Simple design-efficient calibration estimators for rejective and high-entropy sampling

BY Z. TAN

Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, U.S.A.
ztan@stat.rutgers.edu

SUMMARY

For survey calibration, consider the situation where the population totals of auxiliary variables are known or where auxiliary variables are measured for all population units. For each situation, we develop design-efficient calibration estimators under rejective or high-entropy sampling. A general approach is to extend efficient estimators for missing-data problems with independent and identically distributed data to the survey setting. We show that this approach effectively resolves two long-standing issues in existing approaches: how to achieve design-efficiency regardless of a linear super-population model in generalized regression and calibration estimation and how to find a simple approximation in optimal regression estimation. Moreover, the proposed approach sheds light on several issues which seem not to be well studied in the literature. Examples include use of the weighted Kullback–Liebler distance in calibration estimation, and efficient estimation allowing for misspecification of a nonlinear super-population model.

Some key words: Auxiliary information; Calibration; Empirical likelihood; Missing data; Nonparametric likelihood; Poisson sampling; Rejective sampling; Regression estimator.

I. INTRODUCTION

An important topic for survey sampling is calibration with auxiliary information to improve estimation (e.g., Deville & Sarndal, 1992). Suppose that a probability sample, $s$, is drawn from a finite population of $N$ units, $U = \{1, \ldots, i, \ldots, N\}$. For $i \in U$, let $Y_i$ be the value of a study variable, $X_i$ be that of a vector of auxiliary variables, and $R_i$ be the selection indicator such that $R_i = 1$ if $i \in s$ or $R_i = 0$ otherwise. The value $Y_i$ is measured only for $i \in s$, whereas $X_i$ is measured for $i \in s$ along with auxiliary information: either $X_i$ is known for all $i \in U$ or $\sum_{i \in U} X_i$ is known. For expository convenience, we assume the first situation, called complete auxiliary information (Wu & Sitter, 2001), and consider the second situation as a restricted case such that although $\sum_{i \in U} h(X_i)$ is known for any function $h(x)$, only $\sum_{i \in U} X_i$ is exploited. Then the observed data are $\{(R_i Y_i, X_i, R_i) : i \in U\}$. The objective is to estimate the population mean $\bar{Y} = N^{-1} \sum_{i \in U} Y_i$.

We develop design-efficient calibration estimators for rejective and high-entropy sampling. Here design-efficiency is with respect to the sampling design, regardless of any super-population model. A basic idea is to link survey sampling to missing-data problems in the non-survey setting with independent and identically distributed data (e.g., Robins et al., 1994). In fact, survey calibration described above is immediately suggestive of a missing-data problem as follows (e.g., Tan, 2006, 2010a). Suppose that $\{(Y_i, X_i, R_i) : i = 1, \ldots, N\}$ are independent and identically distributed realizations of $(Y, X, R)$, where $Y$ is a response variable, $X$ is a vector of explanatory variables, and $R$ is the non-missing indicator such that $R = 1$ or 0 if $Y$ is observed or
missing. Then the observed data are \( \{(R_i Y_i, X_i, R_i) : i = 1, \ldots, N\} \), of exactly the same form as in survey sampling. Further, assume that the missing-data mechanism is ignorable (Rubin, 1976), \( \Pr(R = 1 \mid Y, X) = \Pr(R = 1 \mid X) \), and that the propensity score (Rosenbaum & Rubin, 1983), \( p(x) = \Pr(R = 1 \mid X = x) \), is a known function. The objective is typically to estimate the expectation \( \mu = E(Y) \), instead of \( \bar{Y} \).

The preceding discussion shows that survey calibration and the missing-data problem are conceptually similar to each other. In general, there remain important, analytical differences between the two settings in whether \( (Y_1, X_1), \ldots, (Y_N, X_N) \) are fixed or random and whether \( (R_1, \ldots, R_N) \) are independent of each other. Nevertheless, the probabilistic structure of the foregoing missing-data problem is equivalent to that of Poisson sampling with a nonparametric super-population model, where \( (Y_1, X_1), \ldots, (Y_N, X_N) \) are considered an independent and identically distributed sample from a super-population and \( (R_1, \ldots, R_N) \) are assumed independent given \( (Y_1, X_1), \ldots, (Y_N, X_N) \). Then Poisson sampling is essentially equivalent to the missing-data problem with \( (Y_1, X_1), \ldots, (Y_N, X_N) \) fixed or conditioned on. A drawback of Poisson sampling is that the sample size, \( \sum_{i=1}^{N} R_i \), is random. Therefore, it is important to consider rejective sampling (Hajek, 1964, 1981), which is Poisson sampling restricted to or conditioned on a fixed sample size.

From these considerations, we propose the following approach:

(i) to employ rejective sampling or high-entropy sampling which is asymptotically equivalent to rejective sampling, and
(ii) to develop efficient estimation, by constructing efficient estimators for the missing-data problem and then extending those estimators to survey sampling.

In this approach, the estimators are transferred from the missing-data problem to the survey setting and then studied fully under the latter probabilistic structure.

Rejective sampling is suggested above for its close connection to the missing-data problem. Moreover, there are several reasons why rejective sampling is desirable. First, fast algorithms have recently been proposed for implementing rejective sampling (Chen et al., 1994; Tillé, 2006). Second, rejective sampling achieves maximum entropy subject to fixed inclusion probabilities for a fixed sample size (Hajek, 1981). High entropy means that all samples have probabilities as uniform as allowed by the constraints, where low entropy indicates that a few samples have disproportionately large probabilities. Third, the structure of rejective sampling as conditional Poisson sampling facilitates the development of simple, asymptotic approximations (Hajek, 1964, 1981). Finally, asymptotic results can be generalized from rejective sampling to other high-entropy sampling designs, including Rao–Sampford sampling, such that the Kullback–Liebler divergence from rejective sampling tends to zero (Berger, 1988a,b, 2005).

Consider linear calibration where only the population totals of auxiliary variables \( X \) are exploited. We adopt the calibrated regression and likelihood estimators of Tan (2006, 2010a) from the missing-data problem. The resulting estimators take the simple form of generalized regression and calibration estimators (Särndal et al., 1992; Deville & Särndal, 1992), but are asymptotically design-efficient, regardless of a linear, super-population model of \( Y_i \) given \( X_i \), similarly to an optimal regression estimator (Fuller & Isaki, 1981; Montanari, 1987; Rao, 1994). Therefore, our approach effectively resolves two long-standing issues in existing approaches. First, generalized regression and calibration estimation gives a class of simple, consistent estimators, depending on first-order and not second-order inclusion probabilities, but it is unclear how to achieve design-efficiency. Second, the optimal regression estimator is asymptotically design-efficient, but cumbersome as it depends on first-order and second-order inclusion probabilities for a general sampling scheme including rejective sampling.
Calibration estimators for rejective sampling

In the presence of complete auxiliary information, the model-calibration approach of Wu & Sitter (2001) consists of two steps, similarly to locally efficient estimation in the missing-data problem (e.g., Robins et al., 1994): a possibly nonlinear regression model is fitted for $Y_i$ given $X_i$, and then linear calibration is performed using the population total of the fitted values. The resulting estimator of $\tilde{Y}$ is efficient in the sense of minimizing the anticipated asymptotic variance if the regression model is correctly specified (Wu, 2003). However, it is possible to improve efficiency when the regression model may be misspecified. For the second step, linear calibration can be done by applying the design-efficient estimators discussed above. For the first step, the regression model can be estimated to enhance efficiency of the final estimator of $\tilde{Y}$, similarly as in the method of efficiency maximization in the missing-data literature (Rubin & van der Laan, 2008; Tan, 2008; Cao et al., 2009).

Our calibrated estimators, though constructed under rejective sampling, can be adopted and evaluated under a general sampling scheme different from rejective sampling. Under suitable regularity conditions, the estimators remain consistent and asymptotically normal, but may not be design-efficient, similarly to existing generalized regression and calibration estimators (Sarndal et al., 1992; Deville & Sarndal, 1992).

2. Existing methods

There is a vast literature on calibration methods for incorporating auxiliary information and for handling non-response in survey sampling (e.g., Fuller, 2002; Kott, 2009). We only discuss methods for linear calibration directly related to our approach in §3. Throughout, $E_r$, $\text{var}_r$, and $\text{cov}_r$ denote the expectation, variance, and covariance under the sampling design, with $r$ indicating the sampling randomization. The first-order inclusion probabilities are $\pi_i = \text{pr}(R_i = 1)$.

