $f_h = n_h/N_h$ is the sampling fraction in the h-th stratum, which must be included to account for the sampling plan. RAO and WU (1985) show that replacing $\overline{\widehat{\theta}}_h$ in (4.1) by $n^{-1} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \widehat{\theta}_{h,-i}$ or by $H^{-1} \sum_{h=1}^{H} \overline{\widehat{\theta}}_h$ has little effect on the performance of v_J in the sense that the three versions are second order asymptotically equivalent; KOVAR *et al.* (1988) observe that the resulting three variance estimators are also comparable for finite samples.

A third, practical, complication is that the delete-one jackknife is computationally intensive because a total of n deletions is required; the resulting computational burden is

often too large for this simple jackknife to be useful. Deletion of d sample elements at a time Shao and Wu (1989) would involve $\binom{n}{d}$ recomputations of the statistic, and so is even more infeasible, but it is common to reduce the number of replications needed by splitting stratum h into m_h disjoint blocks of d observations, so that $dm_h = n_h$, and then systematically recomputing the statistic with each of these blocks deleted. The total number of recomputations is thus $m = \sum_{h=1}^{H} m_h$ rather than n or $\binom{n}{d}$. Deletion from the hth stratum of a block whose units have indexes $i \in \mathcal{S}$ changes the weight for the jth unit in the kth stratum to

$$\omega_{kj}^{(h\mathcal{S})} = \begin{cases} \omega_{kj}, & \text{if } k \neq h, \\ \omega_{hj} n_h / (n_h - |\mathcal{S}|), & \text{if } k = h, j \notin \mathcal{S}, \\ 0, & k = h, j \in \mathcal{S}, \end{cases}$$

where $|\mathcal{S}|$ represents the number of elements in $\mathcal{S}$; here d. Variants of this grouped jackknife procedure choose the block members randomly at each recomputation, with or without replacement (Krewski and Rao, 1981; Shao and Tu, 1995, Section 5.2.2). Since sampling with replacement is simple and produces comparable results to sampling without replacement, the sampling with replacement 'delete-d' jackknife seems preferable. For any of these the variance estimate is

$$v_{J,-d} = \sum_{h=1}^{H} \frac{(1 - f_h)(n_h - d)}{n_h} \sum_{i=1}^{m_h} (\widehat{\theta}_{h,-i}^{d} - \overline{\widehat{\theta}_h^{d}})^2,$$

where $\widehat{\theta}_{h,-i}^{d}$ is the estimator computed from the data with the i-th block of size d excluded from the h-th stratum, and $\overline{\widehat{\theta}_h^{d}}$ is the average of these for that stratum. Minor variations on the above formulae are needed when n_h is not exactly divisible by d, in which case one block has size in the range $1, \ldots, d$.

The delete-d jackknife has the further benefit of extending the range of functions to which the jackknife may be applied. As mentioned above, it turns out that if $d, m_h \to \infty$ together as $n_h \to \infty$, then the block jackknife gives valid variance estimators (Shao and Wu, 1989). As with other subsampling estimators, however, it is not clear how to choose d and m_h when n_h is finite. The less smooth the statistic of interest, the larger d needs to be, at least in theory, but a reliable finite-sample method for choosing d remains elusive.

Kim and Sitter (2003) describe an approach to reducing the number of recomputations for a jackknife variance estimator used with two-phase sampling, and give other references to this topic.

4.3 Imputation

Rao and Shao (1992) describe a consistent version of the delete-one jackknife variance estimator using a particular hot deck imputation mechanism to account for non-response; see also Fay (1996) for a wider perspective, and Chen and Shao (2001), who show that this approach fails for another popular imputation scheme, nearest-neighbour imputation. The key idea is that as the value of a responding unit affects both that unit and any

imputed values that use its value, deletion of such a unit must account for this; deletion of a non-responding unit however affects the estimator in the usual way. This can be extended to stratified multi-stage sampling by deletion of each of the primary sampling units. Below we use a slight generalization of this idea by allowing block deletion, that is, by re-imputing each time a block is deleted, using the respondents of the remaining data to estimate the model used for deterministic imputation. As this is simply a version of the estimator based on the respondent data but accounting for the effect of non-response, no new theory is required when the imputation model is adequate.

Chapter 5

Balanced Repeated Replication

5.1 Basic ideas

Balanced half-sampling (McCarthy, 1969) is the simplest form of balanced repeated replication. It was originally developed for stratified multistage designs with two primary sampling units drawn with replacement in the first stage. When a sample consists of two observations in each of H strata, a half-sample is formed by taking one observation from each stratum. The weight ω_{hi} attached to the i-th unit in stratum h is changed to $2\omega_{hi}$ if that unit is included in the half-sample, and to 0 if not, and the statistic is recomputed with the half-sample and these new weights. There is a total of 2^H such replicate statistics, whose results can be combined to estimate the variance of the original statistic. This is clearly computationally infeasible unless H is small, but ideas from experimental design allow the same variance estimate to be obtained with fewer replications, provided they are balanced. Let $\alpha_{hr} = \pm 1$ according to whether the first or second of the two units in the h-th stratum is to be retained in the r-th replicate, and let $\widehat{\theta}_r$ be the value of the estimator $\widehat{\theta}$ computed using the r-th half-sample and its adjusted weights. Then the set of R half-samples is balanced if

$$\sum_{r=1}^{R} \alpha_{hr}\alpha_{kr} = 0, \quad \text{for all} \quad h \neq k, k, h = 1, \ldots, H,$$

and in this case the balanced repeated replication variance estimator

$$v_{\text{BRR}} = R^{-1} \sum_{r=1}^{R} (\widehat{\theta}_r - \widehat{\theta})^2 \tag{5.1}$$

produces the same variance estimate as does computation using the full set of 2^H half-samples, at least for linear statistics such as the total or average. For such statistics, v_{BRR} equals both the jackknife estimator v_J and the linearization estimator. If desired, $\widehat{\theta}$ in (5.1) can be replaced by $R^{-1}\sum_{r=1}^{R}\widehat{\theta}_r$ without changing the asymptotic properties of the variance estimator, while other variants are also possible. When there are no missing observations, v_{BRR} is consistent for the variances of nonlinear functions Krewski and Rao (1981). It turns out that v_{BRR} is stochastically close to the linearization variance

estimator, but not so close to it as is the jackknife variance estimator. The design-consistency of v_{BRR} for estimators of quantiles and those based on quantiles has been established by SHAO and WU (1992) and SHAO and RAO (1994).

A balanced set of half-samples can be obtained for any H by selecting R to be the smallest multiple of four that is greater than or equal to H; thus $H \leq R \leq H+3$. Appropriate values for α_{hr} are given by the entries of an $R \times R$ Hadamard matrix. For instance, the Hadamard matrix useful for $H = 7$ is the 8×8 matrix

$$\alpha = \begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix},$$

from which the half-samples are identified by taking α_{hr} to be the r-th element of row h. Thus strata correspond to rows and replicates to columns; for example the first column identifies the first of the eight replicates as including the first unit from strata 1, 2, 3 and 5, and and second unit from strata 4, 6 and 7. This is then repeated for each of the eight columns, with the elements of the matrix giving the appropriate half-sample for the replicate that corresponds to each column. Appendix A of WOLTER (1985) lists such matrices up to $R = 100$ and gives references to means of constructing larger ones.