Cassel et al. (1976) and Sarndal et al. (1992) developed a class of generalized regression estimators of $\tilde{Y}$:

$$\hat{\mu}_{\text{GREG}} = \frac{1}{N} \sum_{i \in S} Y_i \pi_i^{-1} - \frac{1}{N} \left( \sum_{i \in S} X_i^T \pi_i^{-1} - \sum_{i \in U} X_i^T \right) \hat{\beta}_q,$$

where $\hat{\beta}_q = (\sum_{i \in S} \pi_i^{-1} q_i X_i X_i^T)^{-1}(\sum_{i \in S} \pi_i^{-1} q_i X_i Y_i)$ and $q_i > 0$ are weights to be specified. In general, different choices of $q_i$ result in different variances. The standard choice is $q_i = 1$ for $i \in s$. The choice advocated in Sarndal et al. (1992) is $q_i = \sigma_i^{-2}$, motivated by a linear model

$$E_r(Y_i | X_i) = X_i^T \beta, \quad \text{var}_r(Y_i | X_i) = \sigma_i^2. \quad (1)$$

Throughout, $E_r$ and $\text{var}_r$ denote the expectation and variance under a super-population model. The choice $q_i = \sigma_i^{-2}$ might be considered optimal in minimizing the variance under model (1). But if model (1) holds, then the anticipated variance of $\hat{\mu}_{\text{GREG}}$, $E_r E_r \{ (\hat{\mu}_{\text{GREG}} - \tilde{Y})^2 \}$, is approximately $N^{-2} \sum_{i \in U} \{ (\pi_i^{-1} - 1) \sigma_i^2 \}$, regardless of the choice of $q_i$ (Wright, 1983). Therefore, any choice of $q_i$ is asymptotically as efficient as $q_i = \sigma_i^{-2}$ if model (1) holds. It is important to investigate how to choose $q_i$ for efficient estimation when model (1) may be misspecified.

Calibration estimation (Deville & Sarndal, 1992) gives an even broader class of estimators than $\hat{\mu}_{\text{GREG}}$. Let $d_i = \pi_i^{-1}$. A calibration estimator of $\tilde{Y}$ is $\hat{\mu}_{\text{calc}} = N^{-1} \sum_{i \in S} w_i Y_i$, where $w_i$’s are defined to minimize a distance, $\sum_{i \in S} G_i(w_i, d_i)$, subject to $\sum_{i \in S} w_i X_i = \sum_{i \in U} X_i$. Three examples of $G_i(w_i, d_i)$ are $(w_i - d_i)^2/(2d_iq_i)$, $(w_i \log(w_i/d_i) - w_i - d_i)/q_i$, and $\{-d_i \log(w_i/d_i) + w_i - d_i\}/q_i$, where $q_i > 0$ are weights to be specified. For convenience, we refer to the corresponding distances $\sum_{i \in S} G_i(w_i, d_i)$ as the weighted chi-squared, reverse Kullback–Liebler, and Kullback–Liebler distances. The weighted chi-squared distance
leads to weights of form \( w_i = d_i (1 + q_i X_i^T \lambda) \). Then \( \hat{\mu}_{\text{cal}} \) is algebraically identical to \( \hat{\mu}_{\text{aere}} \) using the same \( q_i \). The weighted reverse Kullback–Liebler distance leads to weights of form \( w_i = d_i \exp(q_i X_i^T \lambda) \). Then \( \hat{\mu}_{\text{cal}} \) is called a generalized raking estimator (Deville et al., 1993).

The weighted Kullback–Liebler distance leads to weights of form \( w_i = d_i / (1 + q_i X_i^T \lambda) \). With \( q_i = 1 \ (i \in s) \), if \( \sum_{i \in s} w_i = N \) is included as a calibration equation, then minimizing the Kullback–Liebler distance is equivalent to maximizing \( \sum_{i \in s} d_i \log(w_i) \) and hence \( \hat{\mu}_{\text{cal}} \) coincides with the pseudo empirical likelihood estimator (Chen & Sitter, 1999) discussed below. For the two types of Kullback–Liebler distances, but not the chi-squared distance, the weights \( w_i \) are always positive, which is desirable in finite samples.

Under suitable regularity conditions, calibration estimators \( \hat{\mu}_{\text{cal}} \) using different distance measures but the same \( q_i \) admit the same asymptotic expansion (Deville & Sarndal, 1992)

\[
\frac{1}{N} \sum_{i \in s} \frac{Y_i}{\pi_i} - \frac{1}{N} \left( \sum_{i \in s} \frac{X_i^T}{\pi_i} - \sum_{i \in U} X_i^T \right) \beta_q + o_p(n^{-1/2}),
\]

where \( \beta_q = \left( \sum_{i \in U} q_i X_i X_i^T \right)^{-1} \left( \sum_{i \in U} q_i X_i Y_i \right) \) and \( n \) is the sample size or the expected sample size. This result confirms the foregoing discussion on choices of \( q_i \) for \( \hat{\mu}_{\text{aere}} \). If model (1) holds, then \( \beta_q \) converges in probability under model (1) to \( \beta \) and hence the anticipated variance of (2) is asymptotically the same for different choices of \( q_i \).

Consider another class of estimators of \( \bar{\mu} \),

\[
\frac{1}{N} \sum_{i \in s} \frac{Y_i}{\pi_i} - \frac{1}{N} \left( \sum_{i \in s} \frac{X_i^T}{\pi_i} - \sum_{i \in U} X_i^T \right) b,
\]

where \( b \) is a vector of arbitrary constants. By expansion (2), this class contains all calibration estimators \( \hat{\mu}_{\text{cal}} \) up to asymptotic equivalence. The optimal choice of \( b \) in minimizing the variance of (3) is \( \beta_{\text{opt}} = \text{var}^{-1} \left( \sum_{i \in s} \pi_i^{-1} X_i \right) \text{cov} \left( \sum_{i \in s} \pi_i^{-1} X_i, \sum_{i \in s} \pi_i^{-1} Y_i \right) \). Estimating the variance and covariance matrices in \( \beta_{\text{opt}} \) and then substituting the estimator of \( \beta_{\text{opt}} \) for \( b \) in (3) yields an optimal regression estimator, which is asymptotically equivalent to (3) with \( b = \beta_{\text{opt}} \) (Fuller & Isaki, 1981; Montanari, 1987; Rao, 1994). Therefore, the optimal regression estimator asymptotically attains the minimum variance of estimators (3), regardless of model (1). On the other hand, this estimator is cumbersome with double summation, except in special cases such as Poisson sampling in §3.2 and stratified simple random sampling.

There has been considerable research on comparisons and connections between the generalized regression and calibration approach and the optimal regression approach (e.g., Montanari, 1998; Andersson & Thorburn, 2006). Our review is focused on the main features of the two approaches. The first provides simple, but not necessarily design-efficient estimators, whereas the second gives design-efficient but in general complicated estimators.

Finally, there are empirical likelihood methods for survey calibration, including those in Hartley & Rao (1968) and Chen & Qin (1993) for simple random sampling and Chen & Sitter (1999) and Kim (2009) for unequal probability sampling. As mentioned above, Chen & Sitter’s (1999) estimator of \( \bar{\mu} \) is \( \hat{\mu}_{\text{cal}} = N^{-1} \sum_{i \in s} w_i Y_i \), where \( w_i \)’s are defined to maximize the pseudo empirical likelihood \( \sum_{i \in s} d_i \log w_i \) subject to \( \sum_{i \in s} w_i = N \) and \( \sum_{i \in s} w_i X_i = \sum_{i \in U} X_i \). These methods do not, in general, produce efficient estimators except for simple random sampling or in the case of negligible sampling fractions. A heuristic reason for the lack of efficiency is that such pseudo likelihoods do not capture the probabilistic structure of sampling. In fact, as discussed by Rao (1997), design-based likelihood inference for survey data is challenging because the usual likelihood provides no information on \( \{ Y_i : i \notin s \} \) and hence on \( \bar{\mu} \).
3. LINEAR CALIBRATION

3.1. Missing-data problem

Consider the missing-data problem described in §1. To mimic survey sampling with known inclusion probabilities, the propensity score $p(x) = \Pr(R = 1 \mid X = x)$ is assumed to be a known function. In a real missing-data situation, the function $p(x)$ must be estimated from the data $\{(X_i, R_i) : i = 1, \ldots, N\}$. There are interesting issues related to estimated propensity scores and doubly robust estimation against misspecification of propensity score models, which are not addressed here. See, for example, Robins et al. (1994) and Tan (2006, 2007, 2010a).

For the missing-data problem, it is possible to develop efficient estimation using familiar statistical tools for independent and identically distributed data. Let $\eta = p^{-1}(X)RY$ and $\xi = \{p^{-1}(X)R - 1\}X$. Then $\mu = E(\eta)$ and $0 = E(\xi)$. By the method of control variates (Hammersley & Handscomb, 1964), a regression estimator of $\mu$ is

$$\hat{\mu}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i}{p(X_i)} - \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{R_i}{p(X_i)} - 1 \right\} X^T \hat{\beta},$$

where $\hat{\beta}$ is the least-square estimator in the linear regression of $\eta$ on $\xi$:

$$\hat{\beta} = \left[ \sum_{i=1}^{N} \left\{ \frac{R_i}{p(X_i)} - 1 \right\} X_i X_i^T \right]^{-1} \left[ \sum_{i=1}^{N} \frac{R_i}{p(X_i)} \left\{ \frac{R_i}{p(X_i)} - 1 \right\} X_i Y_i \right].$$

Alternatively, the method of empirical likelihood (Owen, 2001) can be used, by regarding $E(\xi) = 0$ as a constraint on the joint distribution of $(\eta, \xi)$. In the Supplementary Material, the empirical likelihood estimator of $\mu$ is shown to be

$$\hat{\mu}_{\text{lik}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\eta_i}{1 + \lambda \xi_i} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i}{\varpi(X_i; \lambda)},$$

where $\varpi(X; \lambda) = p(X) + \{1 - p(X)\}X^T \lambda$ and $\hat{\lambda}$ is a maximizer of

$$\ell(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \left[ R_i \log \varpi(X_i; \lambda) + (1 - R_i) \log \{1 - \varpi(X_i; \lambda)\} \right]$$

subject to $\varpi(X_i; \lambda) > 0$ if $R_i = 1$ and $\varpi(X_i; \lambda) < 1$ if $R_i = 0$. Setting the gradient of $\ell(\lambda)$ to 0 shows that $\hat{\lambda}$ is a solution to

$$0 = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{R_i - \varpi(X_i; \lambda)}{\varpi(X_i; \lambda) \{1 - \varpi(X_i; \lambda)\}} \{1 - p(X_i)\} X_i \right].$$

The same estimator $\hat{\mu}_{\text{lik}}$ is derived by the method of nonparametric likelihood in Tan (2006) provided that 1 is included in $X$. In this case, $\hat{\lambda}$ satisfies $1 = N^{-1} \sum_{i=1}^{N} \varpi^{-1}(X_i; \lambda) R_i$ in addition to (5), and $\hat{\mu}_{\text{lik}}$ and $\hat{\mu}_{\text{reg}}$ are invariant under a location transformation $(Y, \mu) \mapsto (Y + c, \mu + c)$. See Tan (2010b, 2011) for further discussions on the relationship between the methods of empirical and nonparametric likelihood in missing-data problems.

Neither $\hat{\mu}_{\text{reg}}$ nor $\hat{\mu}_{\text{lik}}$ is calibrated on $X$. Throughout, an estimator is said to be calibrated on $X$ if applying the estimator with $Y$ replaced by $X$ yields exactly $X = N^{-1} \sum_{i=1}^{N} X_i$. Remarkably, there are simple modifications of $\hat{\mu}_{\text{reg}}$ and $\hat{\mu}_{\text{lik}}$ to achieve calibration on $X$, without affecting
first-order asymptotic variances. The calibrated regression estimator of Tan (2006) is
\[
\hat{\mu}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i}{p(X_i)} - \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{R_i}{p(X_i)} - 1 \right\} X_i^T \hat{\beta},
\]
where \( \hat{\beta} \) is a coefficient vector different from \( \hat{\beta} \):
\[
\hat{\beta} = \left[ \sum_{i=1}^{N} \frac{R_i}{p(X_i)} \left\{ \frac{1}{p(X_i)} - 1 \right\} X_i X_i^T \right]^{-1} \left[ \sum_{i=1}^{N} \frac{R_i}{p(X_i)} \left\{ \frac{1}{p(X_i)} - 1 \right\} X_i Y_i \right].
\]

See Rubin & van der Laan (2008) and Tan (2008) for related discussions. The calibrated likelihood estimator of Tan (2010a) is
\[
\hat{\mu}_{\text{lik}} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i}{\varpi(X_i; \lambda)},
\]
where \( \varpi(X; \lambda) = p(X) + \{1 - p(X)\} X^T \lambda \) as before and \( \hat{\lambda} \) is a solution to
\[
0 = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{R_i}{\varpi(X_i; \lambda)} - 1 \right\} X_i,
\]
subject to \( \varpi(X_i; \lambda) > 0 \) if \( R_i = 1 \) for \( i = 1, \ldots, N \). Equation (6) is obtained by setting \( \{1 - p(X_i)\}/\{1 - \varpi(X_i; \lambda)\} \) to 1 in (5). If 1 is included in \( X \), then \( 1 = N^{-1} \sum_{i=1}^{N} \varpi^{-1}(X_i; \lambda) R_i \), and \( \hat{\mu}_{\text{lik}} \) and \( \hat{\mu}_{\text{reg}} \) are location invariant. It is interesting to mention that calibration on \( X \) implies double robustness such that an estimator of \( \mu \) remains consistent if either the propensity score \( p(x) \) or the linear model \( E(Y \mid X = x) = x^T \beta \) is correctly specified. In fact, both estimators \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \) were originally constructed to achieve this property in Tan (2006, 2010a).

There are further differences between the calibrated estimators \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \) and the non-calibrated estimators \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \) in connection with survey sampling. The estimators \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \), unlike \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \), turn out to depend on \( (X_1, \ldots, X_N) \) only through \( \{X_i : R_i = 1, i = 1, \ldots, N\} \) and \( \hat{\lambda} \), and therefore are applicable as long as \( \{(Y_i, X_i) : R_i = 1, i = 1, \ldots, N\} \) and \( \hat{\lambda} \) are available. This property seems incidental in the missing-data problem, but is crucial in the survey setting where \( \hat{\lambda} \) is known but \( (X_1, \ldots, X_N) \) may be inaccessible.

We point out that \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \) can be algebraically identified in the class of calibration estimators \( \hat{\mu}_{\text{cal}} \) including \( \hat{\mu}_{\text{reg}} \) for survey sampling, in spite of different arguments behind these estimators. In fact, \( \hat{\mu}_{\text{reg}} \) is identical to \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \) is identical to \( \hat{\mu}_{\text{cal}} \) for the weighted Kullback–Liebler distance, with \( d_i = p^{-1}(X_i) \) and \( q_i = p^{-1}(X_i) - 1 \). The first relationship follows directly by the formulas of \( \hat{\beta} \) and \( \hat{\beta}_q \). The second relationship follows because equation (6) for \( \hat{\lambda} \) is equivalent to the calibration equation \( \sum_{i=1}^{N} R_i w_i X_i = \sum_{i=1}^{N} X_i \) with \( w_i = d_i/(1 + q_i X_i^T \lambda) \) = \( [p(X_i) + \{1 - p(X_i)\} X_i^T \lambda]^{-1} \). These connections shed new light on both the calibrated estimators \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \) in the missing-data problem and the calibration estimators \( \hat{\mu}_{\text{cal}} \) in survey sampling. In particular, the derivation of \( \hat{\mu}_{\text{lik}} \) by empirical or nonparametric likelihood gives strong support for using the weighted Kullback–Liebler distance in \( \hat{\mu}_{\text{cal}} \).

We summarize asymptotic properties of the foregoing estimators in the missing-data problem (Tan, 2006, 2010a). Assume that \( p(X) > p_0 \) almost surely for a constant \( 0 < p_0 < 1 \), \( E(Y) < \infty \), \( E(||X||^2) < \infty \), and \( V \) is nonsingular, where \( ||X|| = (X^T X)^{1/2} \) and \( V = E(\xi \xi^T) = E\{p^{-1}(X) - 1\} X X^T \). The estimators \( \hat{\mu}_{\text{reg}}, \hat{\mu}_{\text{lik}}, \hat{\mu}_{\text{reg}}, \) and \( \hat{\mu}_{\text{lik}} \) admit the asymptotic
expansion
\[
\frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i}{p(X_i)} - \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{R_i}{p(X_i)} - 1 \right\} X_i^T \beta + o_p(N^{-1/2})
\]  
(7)

where \( \beta = V^{-1}U \) and \( U = E(\xi \eta) = E[(p^{-1}(X) - 1)XY] \). Each estimator asymptotically attains the minimum variance of estimators of \( \mu \) in the form similar to (3):

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i}{p(X_i)} - \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{R_i}{p(X_i)} - 1 \right\} X_i^T b,
\]

(8)

where \( b \) is a vector of arbitrary constants.