Calibration is applied to a half-sample by calibrating the rescaled weights using the population margins.

5.2 Strata of size $n_h > 2$

Two main generalizations to surveys with more than $n_h = 2$ observations per stratum have been proposed. The first, investigated by GURNEY and JEWETT (1975), GUPTA and NIGAM (1987), WU (1991) and SITTER (1993), uses orthogonal arrays but requires a large number of replicates, making it impractical for many applications.

The second generalization, a simpler more pragmatic approach, is to group the primary sampling units in each stratum into two groups of sizes $m_h = \lfloor n_h/2 \rfloor$ and $n_h - m_h$, and to apply balanced repeated replication using the groups rather than individual units (RAO and SHAO, 1996; WOLTER, 1985, Section 3.7). Under this grouped balanced repeated replication scheme the weights for the r-th replicate are adjusted using the formula

$$\omega_{hi}^r = \begin{cases} \omega_{hi}\left[1 + \left\{\frac{(n_h - m_h)(1-f_h)}{m_h}\right\}^{1/2}\right], & \alpha_{hr} = 1, \\ \omega_{hi}\left[1 - \left\{\frac{m_h(1-f_h)}{n_h - m_h}\right\}^{1/2}\right], & \alpha_{hr} = -1, \end{cases} \tag{5.2}$$

where the factor $(1 - f_h)$ is an adjustment for the effect of sampling without replacement and the condition $m_h \leq n_h/2$ ensures that the adjusted weights are positive. The choice

$m_h = n_h/2$ leads to the original weight adjustment of $2\omega_{hi}$ or 0 when f_h can be neglected. The variance estimator is still defined by (5.1), but one effect of the grouping is that it does not reduce to the standard variance estimator in the linear case.

5.3 Improved estimators

The balanced repeated replication variance estimator v_{BRR} can be highly variable, for two main reasons. The first is that the distribution of (5.1) is approximately that of a scaled χ^2_{R-1} variable, that is, its distribution is $c\,\chi^2_{R-1}$, where $c > 0$ is constant. Thus the coefficient of variation of (5.1) is roughly $(2/R)^{1/2}$; this is appreciable for small R, or equivalently when H is small. The coefficient of variation may be reduced by increasing the degrees of freedom, for example by splitting the H true strata into artificial sub-strata, to which half-sampling is then applied.

The second difficulty is that certain statistics may be highly sensitive to perturbations of the weights, and if so it can be impossible to compute all the replicate estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_R$. This occurs for example with ratio estimators, where the denominator may become zero for some reweightings, and where deleting half the sample may result in smaller domain sample sizes than would have been tolerated in the original survey. A solution to this suggested by Robert Fay of the US Bureau of the Census Dippo *et al.* (1984); Fay (1989) is to use a milder reweighting scheme, whereby a factor ε is chosen, with $0 < \varepsilon \leq 1$, and the weights are perturbed to

$$\omega^r_{hi} = \begin{cases} \omega_{hi}\left[1 + \varepsilon\left\{\frac{(n_h - m_h)(1-f_h)}{m_h}\right\}^{1/2}\right], & \alpha_{hr} = 1, \\ \omega_{hi}\left[1 - \varepsilon\left\{\frac{m_h(1-f_h)}{n_h - m_h}\right\}^{1/2}\right], & \alpha_{hr} = -1, \end{cases}$$

resulting in replicate estimates $\widehat{\theta}_1(\varepsilon), \ldots, \widehat{\theta}_R(\varepsilon)$ and variance estimator

$$v_{\mathrm{BRR}}(\varepsilon) = \frac{1}{R\varepsilon^2}\sum_{r=1}^{R}\left\{\widehat{\theta}_r(\varepsilon) - \widehat{\theta}\right\}^2.$$

Setting $\varepsilon = 1$ recovers weights (5.2) and variance estimator (5.1), but other values of ε amount to downweighting rather than deleting groups of units, and so greatly reduce instability of the replicate estimates. Rao and Shao (1999) studied theoretical aspects of these weight modifications and considered also their extension to imputation for item non-response. Their main conclusions are: that the milder weighting schemes give consistent variance estimation under the same conditions as the original approach, that is, with $\varepsilon = 1$; that when $\widehat{\theta}$ is a smooth function of totals $v_{\mathrm{BRR}}(\varepsilon)$ converges to the Taylor linearization variance estimator when $\varepsilon \to 0$; that the difference between the two is small when ε is a function of sample size n; and that suitably applied imputation schemes lead to consistent variance estimation using $v_{\mathrm{BRR}}(\varepsilon)$. Numerical work Judkins (1990); Rao and Shao (1999) suggests that good results are obtained with $\varepsilon = 0.5$; in particular the variance estimator of the median using ratio imputation then has much better finite-sample properties than does setting $\varepsilon = 1$.

5.4 Repeatedly grouped balanced repeated replication

When H is large, grouped balanced repeated replication provides a consistent variance estimator, but this can be very variable if $H \ll n$. Moreover it is inconsistent when H is fixed and the $n_h \to \infty$ RAO and SHAO (1996), suggesting that it may work poorly for small H. This has been observed empirically, and, as mentioned above, is due partly to the fact that when H is fixed the variance estimator v_{BRR} has a distribution close to scaled χ^2_{R-1}. To reduce the variability of v_{BRR} and to remove the effect of the particular random split of the stratum units into groups, RAO and SHAO (1996) proposed repeating the method over differently randomly selected groups to provide m estimates of variance, averaging of which will provide a more stable overall variance estimate. This repeatedly grouped balanced repeated replication variance estimator turns out to be consistent as $mR \to \infty$. In particular, choosing m such that $mR \doteq n$ leads to a quite efficient variance estimator, which implies that m should be large when the number of strata H is small. But when H is large, the rule suggests choosing only $m = 1$. The obvious drawback to this is that computing the estimator $\widehat{\theta}$ for $O(n)$ replicates is generally too burdensome for practical application, where typically one must take $mR \ll n$. Simulations suggest that the repeatedly grouped balanced repeated replication variance estimator has similar performance to the grouped jackknife, and that both outperform the grouped balanced repeated replication variance estimator, which has a high variance. Variants of these procedures include random subsampling of units to form groups, followed by application of the usual variance formulae (KOVAR *et al*., 1988).

5.5 Imputation

The work of RAO and SHAO (1999) on the behaviour of $v_{\mathrm{BRR}}(\varepsilon)$ when unit non-response occurs has already been mentioned. SHAO *et al.* (1998) adjust balanced repeated replication to the presence of nonresponse, by taking into account a deterministic or random imputation mechanism. Under a general stratified multistage sampling design, they establish consistency of the adjusted balanced repeated replication variance estimators for functions of smooth and nonsmooth statistics.

Chapter 6

Bootstrap

6.1 Basic ideas

The bootstrap has been extensively studied for independent and identically distributed data, in which case sampling is done with replacement, but its adaptation to the complex survey setting is not straightforward, because the probability sampling scheme of the survey induces dependence. Calibration and imputation must also be taken into account when bootstrapping the survey sample.