3.2. Poisson sampling

Now consider Poisson sampling, where \((R_1, \ldots, R_N)\) are independent Bernoulli random variables with draw probabilities \((p_1, \ldots, p_N)\). For clarity, we speak of \( p_i \) instead of \( \pi_i \), although \( \pi_i = p_i \) in Poisson sampling. As discussed in §1, Poisson sampling is essentially equivalent to the missing-data problem with \((Y_1, X_1), \ldots, (Y_N, X_N)\) fixed or conditioned on, although there remains the difference that \((Y_i, X_i)\)'s here may not be a random sample from a probability distribution. The basic equations \( \mu = E(\eta) \) and \( 0 = E(\xi) \) in the missing-data problem lead to \( \bar{Y} = E_r(N^{-1} \sum_{i=1}^{N} p_i^{-1} R_i Y_i) \) and \( 0 = E_r \{ N^{-1} \sum_{i=1}^{N} (p_i^{-1} R_i - 1) X_i \} \) in Poisson sampling.

To estimate \( \bar{Y} \), we apply the calibrated estimators \( \mu_{\text{reg}} \) and \( \mu_{\text{lik}} \) with \( p(X_i) \) replaced by \( p_i \). The two estimators take the form of calibration estimators \( \mu_{\text{cal}} \) with \( q_i = p_i^{-1} - 1 \) as discussed in §3.1. Assume that the expected sampling fraction is asymptotically non-negligible: \( \lim inf_{N \to \infty} N^{-1} \sum_{i=1}^{N} p_i > 0 \). Then asymptotic expansion (2) for the two estimators gives

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i}{p_i} - \frac{1}{N} \sum_{i=1}^{N} \left( \frac{R_i}{p_i} - 1 \right) X_i^T \beta_p + o_p(N^{-1/2}),
\]

(9)

where \( \beta_p = V_p^{-1}U_p \), \( V_p = N^{-1} \sum_{i=1}^{N} (p_i^{-1} - 1) X_i X_i^T \), and \( U_p = N^{-1} \sum_{i=1}^{N} (p_i^{-1} - 1) X_i Y_i \). Expansion (9) is identical to (7) in the missing-data problem except for the difference between \( \beta_p \) and \( \beta \). The vector \( \beta_p \) equals \( \beta_{\text{reg}} \) in §2 because \( V_p = N \text{var}(N^{-1} \sum_{i=1}^{N} p_i^{-1} R_i X_i) \) and \( U_p = N \text{cov}(N^{-1} \sum_{i=1}^{N} p_i^{-1} R_i X_i, N^{-1} \sum_{i=1}^{N} p_i^{-1} R_i Y_i) \) by independence of \((R_1, \ldots, R_N)\). Then \( \mu_{\text{reg}} \) and \( \mu_{\text{lik}} \) asymptotically attain the minimum variance of estimators (3). Therefore, \( \mu_{\text{reg}} \) and \( \mu_{\text{lik}} \) are efficient in the class of estimators (3) or (8), whether inference is conditional on \((Y_i, X_i)\)’s in Poisson sampling or unconditional in the missing-data problem.

By the preceding discussion, \( \mu_{\text{reg}} \) coincides with the optimal regression estimator known for Poisson sampling (e.g., Samdal, 1996, Result 9.1). It is interesting that \( \mu_{\text{reg}} \) appears more natural than \( \hat{\mu}_{\text{reg}} \) in Poisson sampling, because \( N^{-1} \sum_{i=1}^{N} p_i^{-1} R_i (p_i^{-1} - 1) X_i X_i^T \) appears a more natural estimator of \( V_p \) than \( N^{-1} \sum_{i=1}^{N} (p_i^{-1} R_i - 1)^2 X_i X_i^T \) and hence \( \beta \) appears a more natural estimator of \( \beta_p \) than \( \hat{\beta} \). In contrast, \( \mu_{\text{reg}} \) appears more natural than \( \mu_{\text{reg}} \) in the missing-data problem, because \( \mu_{\text{reg}} \) is a direct application of the classical regression estimator but \( \hat{\mu}_{\text{reg}} \) is derived as a modification of \( \hat{\mu}_{\text{reg}} \) to achieve calibration on \( X \) by Tan (2006).

3.3. Rejective sampling

Consider rejective sampling, which is Poisson sampling restricted to or conditioned on a fixed sample size. By Hajek (1964), a sampling design is called rejective sampling of size \( n \) with draw probabilities \((p_1, \ldots, p_N)\) if
\[ \Pr(R_1 = r_1, \ldots, R_N = r_N) = \begin{cases} c \prod_{i=1}^{N} p_i \prod_{i=0}^{c} (1 - p_i), & \sum_i r_i = n, \\ 0, & \text{otherwise}, \end{cases} \]  

(10)

where \( c \) is a constant. There exist infinitely many choices of \((p_1, \ldots, p_N)\) leading to the same sampling design. In fact, if \( p_i'/(1 - p_i') \propto p_i/(1 - p_i) \) for \( i = 1, \ldots, N \), then rejective sampling with probabilities \((p_1', \ldots, p_N')\) is identical to that with probabilities \((p_1, \ldots, p_N)\). The representation (10) is called canonical if \( \sum_{i=1}^{N} p_i = n \). It is important to distinguish draw probabilities \( p_i \) in the underlying Poisson sampling and inclusion probabilities \( \pi_i \) in the actual rejective sampling. As mentioned in §1, there are fast algorithms to compute \( p_i \) from \( \pi_i \) and vice versa and to draw rejective samples (Chen et al., 1994; Tillé, 2006).

To estimate \( \bar{Y} \), we include \( \pi_i \) as a component of \( X_i \), which is feasible because \( \sum_{i=1}^{N} \pi_i = n \) is known, and apply \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \) with \( p(X_i) \) replaced by \( \pi_i \). One of our main objectives is to show that the resulting estimators, denoted by \( \hat{\mu}_{\text{reg}, \pi} \) and \( \hat{\mu}_{\text{lik}, \pi} \), are simple and efficient:

(i) they are as simple as calibration estimators \( \hat{\mu}_{\text{case}} \) and \( \hat{\mu}_{\text{cal}} \) with \( q_i = \pi_i^{-1} - 1 \);

(ii) they attain the minimum asymptotic variance of estimators (3) and hence are asymptotically as efficient as an optimal regression estimator.

Property (i) holds as discussed in §3.1. A remarkable byproduct of including \( \pi_i \) in the auxiliary variables \( X_i \) is that \( \hat{\mu}_{\text{reg}, \pi} \) can then be recast in the cosmetic form of linear prediction estimators (Sarndal & Wright, 1984):

\[
\hat{\mu}_{\text{reg}, \pi} = \frac{1}{N} \sum_{i=1}^{N} X_i \hat{\beta} + \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} (Y_i - X_i \hat{\beta})
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} X_i \hat{\beta} + \frac{1}{N} \sum_{i=1}^{N} R_i (Y_i - X_i \hat{\beta}) = \frac{1}{N} \sum_{i \in \pi} Y_i + \frac{1}{N} \sum_{i \notin \pi} X_i \hat{\beta}.
\]

The second equation follows because \( N^{-1} \sum_{i=1}^{N} R_i \pi_i^{-1}(\pi_i^{-1} - 1)(Y_i - X_i \hat{\beta})X_i = 0 \) and \( \pi_i \) is a component of \( X_i \). There are previous estimators of the linear prediction form in Brewer (1979), Fuller & Isaki (1981), and Kim & Park (2006) among others. Moreover, the use of \( \pi_i \) as a calibration variable is previously considered by, for example, Wu & Rao (2006). The importance of incorporating design variables such as \( \pi_i \) is discussed by Montanari & Ranalli (2002) for generalized regression estimation. Nevertheless, our approach seems unique in effectively combining the choice \( q_i = \pi_i^{-1} - 1 \) and the inclusion of \( \pi_i \) in \( X_i \).

To show property (ii), we develop formal asymptotic theory and provide an interesting interpretation based on conditional inference. We adopt the asymptotic framework without imposing any super-population (e.g., Hajek, 1964, 1981). A sequence of rejective samples \( \{s(t)\} \) is drawn from a sequence of finite populations \( \{U(t)\} \) of sizes \( N(t) \), where \( N(t) \to \infty \) as \( t \to \infty \). All limiting processes are defined as \( t \to \infty \) or, equivalently, \( N(t) \to \infty \). For simplicity, the index \( t \) is suppressed in the notation. Consider the following regularity conditions:

(C1) \( \min(\pi_1, \ldots, \pi_N) > \pi_0 \) for a constant \( 0 < \pi_0 < 1 \);

(C2) \( N^{-1} \sum_{i=1}^{N} |Y_i|^{2+\gamma_1} = O(1) \) for a constant \( \gamma_1 > 0 \);

(C3) \( N^{-1} \sum_{i=1}^{N} \|X_i\|^{2+\gamma_2} = O(1) \) for a constant \( \gamma_2 > 0 \);

(C4) \( \lim_{N \to \infty} \lambda_{\min}(V_{\pi}) > 0 \), where \( V_{\pi} = N^{-1} \sum_{i=1}^{N} (\pi_i^{-1} - 1)X_iX_i^T \) and \( \lambda_{\min}(\cdot) \) denotes the smallest eigenvalue.