The bootstrap idea is to mimic how the original data were generated. Like the balanced repeated replication and the jackknife methods, the bootstrap involves recomputing the statistic, now using resampling from an estimated population $\widehat{F}$ to obtain bootstrap samples that may be represented by $\widehat{F}^*$, giving corresponding statistics $\widehat{\theta}^* = t(\widehat{F}^*)$. Repeating this process R times independently, the bootstrap estimate of variance is given by

$$v_B = (R-1)^{-1} \sum_{r=1}^{R} (\widehat{\theta}_r^* - \overline{\widehat{\theta}^*})^2, \quad \overline{\widehat{\theta}^*} = R^{-1} \sum_{r=1}^{R} \widehat{\theta}_r^*.$$

For stratified data, the resampling is performed independently within each stratum. The standard bootstrap uses sampling with replacement, corresponding to independent sampling from an original population, but this does not match the without-replacement sampling generally used in the survey context, and the result is that the finite sampling correction is missed, leading to a biased variance estimator. The simplest context in which to see this is when estimating the mean of a population of size N by equal-probability sampling without replacement. Then an unbiased estimator of the design variance of the sample average $\overline{y} = n^{-1} \sum_{i \in \mathcal{S}} y_i$, where $\mathcal{S} \subset \{1, \ldots, N\}$ contains the n sampled indices, is

$$(1-f)\frac{s^2}{n}, \quad s^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (y_i - \overline{y})^2,$$

where $f = n/N$ is the sampling fraction. The simplest bootstrap scheme takes samples of size n with replacement from $\{y_i : i \in \mathcal{S}\}$, and uses the sampling properties of the

bootstrap average $\overline{y}^*$ to estimate those of $\overline{y}$ under repetition of the original sampling design. The bootstrap variance of $\overline{y}^*$ is

$$\mathrm{var}^*(\overline{y}^*) = \frac{n-1}{n}\frac{s^2}{n},$$

however, and in general this does not equal $\mathrm{var}(\overline{y})$ unless n is large and the sampling fraction is small, or unless by happenstance $f = 1/n$. Essentially the same problem arises with stratified sampling, in which case the variance of an overall estimator is a sum of contributions from individual strata; see (4.1). This implies that the standard bootstrap variance estimator may be inconsistent under sampling plans under which $H \to \infty$ but the sampling fractions f_h are bounded away from zero. One viewpoint is that this asymptotic failure is unimportant if the bootstrap variance estimator has low mean squared error for the type of survey met in practice. This failure of the conventional bootstrap has spurred a good deal of work on modified bootstraps, which we outline below.

EFRON (1982) recognized the scaling problem of small strata, and suggested the simple solution of taking bootstrap samples of size $n_h - 1$, but this does not deal with the effect of non-negligible sampling fractions f_h, for which more elaborate approaches are needed, and which we now describe.

6.2 Without-replacement bootstrap

The simplest and most satisfactory approach in the case of an unstratified sample is to create a pseudopopulation using N/n replicates of the sample, and then to sample from this artificial population without replacement (GROSS, 1980, and CHAO and LO, 1985). For stratified samples (BICKEL and FREEDMAN, 1984, BOOTH *et al.*, 1994, and BOOTH and HALL, 1994), the h-th stratum is replicated N_h/n_h times, and then stratified sampling without replacement is applied separately in each stratum. BICKEL and FREEDMAN (1984) generalize the method to the situation where N_h/n_h is not an integer, by randomization between two different population sizes, and derive some theoretical properties. MCCARTHY and SNOWDEN (1985) give some situations where this approach is not feasible, e.g., when $f_h^3 < 1/N_h$ for some h.

SITTER (1992b) modifies the bootstrap method to circumvent the problems of the variance estimator of BICKEL and FREEDMAN (1984) by selecting without-replacement samples of size m_h in each pseudopopulation of size $k_h n_h = N_h$, where

$$m_h = n_h - (1 - f_h) \quad \text{and} \quad k_h = (1 - \tfrac{1-f_h}{n_h})/f_h, \tag{6.1}$$

in order to obtain an unbiased and consistent variance estimator, at least for a linear statistic. SITTER (1992b) gives details of the adjustments needed for non-integer m_h and k_h.

When the inverse sampling fraction N_h/n_h is non-integer, all these methods require randomisation between pseudopopulations whose sizes bracket N_h. The details of the randomisation are important, and in addition to being somewhat clumsy and awkward to program, mistakes can lead to failure of the methods PRESNELL and BOOTH (1994). Moreover the need for a pseudopopulation for each stratum entails additional storage, which is not practical for large surveys.

6.3 With-replacement bootstrap

Without creating a pseudopopulation, McCarthy and Snowden (1985) propose to mimic the original sampling design by selecting samples of size m_h with replacement in each stratum with

$$m_h = (n_h - 1)/(1 - f_h) \tag{6.2}$$

so as to get an unbiased and consistent variance estimator for a linear estimator. When m_h is not an integer, a randomization is again required.

6.4 Rescaling bootstrap

Rao and Wu (1988) also propose to bootstrap with replacement, but they match the standard variance by rescaling the bootstrap sample of size m_h according to

$$y_h^* = \overline{y}_h + \sqrt{m_h(1 - f_h)/(n_h - 1)}(\overline{y}_h^* - \overline{y}_h),$$

where m_h can be any positive integer. This includes Efron (1982)'s method as a special case. They show that the distribution estimators are consistent. To match third order moments, Rao and Wu (1988) suggest taking

$$m_h = \lfloor (1 - f_h)(n_h - 2)^2/(1 - 2f_h)/(n_h - 1) \rfloor,$$

which avoids the problem of non-integer m_h as in (6.1) and (6.2). The method also has the advantage of not creating and storing pseudopopulations, but it has some potential drawbacks, notably that for $m_h \geq n_h$, it can yield impossible values of the bootstrapped statistic $\widehat{\theta}^*$.

6.5 Mirror-match bootstrap

The mirror-match bootstrap (Sitter, 1992a) is a hybrid between the with replacement bootstrap and the without replacement bootstrap methods. In each stratum, it draws a sample of size $n_h^* < n_h$ without replacement $k_h = n_h(1 - f_h^*)/n_h^*/(1 - f_h)$ times. To match third order moments, Sitter (1992a) suggests $n_h^* = f_h n_h$. The mirror-match bootstrap variance estimators is consistent for linear statistics. It is equivalent to the with-replacement bootstrap of McCarthy and Snowden (1985) if $n_h^* = 1$. A drawback is that a randomization between the bracketing integers is needed when k_h is not an integer (Sitter, 1992a).

6.6 Non-response

When responses are missing, the imputation mechanism must be applied to each resample $\widehat{F}^*$ (Shao and Sitter, 1996). Thus we must re-impute repeatedly using the respondents

of the bootstrapped sample to fit the imputation model and then impute the nonrespondents of the bootstrap sample. The method is therefore rather computer intensive, but on the other hand gives consistent variance estimators for medians and other estimators based on quantiles.

6.7 Comments

Section 3.7 of DAVISON and HINKLEY (1997) contains additional comments and examples. SITTER (1992b) used simulation to compare various bootstrap methods with and without replacement, and found that they all perform well and are comparable. The considerable added complexity of the mirror-match and rescaling methods leads us to prefer simple bootstrap resampling in cases where the sampling fraction is small, as is generally the case for DACSEIS, and otherwise to use the pseudopopulation approach.

LAHIRI (2003) and SHAO (2003) review how bootstrap methods may be applied to complex survey data, while RUST and RAO (1996) describe the application of replication methods in this context.

PRESNELL and BOOTH (1994) point out theoretical difficulties with many of the published approaches.