These conditions are directly informative on the finite population and the inclusion probabilities.

Recall that \( \pi_i \) is a component of \( X_i \). Then condition (C4) implies

\( \text{(C4') lim}_{N \to \infty} N^{-1} \sum_{i=1}^{N} \pi_i(1 - \pi_i) > 0 \)
If conditions (C1), (C2), and (C4') hold, then

Assume that \( \sum_{i=1}^{N} \pi_i (1 - \pi_i) \to \infty \), a basic assumption in Hajek’s asymptotic theory on rejective sampling. By either (C1) or (C4'), the sampling fraction is asymptotically non-negligible: \( \liminf_{N \to \infty} n/N = \liminf_{N \to \infty} N^{-1} \sum_{i=1}^{N} \pi_i > 0 \).

We first give sufficient conditions to apply Hajek’s (1964) asymptotic results on the Horvitz–Thompson estimator \( \hat{\mu}_{ht} = N^{-1} \sum_{i=1}^{N} \pi_i^{-1} R_i Y_i \).

**Lemma 1.** If conditions (C1), (C2), and (C4') hold, then \( \text{var}_r(\hat{\mu}_{ht}) = N^{-1} v_0 + o(N^{-1}) \) and hence \( \hat{\mu}_{ht} = \bar{Y} + O_p(N^{-1/2}) \), where \( v_0 = N^{-1} \sum_{i=1}^{N} (\pi_i^{-1} - 1) (Y_i - \beta_0 \pi_i)^2 \) and \( \beta_0 = \{ \sum_{i=1}^{N} \pi_i (1 - \pi_i) \}^{-1} \{ \sum_{i=1}^{N} (1 - \pi_i) Y_i \} \). If, further, \( \liminf_{N \to \infty} v_0 > 0 \), then \( N^{1/2} v_0^{-1/2}(\hat{\mu}_{ht} - \bar{Y}) \to N(0, 1) \), the standard normal distribution, in distribution.

We now state the asymptotic results on the calibrated estimators \( \hat{\mu}_{cal} \) and \( \hat{\mu}_{lik} \). Asymptotic expansion (11) can be seen from (2) for \( \hat{\mu}_{cal} \), but is shown under weaker and more elementary conditions than in Deville & Sarndal (1992).

**Theorem 1.** Assume that \( \pi_i \) is a component of \( X_i \). If conditions (C1)–(C4) hold, then \( \hat{\mu}_{reg} \) and \( \hat{\mu}_{lik} \) admit the asymptotic expansion

\[
\hat{\mu}_{reg} = \bar{Y} + O_p(N^{-1/2}),
\]

where \( \beta = V^{-1}_{ii} X_i \), \( V = N^{-1} \sum_{i=1}^{N} (\pi_i^{-1} - 1) X_i Y_i \), and \( V \) is defined in (C4). If, further, \( \liminf_{N \to \infty} v_0 > 0 \), then \( N^{1/2} v_0^{-1/2}(\hat{\mu}_{reg,p} - \bar{Y}) \to N(0, 1) \) and \( N^{1/2} v_0^{-1/2}(\hat{\mu}_{lik,p} - \bar{Y}) \to N(0, 1) \) in distribution, where

\[
v_0 = N^{-1} \sum_{i=1}^{N} (\pi_i^{-1} - 1)(Y_i - X_i X_i)^2.
\]

The two estimators attain the minimum asymptotic variance of estimators (3).

We provide an interpretation of the asymptotic variance \( v_0 \) in Theorem 1 based on conditional inference: inference from rejective sampling of size \( n \) is equivalent to conditional inference from Poisson sampling given that the sample size equals \( n \). Assume the canonical representation (10) with \( \sum_{i=1}^{N} R_i = n \). Consider the following estimator \( \hat{\mu}_{reg,p} \) include \( p_i \) as a component of \( X_i \) and apply \( p(X_i) \) replaced by \( p_i \), instead of \( \pi_i \). Then the asymptotic variance of \( N^{1/2}(\hat{\mu}_{reg,p} - \bar{Y}) \) under rejective sampling is the conditional asymptotic variance given \( \sum_{i=1}^{N} R_i = n \) under Poisson sampling. But \( N^{1/2}(\hat{\mu}_{reg,p} - \bar{Y}) \) is asymptotically uncorrelated with \( N^{1/2}(\sum_{i=1}^{N} R_i - n) = N^{-1/2} \sum_{i=1}^{N} (R_i - p_i) \) under Poisson sampling, with the asymptotic covariance

\[
N^{-1} \sum_{i=1}^{N} (p_i^{-1} - 1)(Y_i - X_i X_i^T) = 0,
\]

because \( N^{-1} \sum_{i=1}^{N} (p_i^{-1} - 1)(Y_i - \beta_p X_i) = 0 \) and \( p_i \) is a component of \( X_i \). Therefore, \( N^{1/2}(\hat{\mu}_{reg,p} - \bar{Y}) \) has the same asymptotic variance \( v_p \) under Poisson sampling and under rejective sampling. Finally, \( N^{1/2}(\hat{\mu}_{reg,p} - \bar{Y}) \) has asymptotic variance \( v_p \) because \( \hat{\mu}_{reg,p} = \hat{\mu}_{reg} + O_p(N^{-1/2}) \) and \( v_p = v_p + o(1) \), by the fact that the relative
errors of approximating \( \pi_i \)'s by \( p_i \)'s are uniformly of second order: \( \max_{i=1,\ldots,N} |(\pi_i/p_i) - 1| = O(N^{-1}) \) (Hajek 1981, Theorem 7.3).

The foregoing argument is even useful for understanding Hajek’s (1964) variance approximation for \( \hat{\mu} \) in Lemma 1. In fact, Lemma 1 is a special case of Theorem 1 where \( X_i = \pi_i \) is the only auxiliary variable and hence \( \beta_x = \beta_0 \) and \( v_x = v_0 \). The estimator \( \hat{\mu}_{\text{reg}, \pi} \) immediately reduces to \( \hat{\mu}_{\text{str}}, \) because \( N^{-1} \sum_{i=1}^{N} (R_i - \pi_i) = 0 \). Moreover, \( \hat{\mu}_{\text{lik}, \pi} \) reduces to \( \hat{\mu}_{\text{str}}, \) because equation (6) is satisfied by \( \lambda = 0 \). Such an argument is implicitly exploited in Hajek (1964). For example, Hajek (1964, §3) discussed an important quantity \( T \) in his proofs. But \( T/N \) is exactly the leading term in expansion (9) with \( \pi_i = p_i \) under Poisson sampling.

Fuller (2009) studied a rejective scheme such that a sample selected by a basic scheme is rejected until \( \sum_{i \in s} p_i^{-1} Z_i \) is within a specified distance from \( \sum_{i \in U} Z_i \) for an auxiliary vector \( Z_i \), where \( p_i \) is the \( i \)th inclusion probability in the basic scheme. Under several regularity conditions, Fuller showed that the asymptotic variance of a particular regression estimator, obtained using \( p_i \)'s, under the rejective scheme is the same as that of the regression estimator under the basic scheme. This result is similar to our result that \( \hat{\mu}_{\text{reg}, \pi} \) has the same asymptotic variance under Poisson sampling and under rejective sampling by conditional inference.

3.4. Simple random sampling

Consider simple random sampling without replacement as a special case of rejective sampling with \( \pi_1 = \cdots = \pi_N = n/N \). We illustrate how the calibrated estimators \( \hat{\mu}_{\text{reg}} \) and \( \hat{\mu}_{\text{lik}} \) and the asymptotic results are related to existing ones.

First, consider the simple case where \( X_i = \pi_i \) or equivalently \( X_i = 1 \). The estimators \( \hat{\mu}_{\text{reg}}, \hat{\mu}_{\text{lik}}, \hat{\mu}_{\text{reg}, \pi} \), and \( \hat{\mu}_{\text{lik}, \pi} \) are all reduced to the sample average \( \bar{y} = n^{-1} \sum_{i=1}^{N} R_i Y_i \). The exact variance of \( \bar{y} \) is \( n^{-1}(1 - n/N)S_{Y \cdot Y} \), where \( S_{Y \cdot Y} = (N - 1)^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \). For comparison, the asymptotic variance, \( N^{-1} v_{0} \), is

\[
\frac{1}{N} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( \frac{N}{n} - 1 \right) (Y_i - \bar{Y})^2 \right\} = \frac{1}{n} \left( 1 - \frac{n}{N} \right) \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \right\},
\]

which differs from the exact variance by \( O(N^{-2}) \). The finite population correction factor, \( 1 - n/N \), is derived from the weight \( \pi_i^{-1} - 1 = N/n - 1 \).

Next, consider the general case where \( X_i = (\pi_i, Z_i^T) \) with non-constant \( Z_i \). By simple algebra, the calibrated estimator \( \hat{\mu}_{\text{reg}, \pi} \), not \( \hat{\mu}_{\text{reg}} \), is reduced to the classical, optimal regression estimator (Cochran, 1977), \( \hat{\beta}_0 = \beta_0^\pi (\bar{z} - \bar{Z}) \), where \( \bar{z} = n^{-1} \sum_{i=1}^{N} R_i Z_i, \bar{Z} = N^{-1} \sum_{i=1}^{N} Z_i \), and \( \hat{\beta}_1 = (\sum_{i=1}^{N} R_i Z_i (Z_i - \bar{Z})^T)^{-1} \{\sum_{i=1}^{N} R_i Z_i (Y_i - \bar{Y})\} \). Moreover, \( \hat{\mu}_{\text{lik}, \pi} \), not \( \hat{\mu}_{\text{lik}} \), is reduced to the empirical likelihood estimator of Chen & Qin (1993), \( n^{-1} \sum_{i=1}^{N} (1 + \gamma^\chi (Z_i - \bar{Z}))^{-1} R_i Y_i \), where \( \gamma \) is determined by

\[
\hat{Z} = \frac{1}{n} \sum_{i=1}^{N} \frac{R_i Z_i}{1 + \gamma^\chi (Z_i - \bar{Z})}.
\]

Comparison of this equation with (6) shows that \( \gamma = (N/n - 1) \lambda_1 \) and \( 0 = (N/n - 1)(\lambda_0 n/N + \lambda_1^T \bar{Z}) \), where \( \lambda = (\lambda_0, \lambda_1^T)^T \). The asymptotic variance of \( \hat{\mu}_{\text{reg}} \) or \( \hat{\mu}_{\text{lik}} \) is

\[
\frac{v_{0}}{N} = \frac{1}{n} \left( 1 - \frac{n}{N} \right) \left[ \frac{1}{N} \sum_{i=1}^{N} \left\{ Y_i - \bar{Y} - \beta_1^\chi (Z_i - \bar{Z}) \right\}^2 \right],
\]

where \( \beta_1 = \{\sum_{i=1}^{N} Z_i (Z_i - \bar{Z})^T\}^{-1} \{\sum_{i=1}^{N} Z_i (Y_i - \bar{Y})\} \). As in the simple case of \( X_i = 1 \), the finite population correction factor is linked to the weight \( \pi_i^{-1} - 1 \).
3.5. High-entropy sampling

A sampling design of fixed size $n$ without replacement corresponds to a probability distribution, $P(s)$, on the set $S$ of all possible samples $s$ of $n$ distinct units from $U$. The entropy of the sampling design $P(s)$ is $H(P) = -\sum_{s\in S} P(s) \log\{P(s)\}$. Rejective sampling, denoted by $R(s)$, is the unique sampling design which maximizes the entropy among all fixed-size sampling designs without replacement with fixed first-order inclusion probabilities (Hajek, 1981). Then a sampling design with a high entropy is close to rejective sampling. The closeness in approximating $P(s)$ by $R(s)$ can be measured by the Kullback–Liebler divergence

$$D(P\|R) = \sum_{s\in S} P(s) \log \frac{P(s)}{R(s)}.$$  

Berger (1988a) studied asymptotic normality of the Horvitz–Thompson estimator $\hat{\mu}_{HT}$ for high-entropy sampling with the Kullback–Liebler divergence tending to zero. By building on his work, we generalize Theorem 1 on the calibrated estimators $\hat{\mu}_{reg,\pi}$ and $\hat{\mu}_{lik,\pi}$ from rejective sampling to high-entropy sampling. Recall that conditions (C1)–(C4) involve only the finite population and the first-order inclusion probabilities.

**Theorem 2.** Suppose that $P(s)$ is a sampling design with the first-order inclusion probabilities $(\pi_1, \ldots, \pi_N)$. If there exists rejective sampling $R(s)$ such that $D(P\|R) \to 0$ as $N \to \infty$, then Theorem 1 holds under $P(s)$.

For Rao–Sampford sampling $P(s)$ with inclusion probabilities $(\pi_1, \ldots, \pi_N)$, let $R(s)$ be rejective sampling with draw probabilities $p_i(R)$ equal to $\pi_i$ ($i = 1, \ldots, N$). Berger (1988a) showed that $D(P\|R) \to 0$ if $\sum_{i=1}^{N} \pi_i (1 - \pi_i) \to \infty$, which is already implied by condition (C4) because $\pi_i$ is a component of $X_i$. Therefore, Theorem 1 is directly valid for Rao–Sampford sampling.

For successive sampling $P(s)$ (Hajek, 1981), Berger (1988a) showed that there exists rejective sampling $R(s)$ such that $D(P\|R) \to 0$ if $n^2/N \to 0$, which violates $\lim \inf_{N \to \infty} n/N > 0$ under condition (C4'). It may be of interest to study asymptotic theory for rejective sampling when $n/N \to 0$ but $n \to \infty$ or for successive sampling when $n/N \not\to 0$.

Berger (1988b, 2005) discussed the Kullback–Liebler divergence and the validity of Hajek’s (1964) variance approximation for other sampling designs including randomized systematic sampling. Further research on this topic is desired.

4. Model calibration

Consider the situation of complete auxiliary information, where $X_i$ is known for each $i \in U$. Then for any function $h(x)$, $\sum_{i=1}^{N} h(X_i)$ is known and hence $h(X_i)$ serves as an auxiliary variable for linear calibration as discussed in §3. To use such information, the model-calibration approach of Wu & Sitter (2001) consists of two steps. First, a possibly nonlinear regression model is specified:

$$E_{\epsilon}(Y_i \mid X_i) = \mu(X_i; \theta), \quad \text{var}_{\epsilon}(Y_i \mid X_i) = \sigma^2 \nu(X_i; \theta), \quad (12)$$

where $\theta$ and $\sigma^2$ are unknown parameters and $\mu(\cdot; \theta)$ and $\nu(\cdot)$ are known functions. Let $\hat{\theta}$ be the maximum weighted quasi-likelihood estimator of $\theta$ solving

$$0 = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} \frac{\partial \mu(X_i; \theta)}{\partial \theta} Y - \mu(X_i; \theta) \nu(X_i; \theta).$$
Second, a calibration estimator, for example, \( \hat{\mu}_{\text{cal}} \), is applied with \( X_i \) replaced by the fitted value \( \mu(X_i; \hat{\theta}) \) or the vector \( \{1, \mu(X_i; \hat{\theta})\}^T \) with 1 included. The resulting estimator, \( \hat{\mu}_{W,S} \), attains the minimum anticipated asymptotic variance among all calibration estimators \( \hat{\mu}_{\text{cal}} \) with \( X_i \) replaced by any function \( h(X_i) \) if model (12) is correctly specified (Wu, 2003). We point out two possible ways to improve efficiency while allowing that model (12) may be misspecified.