Chapter 7

Empirical Comparisons

7.1 Simple cases

Various simulations have been done to investigate the relative performance of the variance estimators discussed. Kovar *et al.* (1988) considered stratified one stage simple random sampling with replacement based on pseudo-data that resemble real populations with $H = 32$ strata from which samples of sizes n_h either 2 or 5 are taken. The parameters they consider are the ratio of population means, the regression coefficient, the correlation coefficient between the two variables, and the population median. Based on a measure of bias and stability, they compared the jackknife, the balanced repeated replication, the with-replacement bootstrap by rescaling and the linearization methods. They observe that the linearization and jackknife variance estimators perform equivalently well and have the best performance in terms of relative bias and stability, that balanced repeated replication has the second-best performance, and that the rescaling bootstrap of Rao and Wu (1988), the only bootstrap method they consider, has the worst performance overall. Balanced repeated replication and this bootstrap tend to overestimate the variances of nonlinear statistics. They also consider the coverage of confidence intervals, and conclude that jackknife and linearization intervals tend to under-cover, while those based on balanced repeated replication tend to over-cover. Studentized bootstrap confidence intervals improve the situation, but also tend to over-cover.

Rao and Shao (1996) compare the group balanced repeated replication and the repeatedly group balanced repeated replication methods when $n_h = 48$ in each of the $H = 5$ strata; the sampling fraction was roughly 1/10. Single stage simple random sampling with replacement was used within each stratum. They too considered various parameters, and observed that the jackknife works best when the point estimator is smooth but works poorly for the sample median; from a theoretical viewpoint this is unsurprising as the jackknife variance estimator is inconsistent for a sample quantile. The group balanced repeated replication method has a small relative bias, but is very unstable. This instability is corrected by repeatedly-grouped balanced repeated replication, but at the cost of more computation.

Sitter (1992b) compared various bootstrap methods, namely the with-replacement bootstrap, the rescaling bootstrap, a version of the mirror-match bootstrap and a version of

the without-replacement bootstrap, under stratified one stage simple random sampling without replacement, and concluded that the bootstrap methods are comparable for estimation of the variance of the median.

7.2 Calibration

Dutch study

As part of the DACSEIS project, BOONSTRA and NIEUWENBROEK (2003) conducted a comparison of Taylor linearization and balanced repeated replication variances for generalized regression estimators of totals and ratios of totals from two different surveys, one a two-per-stratum design drawn from a population of businesses, and the other a two-stage design drawn from a population of persons. The Taylor linearization formulae used were standard ones, while different variants of balanced repeated replication were used, including grouping of primary sampling units into two groups for each stratum, and use of artificial strata obtained by random division of each stratum of primary sampling units into sub-strata, each of which is divided into two random groups. A similar approach was applied within each of the first-level strata, leading to a more complex half-sampling scheme also involving milder perturbations of the original weights through a Fay factor ε, as outlined in Section 5.3.

The first simulation of BOONSTRA and NIEUWENBROEK (2003) involves 2000 samples taken from 84 strata whose sizes N_h lie in the range 10–20: two observations were taken from each stratum for a total of 168 observations. Calibration using the values of two categorical variables was performed using a regression estimator of the type described in Section 2.2, and variance estimators were computed using Taylor linearization, using Taylor linearization with regression weights, and using balanced repeated replication with various different Fay factors. Note that these Taylor estimators do not account for the randomness of the regression weights. The broad conclusions are that all the methods slightly underestimate the variances, with the standard Taylor method tending to underestimate by most. The only exception to this is standard balanced repeated replication, which somewhat overestimates the true variances. All the methods yield 95% confidence intervals with coverage in the range roughly 90–93%; this is a consequence of the variance underestimation.

In a second study, BOONSTRA and NIEUWENBROEK (2003) created an artificial population using data from the two-stage Dutch Labour Force Survey and used them to compare various forms of balanced repeated replication. The artificial population comprised 188,216 persons in 70 strata each containing from 10 to several hundred primary sampling units. Each such unit contained from 15 to 40 households. Five hundred samples were drawn according to a two-stage design with simple random sampling without replacement at both stages, with sampling fractions of 1/5 at each stage leading to overall samples of 4% of the population. The variance estimators used were the Taylor estimators corresponding to those mentioned in the preceding paragraph, grouped balanced repeated replication with 72 resamples, and grouped balanced repeated replication with artificial strata for a total of 120 strata and 120 resamples. All the balanced repeated replication estimates used Fay

Table 7.1: Typical performances of standard and balanced repeated replication variance estimators, from Boonstra and Nieuwenbroek (2003). RB is the relative bias and RMSE the root mean squared error of the 500 estimates. BRR denotes balanced repeated replication.

Statistic	Variance estimation method	RB (%)	Relative RMSE (%)
Total	standard	2.2	15.4
	grouped BRR	–0.1	29.2
	grouped BRR, with artificial strata	2.1	22.5
Ratio	standard	0.1	4.8
	standard, with regression weights	–0.4	4.8
	grouped BRR	–0.9	25.0
	grouped BRR, with artificial strata	0.6	13.7

factor $\varepsilon = 0.57$. Table 7.1, which contains some typical results from this study, shows that the grouped balanced repeated replication is appreciably less efficient than is the standard Taylor linearization approach, which is little affected by the use of the calibrated weights. Use of additional artificial strata improves balanced repeated replication slightly, but it remains considerably less efficient than the standard approach.

British and Swiss studies

Two major simulation exercises undertaken to compare the performances of resampling and more standard methods for variance estimation in labour force surveys have used simulated data based on real Swiss and British surveys (Canty and Davison, 1999). These involved overlapping waves of participants, and the statistics considered were totals, ratios, and differences of totals and ratios between two successive waves. Calibration by iterative proportional fitting was used to adjust weights to margins of three categorical variables, terminating after five iterations. The effect of missing data was not considered. Here we summarise results from the UK study, for which the population consisted of about $60,000$ addresses in the first wave, with around new $12,000$ addresses in the second wave used to replace $12,000$ of the first wave addresses; thus around 20% units change between the two waves. There were 30 strata used, each containing some individuals present only in the first wave, some present only in the second wave, and some present in both waves, in proportions $1:1:3$. The sampling fraction was roughly $f = 1/48$, yielding a sample size of 1250 in each sampled wave. A total of 500 samples were taken from the artificial population according to this sampling scheme, and different variance estimates were computed for a variety of smooth statistics, based on each of these samples. As the conclusions were similar for all of the statistics considered, the discussion below describes only variance estimation for totals and for difference of totals between the two waves. The estimator of the total used had form

$$\widehat{\tau} = \sum_{h,j,k} w_{hjk} y_{hjk},$$

where the sum is over strata h, addresses j, and admissible persons at that address k.