First, the estimator \( \hat{\mu}_{\text{reg},\pi} \) or \( \hat{\mu}_{\text{lik},\pi} \) can be used with \( X_i \) replaced by \( \{\pi_i, 1, \mu(X_i; \hat{\theta})\}^T \) or, if \( \pi_i \) is constant, \( \{1, \mu(X_i; \hat{\theta})\}^T \) in the second step of model calibration. Under additional regularity conditions as in Wu & Sitter (2001), Theorem 1 holds with \( X_i \) replaced by \( \{\pi_i, 1, \mu(X_i; \theta^*)\}^T \) in expansion (11) and the asymptotic variance \( v_\pi \), where \( \theta^* \) is the finite population counterpart of \( \theta \) solving \( 0 = N^{-1} \sum_{i=1}^N \{\partial \mu(X_i; \theta)/\partial \theta\} \nu^{-1}(X_i; \theta)\{Y - \mu(X_i; \theta)\} \). Then regardless of model (12), \( \hat{\mu}_{\text{reg},\pi} \) or \( \hat{\mu}_{\text{lik},\pi} \) attains the minimum asymptotic variance under rejective sampling of estimators (3) or

\[
\frac{1}{N} \sum_{i=1}^N \frac{R_i Y_i}{\pi_i} - \frac{1}{N} \sum_{i=1}^N \left( \frac{R_i}{\pi_i} - 1 \right) \{b_1 + 2b \mu(X_i; \theta^*)\},
\]

where \( (b_1, b_2) \) are arbitrary constants. This class contains, up to asymptotic equivalence, both \( \hat{\mu}_{W,S} \) and Cassel et al.’s (1976) generalized difference estimator using \( \mu(X_i; \hat{\theta}) \) as the calibration variable

\[
\hat{\mu}_{\text{con}} = \frac{1}{N} \sum_{i=1}^N \frac{R_i Y_i}{\pi_i} - \frac{1}{N} \sum_{i=1}^N \left( \frac{R_i}{\pi_i} - 1 \right) \mu(X_i; \hat{\theta}).
\]

Therefore, \( \hat{\mu}_{\text{reg},\pi} \) or \( \hat{\mu}_{\text{lik},\pi} \) is asymptotically guaranteed to gain efficiency over \( \hat{\mu}_{W,S} \) and \( \hat{\mu}_{\text{con}} \) under rejective sampling when model (12) may be misspecified. The estimator \( \hat{\mu}_{W,S} \) is not necessarily more efficient that \( \hat{\mu}_{\text{con}} \), as shown in a simulation study in §5. On the other hand, it can be shown that if model (12) is correctly specified, then \( \hat{\mu}_{\text{reg},\pi} \), \( \hat{\mu}_{\text{lik},\pi} \), \( \hat{\mu}_{W,S} \), and \( \hat{\mu}_{\text{con}} \) are asymptotically as efficient as each other under suitable regularity conditions as in Wu (2003). The comparison of \( \hat{\mu}_{\text{reg},\pi} \) and \( \hat{\mu}_{\text{lik},\pi} \) versus \( \hat{\mu}_{W,S} \) and \( \hat{\mu}_{\text{con}} \) is similar to that of \( \hat{\mu}_{\text{reg}} \) versus the estimator of Robins et al. (1994) in the missing-data problem discussed in Tan (2006).

The second method for improving efficiency involves choosing an alternative estimator of \( \theta \) for model (12), similarly as in the method of efficiency maximization in the missing-data literature (Rubin & van der Laan, 2008; Tan, 2008; Cao et al., 2009). First, consider the calibrated estimator

\[
\hat{\mu}_{\text{reg},\pi}(\theta) = \hat{\mu}_{\text{reg},\pi}(\theta) \quad \text{or} \quad \hat{\mu}_{\text{lik},\pi}(\theta) = \hat{\mu}_{\text{lik},\pi}(\theta)
\]

with \( X_i \) replaced by \( \{\pi_i, 1, \mu(X_i; \theta)\}^T \) for fixed \( \theta \). By Theorem 1, the asymptotic variance of each estimator is \( N^{-1} v_\pi(\theta) \) under rejective sampling, where \( v_\pi(\theta, b) = \min_b v_\pi(\theta, b) \) with \( b = (b_0, b_1, b_2)^2 \) and

\[
v_\pi(\theta, b) = \frac{1}{N} \sum_{i=1}^N \left( \pi_i^{-1} - 1 \right) \{Y_i - b_0 \pi_i - b_1 - 2b_2 \mu(X_i; \theta)\}^2.
\]

Recall that \( \beta_\pi = \beta(\pi; \theta) \) in Theorem 1 is a minimizer of \( v_\pi(\theta, b) \), whereas \( \tilde{\beta} = \tilde{\beta}(\theta) \) in \( \hat{\mu}_{\text{reg},\pi} \) is a minimizer of the empirical counterpart of \( v_\pi(\theta, b) \):

\[
\tilde{v}(\theta, b) = \frac{1}{N} \sum_{i=1}^N \frac{R_i}{\pi_i} \left( \frac{1}{\pi_i} - 1 \right) \{Y_i - b_0 \pi_i - b_1 - 2b_2 \mu(X_i; \theta)\}^2.
\]

The estimators \( \hat{\mu}_{\text{reg},\pi}(\theta) \) and \( \hat{\mu}_{\text{lik},\pi}(\theta) \) using \( \tilde{\theta} \) are discussed in the preceding paragraph. Alternatively, consider the estimator \( \tilde{\mu}_{\text{reg},\pi}(\tilde{\theta}) \) or \( \tilde{\mu}_{\text{lik},\pi}(\tilde{\theta}) \) using \( \tilde{\theta} \), where \( \tilde{\theta} \) is a minimizer of the empirical counterpart of \( v_\pi(\theta) \), that is, \( \tilde{\theta} = \min_b \tilde{v}(\theta, b) \) or, equivalently, \( \{\tilde{\theta}, \tilde{\beta}(\tilde{\theta})\} \) is jointly
a minimizer of $\bar{v}(\theta, b)$. For these estimators, it can be shown that under additional regularity conditions, Theorem 1 holds with $X_i$ replaced by $(\pi_i, 1, \mu(X_i; \theta^1))^T$ in expansion (11) and the asymptotic variance $v_\pi$, where $\theta^1$ is a minimizer of $v_\pi(\theta)$ or, equivalently, $\{\theta^1, \beta_\pi(\theta^1)\}$ is jointly a minimizer of $v_\pi(\theta, b)$. Therefore, $\hat{\mu}_{reg, \pi}(\theta)$ or $\hat{\mu}_{lik, \pi}(\theta)$ attains the minimum asymptotic variance under rejective sampling of estimators in the form

$$\frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} Y_i - \frac{1}{N} \sum_{i=1}^{N} \left( \frac{R_i}{\pi_i} - 1 \right) \{b_1 + b_2 \mu(X_i; \theta)\},$$

where $(b_1, b_2)$ and $\theta$ are arbitrary constants. This class of estimators is larger than (13) in allowing $\theta$ different from $\theta^*$. If model (12) is correctly specified, then $\theta^*$ seems natural, but both $\theta^*$ and $\theta^1$ converge in probability to the true $\theta$ so that $\hat{\mu}_{reg, \pi}(\theta)$ and $\hat{\mu}_{lik, \pi}(\theta)$ are asymptotically as efficient as $\hat{\mu}_{reg, \pi}(\theta)$ and $\hat{\mu}_{lik, \pi}(\theta)$. If model (12) is misspecified, then the latter estimators using $\bar{\theta}$ are asymptotically more efficient under rejective sampling than the former estimators using $\theta$.

The problem of minimizing $\bar{v}(\theta, b)$ is computationally nontrivial. The function $\mu(X_i; \theta)$ is typically specified as $\Psi(\theta_0 + Z_i^T \theta_1)$, where $\Psi(\cdot)$ is an inverse link function, $X_i = (1, Z_i^T)^T$, and $\theta = (\theta_0, \theta_1)^T$. For $\theta_1$ close to 0, $\mu(X_i; \theta)$ is approximately $\exp(\theta_0 + Z_i^T \theta_1)$ and hence linearly dependent with the constant term. Moreover, if $\mu(X_i; \theta) = \exp(\theta_0 + Z_i^T \theta_1)$ as in the log-normal model in §5, then $\exp(\theta_0)$ cannot be separately identified from $b_2$. For simplicity, consider the estimator $\hat{\theta}$ minimizing the empirical variance under rejective sampling of the generalized difference estimator $\hat{\mu}_{GDIF}(\theta)$, with $\theta$ replaced by fixed $\theta$. Equivalently, $\theta$ together with some $b_0$ minimizes

$$\bar{v}\{\theta, (b_0, 0, 1)\} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} \left( \frac{1}{\pi_i} - 1 \right) \{Y_i - b_0 \pi_i - \mu(X_i; \theta)\}^2.$$

By similar reasoning as above, $\hat{\mu}_{GDIF}(\theta)$ attains the minimum asymptotic variance under rejective sampling of estimators $\hat{\mu}_{GDIF}(\theta)$ for arbitrary $\theta$. For the calibration estimators $\hat{\mu}_{WS}, \hat{\mu}_{reg, \pi},$ and $\hat{\mu}_{lik, \pi}$, using the fitted value $\mu(X_i; \theta)$ may still be more efficient than using $\mu(X_i; \theta)$ as an auxiliary variable.

## 5. Simulation study

We conduct a simulation study to compare various methods in the presence of a misspecified super-population model. Two finite populations of size $N = 1000$ are generated and then fixed to study, respectively, linear calibration and model calibration. For each finite population, 50,000 Monte Carlo samples of size $n = 100$ are drawn by rejective sampling using the R package *Sampling* (Tillé, 2011). From each sample, various estimators of $Y$ are then computed.