The variance estimation methods used were:

- a standard method based on a formula obtained by treating $\widehat{\tau}$ as a binomial variable with denominator $n = \sum_h n_h$ the total number of sampled addresses, that is (CANTY and DAVISON, 1999, equation (1))

$$d^2 \frac{\widehat{\tau}(1-\widehat{\tau})}{n} \left(\sum_{h,j,k} w_{hjk} \right),$$

  in which d is a design effect;

- jackknife linearization, in which (3.3) is used with

$$l_{hj} = n_h \sum_k w_{hjk} e_{hjk} - \sum_i \sum_k w_{hik} e_{hik},$$

  where $e_{hjk} = y_{hjk} - \widehat{y}_{hjk}$ is the residual from a regression of the variable y on the columns of the matrix X whose rows contain the calibration variables x_{hjk} using diagonal weight matrix Ω comprising the original weights ω_{hjk} for inclusion of individual k of address j from stratum h, and

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

  are the recalibrated weights;

- the grouped jackknife, with 10 groups in each of the 30 strata, so each variance recalculation involved 300 computations of the statistic;

- grouped balanced repeated replication, with the 30 strata subdivided into groups according to the presence of the unit in the first wave, the second wave, or both waves, giving a total of 90 replicates; and

- the standard bootstrap, appropriate in view of the low sampling fraction, with 100 replicates.

The effect of reweighting was investigated for the jackknife, for balanced repeated replication, and for the bootstrap. Three sets of weights were used: those for the original sample, which are 'incorrect' because they do not account for changes due to resampling, and those obtained after one and after five iterations of the iterative proportional fitting algorithm used for calibration.

Table 7.2 shows typical results for this study. All the resampling methods account for the extra variability introduced by calibration, but the standard method does not, so it tends to underestimate the true variance, though its variance is small. The bootstrap and jackknife linearization have smaller mean squared error than the jackknife and balanced repeated replication. Figure 7.1 shows a graphical summary of corresponding results for the estimation of unemployment rate, a ratio, for which both numerator and denominator are random, explaining the difference in scale between this figure and Table 7.2. The

relatively poor performance of the jackknife and of balanced repeated replication is due to a certain instability of these variance estimators, which sometimes substantially overestimate the true variance, though there seems to be little systematic bias. Although it is of no direct conern here, the perhaps surprisingly good performance of the standard method results from its low variability; it consistently underestimates the true variance, and always by about the same amount.

Table 7.2: Summary statistics for standard errors of the total unemployed and the change in total unemployed (Canty and Davison, 1999). 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	standard	–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	standard	–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

The main advantage of resampling methods is to account for the additional variability introduced by the calibration, but the corresponding disadvantage is that the calibration must be performed for each resampling estimate. This may give a heavy additional computational burden, and when calibration involves iteration it is natural to ask how many iterative steps are needed. This study compared the effects of using no steps, one step, and five steps of the iterative proportional fitting algorithm, when calibrating the resampled datasets. Figure 7.2 shows the results. For the bootstrap and balanced repeated replication, it seems that although there is a difference between using no steps and one step, there is little change afterwards: one iterative step is sufficient. The jackknife shows the rather strange feature that five iterative steps leads to highly variable estimators. Table 7.3 contains the corresponding numbers; the bootstrap has an attractive combination of low overall mean squared error and relative stability to the number of iterations of the calibration procedure.

7.3 Imputation

Kovar *et al.* (1988) performed a simulation study intended to test the balanced repeated replication method adjusted to the imputation mechanism. They considered stratified cluster sampling with $n_h = 2$ for $H = 32$ strata, and missingness rates of 10%, 20%, 30%, 40%, and 50%. For the sample mean and median statistics, they observed that

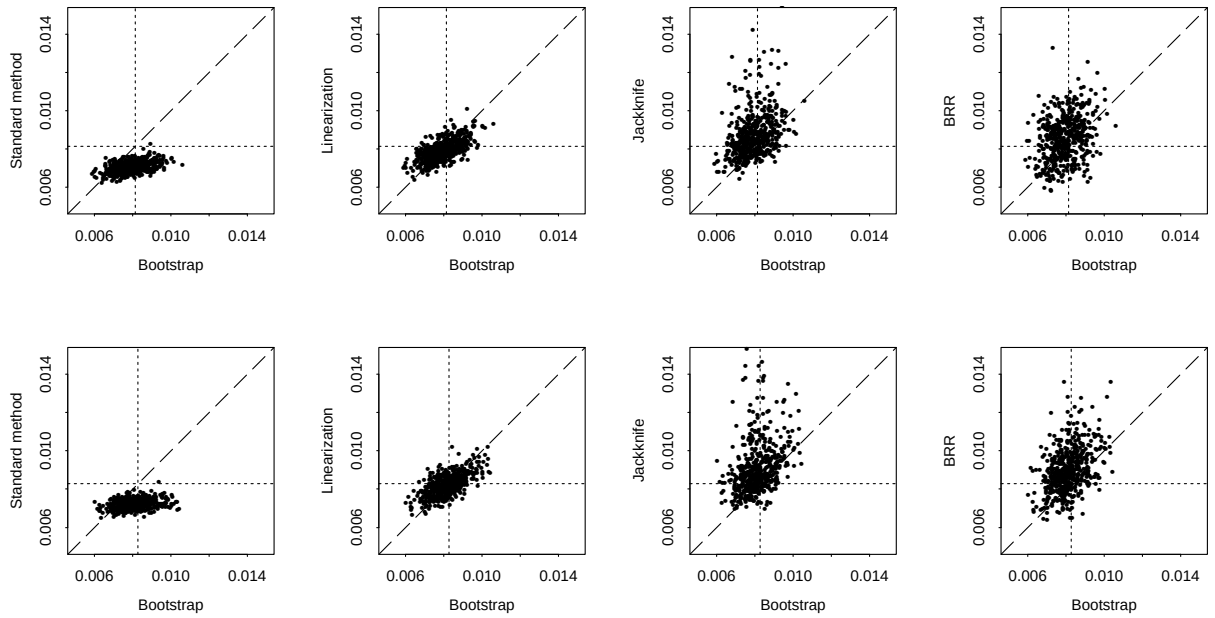

Figure 7.1: Standard errors for the unemployment rate (top row) and the change in unemployment rate (bottom row); the dashed line is $y = x$, the dotted lines are the 'true' sampling standard errors. From CANTY and DAVISON (1999).

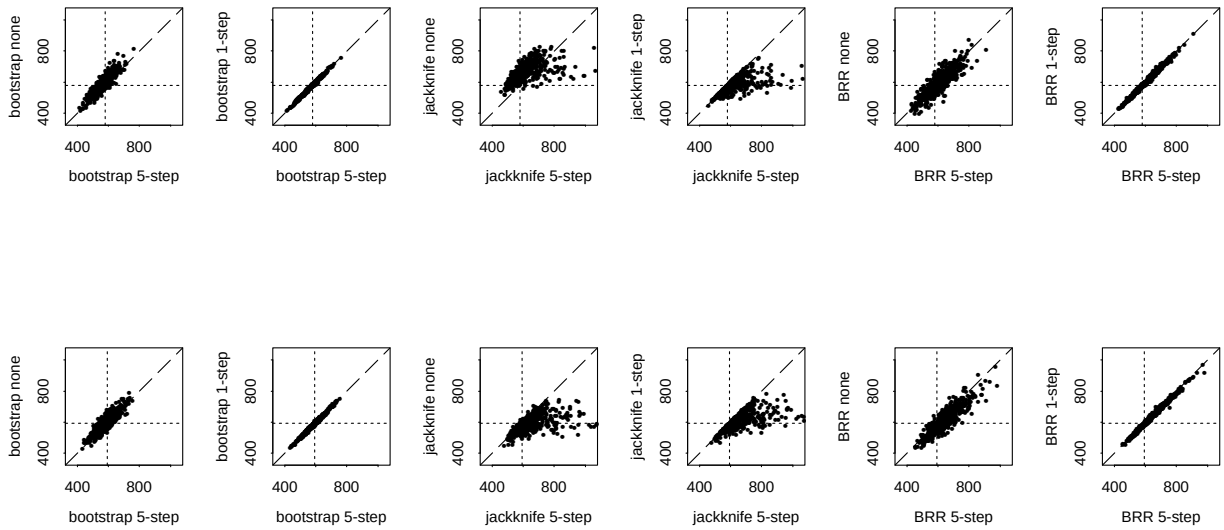

Figure 7.2: Standard errors for the total unemployed in quarter 1 (top row) and the change in total unemployment (bottom row) with different levels of reweighting; the dashes show the line $y = x$ and the dotted lines are the 'true' target sampling standard deviations. From CANTY and DAVISON (1999).

Table 7.3: Summary statistics for resampling standard errors of the total unemployment and its change with different levels of reweighting based on 500 samples (Canty and Davison, 1999). 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

the adjusted balanced repeated replication method works well in terms of relative bias and coefficient of variation, as opposed to the unadjusted one which underestimates the variance of the statistics.

7.4 Imputation and calibration

The simulations discussed above do not cover the case of calibrated and imputed data. Using a realistic simulation based on the Swiss Household Budget Survey from the year 1998 Renfer (2001), we consider the calibrated and imputed Horvitz–Thompson estimator of the total expenditure on bread and cereal products, based on complete data from $N = 9275$ households in $H = 7$ strata of various sizes. Also available on each household is a set of 14 auxiliary variables, of which 10 population margins are known. For the simulation, we consider the $N = 9275$ households as the whole population, for which we know the total expenditure. We then perform stratified random sampling without replacement and with equal inclusion probabilities 1/8 within 6 strata, and 3/8 in the other stratum, giving a sample size of 1332. Item non-response for the response variable is applied using a uniform probability of missingness across the entire sample. On each of the 500 samples simulated, we then calculate the calibrated and imputed Horvitz–Thompson estimates,

and apply various estimation techniques to obtain variances for them.

To match the computational complexity of bootstrap, for which a separate simulation showed that 100 replicates were adequate, we use roughly the same number of block deletions when applying the block jackknife with replacement. This was applied with 13 randomly-selected blocks in each stratum, leading to about 91 computations in all for each jackknife variance estimate.

Two forms of balanced repeated replication were applied, the first using a single random split of each stratum into two halves for each replication; no Fay factor was used but the weights for those observations included in the replicate were multiplied by a factor of two before calibration. The second form, repeatedly-grouped balanced repeated replication, averages over variance estimates from 13 such splits.

The bootstrap used 100 replicates of the calibrated and imputed Horwitz–Thompson estimator, obtained by the procedure of SHAO and SITTER (1996), that is, with missing responses imputed deterministically using a linear model fitted to the bootstrapped full respondents, and with the imputed dataset calibrated to the weights by linear regression.

The jackknife linearization estimator is that given by (3.3) and (3.4).

The standard formulae for multiple imputation were applied, using 30 random imputations from a linear model fitted to the complete data; for parametric imputation we used a homoscedastic normal error model, with the values of the regression parameters and variance changing randomly and independently according to the fitted normal and chi-squared distributions between simulations (RÄSSLER, 2004, Section 3.2); while for nonparametric imputation errors were simulated according to a model-based residual bootstrap (DAVISON and HINKLEY, 1997, p. 262).

Figures 7.3 and 7.4 compare the performances of these variance estimation techniques for missingness rates of 0%, 20%, 40%, and 60%. The block jackknife underestimates the true variances, which are systematically overestimated by the repeatedly-grouped balanced repeated replication. Balanced repeated replication without repeated grouping is clearly highly variable by comparison, in agreement with results of RAO and SHAO (1996), but grouped balanced repeated replication works much better, though it overestimates the variance when there is a high level of missingness. Jackknife linearisation works well for low levels of missingness, and overall it produces variances that are rather too low but quite stable. The bootstrap performs well in terms of both bias and variance, and seems to be the best of the methods. The multiple imputation methods make errors in opposite directions, when substantial fractions of the data are missing. Overall the bootstrap approach of SHAO and SITTER (1996) and the linearization method of Section 3.2 seem best in terms of bias and stability. As far as computation time is concerned, the advantage goes to linearization, which is up to fifty times faster than the other methods included in the study.

A separate simulation was conducted to assess the effect of using artificial strata for balanced repeated replication. Each of the $H = 7$ strata was divided into four parts, and the variance estimators were computed for 1000 samples with full response. Figure 7.5 shows the results. Introduction of the artificial strata greatly improves the balanced repeated replication variance estimators.

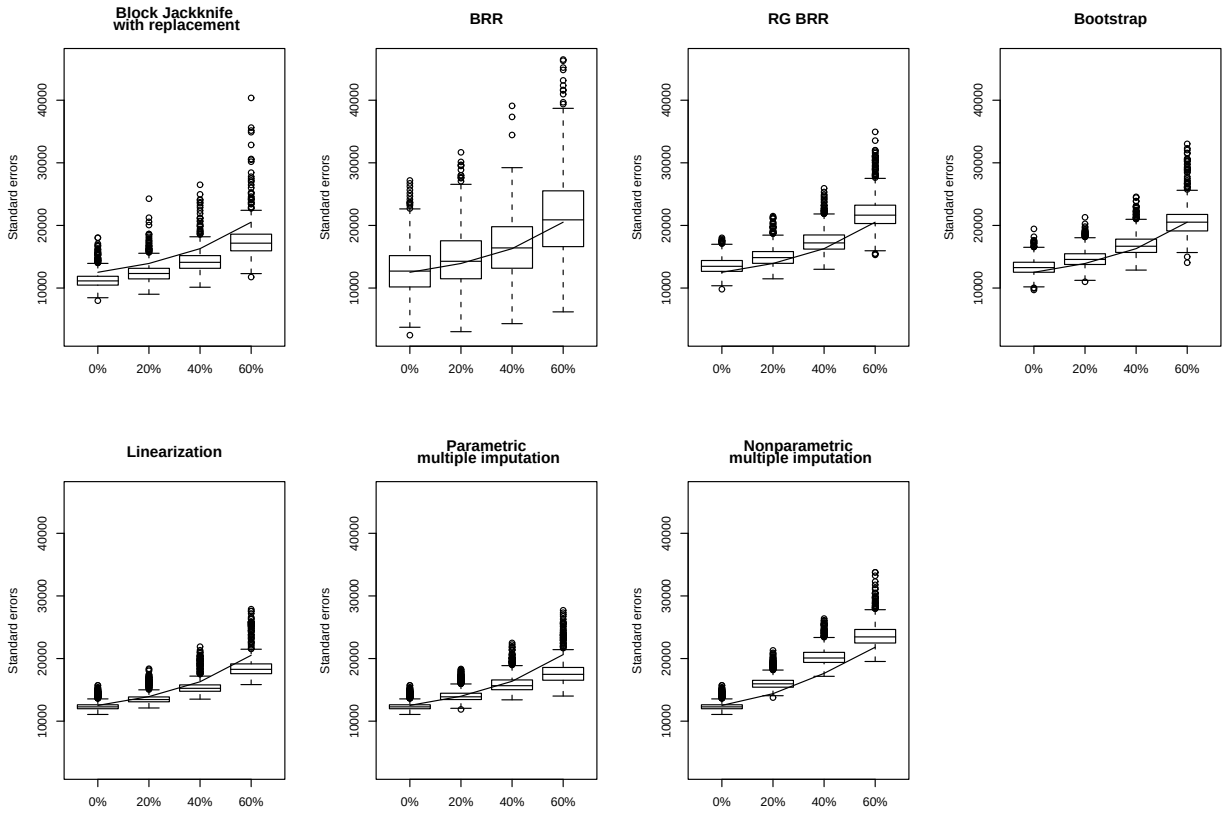

Figure 7.3: Comparison of resampling estimators of variance in the presence of calibration and imputation, as a function of the proportion of missing data. Simulation based on the 1998 Swiss Household Budget Survey. The solid line shows the true variances, estimated from 10,000 simulations, and the boxplots show the variance estimates computed for 500 samples.