For linear calibration, a finite population is generated as an independent and identically distributed sample from the model:

$$W_i \sim \chi^2_2, \quad Z_i = 2 + 0.5 W_i + \delta_i, \quad (14)$$

$$Y_i = 1 + Z_i + \{(Z_i - 3)_+\}^2 + \epsilon_i, \quad (15)$$

where $\delta_i \sim N(0, 1), \epsilon_i \sim N(0, 1)$, and $a_+ = a$ if $a > 0$ or $0$ if $a \leq 0$. The joint distribution of $(W_i, Z_i)$ is taken from Kim (2009). The inclusion probabilities $\pi_i$ are set proportional to $W_i$. The 0%, 25%, 50%, 75%, and 100% quantiles of $\pi_i$ are $7.8 \times 10^{-6}, 0.029, 0.069, 0.14, 0.78$. The sampling weights $\pi_i^{-1}$ are highly variable. The distribution of $Y_i$ is specified to reflect a nonlinear, but monotone, relationship between $Y_i$ and $Z_i$. We compare the four estimators, $\hat{\mu}_{int}, \hat{\mu}_{GDIF},$ with $q_i = 1$, $\hat{\mu}_{reg, \pi}$, and $\hat{\mu}_{lik, \pi}$, in two situations, (I) and (II). For situation (I), $\hat{\mu}_{GDIF}, \hat{\mu}_{reg, \pi},$
other, but also a smaller absolute bias than $\hat{\mu}_{\text{reg, }\pi}$. The ratio of mean squared errors of $\hat{\mu}_{\text{GDIF}}$ versus $\hat{\mu}_{\text{reg, }\pi}$ is $(409/248)^2 = 2.7$ in situation (I) or $(262/209)^2 = 1.6$ in situation (II). This ratio is greater in situation (I) than in (II), perhaps because a linear model of $Y_i$ given $Z_i$ is a more severe departure from the truth in situation (I) than in (II). Third, $\hat{\mu}_{\text{reg, }\pi}$ and $\hat{\mu}_{\text{lik, }\pi}$ have similar variances to each other, but $\hat{\mu}_{\text{lik, }\pi}$ has a much smaller absolute bias than $\hat{\mu}_{\text{reg, }\pi}$.

In Supplementary Material, we present additional results for linear calibration in two cases where the joint distribution of $(Z_i, Y_i)$ remains as before but the variability of the sampling weights $\pi_i^{-1}$ decreases. The ratios of mean squared errors of $\hat{\mu}_{\text{GDIF}}$ versus $\hat{\mu}_{\text{reg, }\pi}$ and $\hat{\mu}_{\text{lik, }\pi}$ remain noticeably above 1, but become smaller as the sampling weights are more homogeneous. Moreover, we report results on variance estimation and confidence interval coverage, using the Horvitz–Thompson variance estimator depending on second-order inclusion probabilities and using the empirical counterpart of asymptotic variance in Theorem 1. Both methods perform reasonably well when sampling weights are moderately or less variable, but lead to under-coverage of confidence intervals in the case of highly variable sampling weights.

For model calibration, a finite population is generated as an independent and identically distributed sample from (14) for $(W_i, Z_i)$ and (15) replaced by

$$\log Y_i = 0.5 + 0.5 Z_i - 0.2 (Z_i - 3) + 0.2 \varepsilon_i,$$

where $\varepsilon_i \sim N(0, 1)$. The distribution of $Y_i$ given $Z_i$ is log-normal, similarly as in Wu & Sitter (2001). Consider the log-linear model (12) for $Y_i$ given $Z_i$: $\mu(Z_i; \theta) = \exp(\theta_0 + Z_i^T \theta_1)$ and $\nu(Z_i; \theta) = \mu^2(Z_i; \theta)$, where $\theta = (\theta_0, \theta_1)^T$. We compare the five estimators, $\hat{\mu}_{\text{tr}}, \hat{\mu}_{\text{GDIF}}, \hat{\mu}_{\text{reg, }\pi}$ with $\hat{\mu}_{\text{GDIF}}, \hat{\mu}_{\text{reg, }\pi}, \hat{\mu}_{\text{lik, }\pi}$ in two situations, (I) and (II). For situation (I), $\hat{\mu}_{\text{GDIF}}, \hat{\mu}_{\text{reg, }\pi}$, and $\hat{\mu}_{\text{lik, }\pi}$ are applied with $X_i = (\pi_i, 1, Z_i, (Z_i - 3)_+)^T$, where $\hat{\theta}$ is the maximum weighted quasi-likelihood estimator. For situation (II), the same estimators as in (I), including $\hat{\mu}_{\text{GDIF}}$, are applied, except that $\mu(Z_i; \hat{\theta})$ is replaced by $\mu(Z_i; \hat{\theta})$ and $\hat{\theta}$ is obtained by efficiency maximization in §4.

The right panel of Table 1 shows the biases and root mean squared errors of the five estimators of $Y = 7.89$. First, the relative performances of $\hat{\mu}_{\text{GDIF}}, \hat{\mu}_{\text{reg, }\pi}, \hat{\mu}_{\text{lik, }\pi}$ are similar to those for linear calibration. Second, the generalized difference estimator $\hat{\mu}_{\text{GDIF}}$ has a mean squared error, smaller than that of $\hat{\mu}_{\text{GDIF}}$ and similar to that of $\hat{\mu}_{\text{reg, }\pi}$ or $\hat{\mu}_{\text{lik, }\pi}$. This comparison confirms that $\hat{\mu}_{\text{GDIF}}$ may not be as efficient as $\hat{\mu}_{\text{GDIF}}$, whereas $\hat{\mu}_{\text{reg, }\pi}$ and $\hat{\mu}_{\text{lik, }\pi}$ are asymptotically no less efficient than $\hat{\mu}_{\text{GDIF}}$. The fact that $\hat{\mu}_{\text{reg, }\pi}$ and $\hat{\mu}_{\text{lik, }\pi}$ are roughly as efficient as $\hat{\mu}_{\text{GDIF}}$ indicates that for this example, there is little efficiency to be gained by linearly transforming the fitted values. Third,

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Linear calibration</th>
<th></th>
<th>Model calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
</tr>
<tr>
<td>$\hat{\mu}_{\text{tr}}$</td>
<td>-2.4</td>
<td>1041</td>
<td>—</td>
</tr>
<tr>
<td>Situation I</td>
<td>Situation II</td>
<td></td>
<td>Situation I</td>
</tr>
<tr>
<td>$\hat{\mu}_{\text{GDIF}}$</td>
<td>-141.3</td>
<td>409</td>
<td>21.9</td>
</tr>
<tr>
<td>$\hat{\mu}_{\text{reg, }\pi}$</td>
<td>-67.3</td>
<td>248</td>
<td>6.8</td>
</tr>
<tr>
<td>$\hat{\mu}_{\text{lik, }\pi}$</td>
<td>-40.6</td>
<td>240</td>
<td>3.5</td>
</tr>
</tbody>
</table>

RMSE, root mean squared error; all figures are multiplied by 1,000.
the mean squared errors of $\hat{\mu}_{\text{case}}$, $\hat{\mu}_{\text{reg}}$, and $\hat{\mu}_{\text{lik}}$ in situation (II) are similar to those in situation (I). For these estimators, little improvement is possible here from using different estimators of $\theta$ to construct fitted values. Nevertheless, $\hat{\mu}_{\text{case}}$ has a noticeably smaller mean squared error in situation (II) than in (I).

6. Conclusion

We develop efficient estimators for both linear and model calibration under rejective and high-entropy sampling. A general approach is to extend theory and methods for missing-data problems with independent and identically distributed data to the survey setting. Previous discussions on connections between survey sampling and missing-data problems have shown how calibration estimation in survey sampling is useful for handling missing-data problems (e.g., Kang & Schafer, 2007; Lumley et al., 2011). Our work shows, in an opposite direction, how efficient estimators can be extended from missing-data problems to survey sampling.

Our development mainly deals with point estimation. For variance estimation, it is straightforward to employ the Horvitz–Thompson variance estimator and, without using second-order inclusion probabilities, the empirical counterpart of asymptotic variance in Theorem 1. Our simulation study suggests that the performances of these methods may deteriorate when sampling weights are highly variable. Further research is warranted to address this issue.

Acknowledgement

The author thanks Jae-Kwang Kim and two referees for valuable comments leading to various improvements. This research was supported by the U.S. National Science Foundation.

Supplementary Material

Supplementary material available at *Biometrika* online includes proofs of Lemma 1 and Theorems 1–2 and additional simulation results mentioned in §5.

References


[Received April 2012. Revised September 2012]