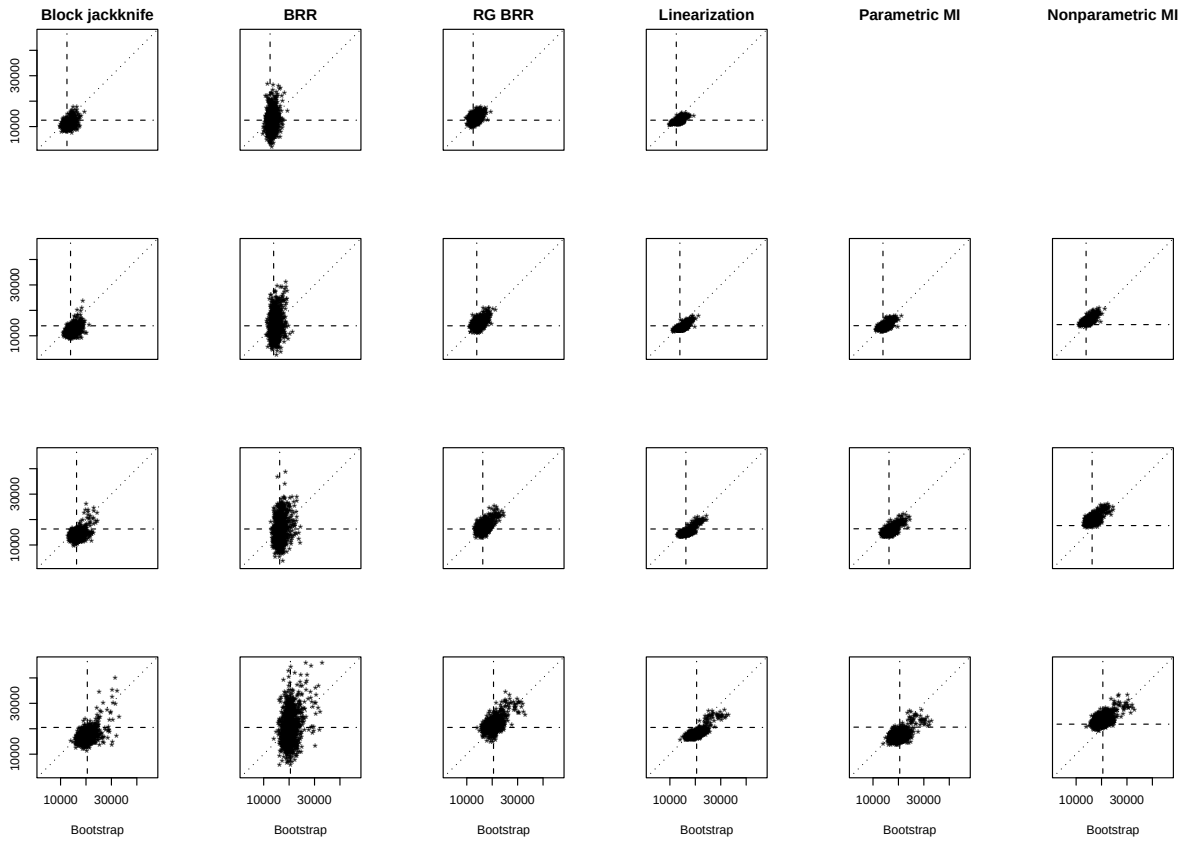

Figure 7.4: Comparison of resampling standard errors in the presence of calibration and imputation, as a function of the proportion of missing data; from top to bottom 0%, 20%, 40%, 60% item non-response. The dashed lines are the 'true' sampling standard errors, and the dotted line shows $x = y$. Simulation based on the 1998 Swiss Household Budget Survey.

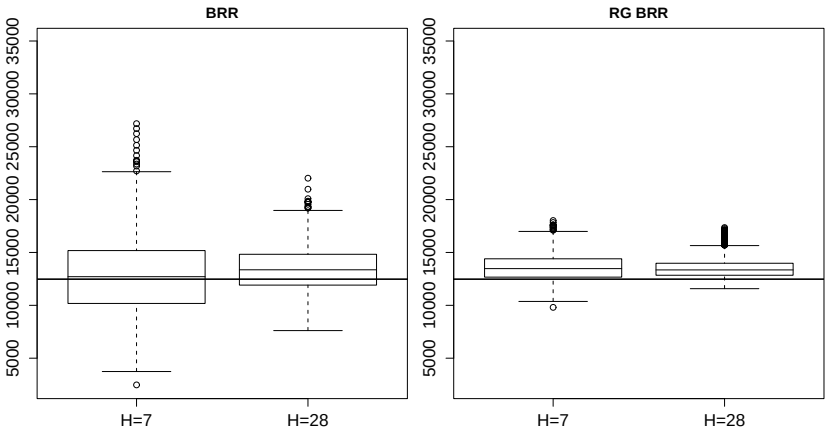

Figure 7.5: Effect of artificial strata on balanced repeated replication variance estimators; those in the right panel use repeated grouping. Simulation of 1000 variance estimators based on the 1998 Swiss Household Budget Survey. In each panel the boxplots show the estimators based on the 7 original strata (left) and when each of these is divided into 4 artificial sub-strata (right).

Chapter 8

Discussion

Overall it seems that bootstrap and linearization variance estimators are most promising for use with complex surveys. The bootstrap has the advantage of being a general-purpose tool which can be applied without much tuning in many situations, and which can be used for both smooth and non-smooth statistics. Moreover and unlike the jackknife or balanced repeated replication, the number of recomputations needed is to a large extent controlled by the user rather than determined by the method. Its main disadvantages are the computational burden that it entails, particularly when used with imputation, and the fact that special programming is needed if it is applied to situations with large sampling fraction f_h. Jackknife linearization demands special computation of influence function values adapted for particular circumstances, but it involves no resampling and so is much quicker than the bootstrap. It cannot safely be applied for non-smooth statistics, but the chain rule can be used to obtain influence values for complex estimators. Balanced repeated replication and the jackknife are almost competitive in some cases, but overall they perform worse than the other methods, and tuning seems to be needed to get the best performance from them.

Jackknife linearization and the bootstrap can be applied for estimators of change

$$\widehat{\theta} = t(\widehat{F}_2, \widehat{F}_3) - t(\widehat{F}_1, \widehat{F}_2),$$

based on panel surveys by splitting the units into three parts: those present only at the first occasion of sampling, those present on second occasion only, and those present on both sampling occasions (Canty and Davison, 1999); these are represented respectively by $\widehat{F}_1$, $\widehat{F}_3$ and $\widehat{F}_2$. Each of the original H strata is split into three parts and the estimator computed accordingly. If we ignore corrections for sampling fractions, the jackknife linearization variance estimator is given by

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

where the h-th stratum contains n_{h1} units present on the first occasion, n_{h2} households present on both occasions, and n_{h3} households present on the second occasion only, and the corresponding empirical influence values are given an obvious notation. A similar argument applies to the bootstrap, for which resampling is applied with these same substrata, and to the other resampling methods.

Acknowledgements

This report was greatly improved by helpful comments from Siegfried Gabler and Chris Skinner.

References

Berger, Y. G. and Skinner, C. J. (2003): Variance estimation for a low income proportion. *Applied Statistics* **52**, 457–468.

Berger, Y. G. and Skinner, C. J. (2004): A jackknife variance estimator for unequal probability sampling. To appear in the Journal of the Royal Statistical Society, Series B.

Bickel, P. J. and Freedman, D. A. (1984): Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics* **12**, 470–482.

Binder, D. A. (1996): Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology* **22**, 17–22.

Boonstra, H. J. H. and Nieuwenbroek, N. (2003): *An empirical comparison of BRR and linearization variance estimators.* DACSEIS.

Booth, J. G., Butler, R. W. and Hall, P. (1994): Bootstrap methods for finite populations. *Journal of the American Statistical Association* **89**, 1282–1289.

Booth, J. G. and Hall, P. (1994): Monte Carlo approximation and the iterated bootstrap. *Biometrika* **81**, 331–340.

Campbell, C. (1980): A different view of finite population estimation. In *Proceedings of the Section on Survey Research Methods*, pp. 319–324. Alexandria, Virginia: American Statistical Association.

Canty, A. J. and Davison, A. C. (1999): Resampling-based variance estimation for labour force surveys. *The Statistician* **48**, 379–391.

Chao, M. T. and Lo, S. H. (1985): A bootstrap method for finite populations. *Sankyā A* **47**, 399–405.

Chen, J. and Shao, J. (2001): Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association* **96**, 260–269.

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

D'Arrigo, J. and Skinner, C. J. (2003): *Variance estimation for estimators subject to raking adjustment: Deliverables 8.1 and 8.2.* DACSEIS.

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

Deville, J. C. (1999): Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* **25**, 193–203.

Deville, J. C. and Särndal, C. E. (1992): Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Deville, J. C., Särndal, C. E. and Sautory, O. (1993): Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013–1020.

Dippo, C. S., Fay, R. E. and Morganstein, D. H. (1984): Computing variances from complex samples with replicate weights. In *Proceedings of the Section on Survey Research Methods*, pp. 489–494. Washington DC: American Statistical Association.

Efron, B. (1979): Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. (1982): *The Jackknife, the bootstrap, and other resampling plans.* Philadephia: SIAM.

Fay, R. E. (1989): Theory and application of replicate weighting for variance calculations. In *Proceedings of the Social Statistics Section*, pp. 212–217. American Statistical Association.

Fay, R. E. (1996): Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* **91**, 490–498.

Gross, S. (1980): Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, pp. 181–184. Alexandria, VA: American Statistical Association.

Gupta, V. K. and Nigam, A. K. (1987): Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika* **74**, 735–742.

Gurney, M. and Jewett, R. S. (1975): Constructing orthogonal replications for standard errors. *Journal of the American Statistical Association* **70**, 819–821.

Hall, P. G. (2003): A short prehistory of the bootstrap. *Statistical Science* **18**, 158–167.

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.

Hartigan, J. A. (1969): Using subsample values as typical values. *Journal of the American Statistical Association* **64**, 1303–1317.

Huber, P. J. (1981): *Robust Statistics.* New York: Wiley.

Judkins, D. R. (1990): Fay's method of variance estimation. *Journal of Official Statistics* **6**, 223–239.

Kim, J. K. and Sitter, R. R. (2003): Efficient replication variance estimation for two-phase sampling. *Statistica Sinica* **13**, 641–653.

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* **16**, 25–45.

Krewski, D. and Rao, J. N. K. (1981): Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* **9**, 1010–1019.

Lahiri, P. (2003): On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* **18**, 199–210.

McCarthy, P. J. (1969): Pseudo-replication: Half samples. *Review of the Interational Statistics Institute* **37**, 239–264.

McCarthy, P. J. and Snowden, C. B. (1985): The bootstrap and finite population sampling. *Vital and Health Statistics* **2**, 2–95.

Miller, R. G. (1974): The jackknife - A review. *Biometrika* **61**, 1–15.

Nielsen, S. F. (2003): Proper and improper multiple imputation (with discussion). *International Statistical Review* **71**, 593–627.

Politis, D. N., Romano, J. P. and Wolf, M. (1999): *Subsampling*. New York: Springer-Verlag.

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

Quenouille, M. H. (1949a): Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B* **11**, 68–84.

Quenouille, M. H. (1949b): Notes on bias in estimation. *Biometrika* **43**, 353–360.

Rao, J. N. K. (1988): Variance estimation in sample surveys. In *Handbook of Statistics, Volume 6*, ed. P. R. K. amd C. R. Rao, pp. 427–447. Amsterdam: Elsevier Science.

Rao, J. N. K. and Shao, J. (1992): Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811–822.

Rao, J. N. K. and Shao, J. (1996): On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association* **91**, 343–348.

Rao, J. N. K. and Shao, J. (1999): Modified balanced repeated replication for complex survey data. *Biometrika* **86**, 403–415.

Rao, J. N. K. and Wu, C. F. J. (1985): Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association* **80**, 620–630.

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.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992): Some recent work on resampling methods for complex surveys. *Survey Methodology* **18**, 209–217.

Rässler, S. (2004): *The Impact of Multiple Imputation for DACSEIS*. DACSEIS Research Paper number 5.

Renfer, J.-P. (2001): *Description and process of the Household and Budget Survey of 1998 (HBS 1998)*. Swiss Federal Statistical Office. 1-19.

Rubin, D. B. (1987): *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1996): Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489. with discussion, 507-515, and rejoinder, 515-517.

Rust, K. F. and Rao, J. N. K. (1996): Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* **5**, 283–310.

Shao, J. (2003): Impact of the bootstrap on sample surveys. *Statistical Science* **18**, 191–198.

Shao, J., Chen, Y. and Chen, Y. (1998): Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association* **93**, 819–831.

Shao, J. and Rao, J. N. K. (1994): Standard errors for low income proportions estimated from stratified multistage samples. *Sankyā B* **55**, 393–414.

Shao, J. and Sitter, R. R. (1996): Bootstrap for imputed survey data. *Journal of the American Statistical Association* **91**, 1278–1288.

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

Shao, J. and Wu, C. F. J. (1989): A general theory for jackknife variance estimation. *Annals of Statistics* **17**, 1176–1197.

Shao, J. and Wu, C. F. J. (1992): Asymptotic properties of the balanced repeated replication method for sample quantiles. *Annals of Statistics* **20**, 1571–1593.

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.

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

Tukey, J. W. (1958): Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics* **29**, 614.

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

Wu, C. F. J. (1991): Balanced repeated replications based on mixed orthogonal arrays. *Biometrika* **78**, 181–188.

Zhang, P. (2003): Multiple imputation: Theory and method. *International Statistical Review* **71**, 581–592.