The Annals of Statistics
1998, Vol. 26, No. 5, 1894–1915

WEIGHTED POLYNOMIAL MODELS AND WEIGHTED SAMPLING SCHEMES FOR FINITE POPULATION[1]

BY SEAN X. CHEN

New York University

This paper outlines a theoretical framework for finite population models with unequal sample probabilities, along with sampling schemes for drawing random samples from these models. We first present four *exact* weighted sampling schemes that can be used for any finite population model to satisfy such requirements as ordered/unordered samples, with/without replacement, and fixed/nonfixed sample size. We then introduce a new class of finite population models called *weighted polynomial models* or, in short, WPM. The probability density of a WPM is defined through a symmetric polynomial of the weights of the units in the sample. The WPM is shown to have been applied in many statistical analyses including survey sampling, logistic regression, case-control studies, lottery, DNA sequence alignment and MCMC simulations. We provide general strategies that can help improve the efficiency of the exact weighted sampling schemes for any given WPM. We show that under a mild condition, sampling from any WPM can be implemented within polynomial time. A Metropolis–Hasting-type scheme is proposed for *approximate* weighted sampling when the exact sampling schemes become intractable for moderate population and sample sizes. We show that under a mild condition, the average acceptance rate of the approximate sampling scheme for any WPM can be expressed in closed form using only the inclusion probabilities.

1. Introduction. Statistical analysis for finite populations often requires the specifications of (1) a probability model for samples from the finite population and (2) a sampling scheme that produces samples according to the probability model. In most situations, it is also assumed that the probability model is invariant under the permutation of the indexes of the units, or equivalently, the units are exchangeable. When little information about the characteristics of the population units is known, an equal probability model is often employed for convenience. For example, when samples of n distinct units are to be drawn from a population of N units, the equal probability model has a density function with the probability $n!(N-n)!/N!$ for each possible sample. The sampling scheme corresponding to an equal probability model is called *simple random sampling* or, in short, SRS, in which each unit is equally likely to be chosen into the sample. Except when dealing with

Received June 1996; revised March 1998.
[1]Supported in part by ARO Grant DAAL-0391-0089 and NSF Grant DMS-94-04396.
AMS 1991 *subject classifications*. Primary 62D05, 62E15; secondary 62E25.
Key words and phrases. Finite population, Metropolis–Hasting algorithm, polynomial theory, survey sampling, weighted sampling.

abstract objects (e.g., computer memory space, playing cards), an equal probability model along with SRS often produces unsatisfactory results (e.g., large bias or variance for the estimate) or sometimes "absurd" results (e.g., negative estimate for variance). Therefore, any auxiliary information about the population units should be utilized when available. If all available auxiliary information can be summarized by a single variable, which will be called *weight*, then the resulting probability model can be called weighted model in general. By analogy, any sampling scheme that draws random samples from a weighted model can be called weighted sampling scheme.

Many existing finite population models are of the weighted model type. In the context of survey sampling [e.g., Hanif and Brewer (1980)], "sampling with probabilities proportional to sizes" (in short, PPS), or in a broader sense, "sampling with unequal probabilities," often results in a weighted model. Some models are in fact of weighted model type but were not recognized by their inventors. In this paper, we present a unified theoretical framework that incorporates all such models.

The use of weighted models is twofold. One is as design tools for data collection and error control. This is the typical situation in survey sampling [e.g., Chen, Dempster and Liu (1994), Lahiri (1951), Singh and Srivastava (1980)], in which data are to be collected on samples that are randomly drawn from a probability model through a sampling scheme. In such cases, both the probability model and the sampling scheme must be specified. The other use of weighted models is as modeling and simulation tools for observed data. For example, the conditional Bernoulli model (a weighted model type) was applied by Chen and Liu (1997) to logistic regression and case-control studies, and by Stern and Cover (1989) in the fitting of Canada's Lotto 6/49 data. In such cases, the probability model must be specified, while the sampling scheme can be used as an option to simulate samples from the weighted model for estimation and hypothesis testing of parameters and lack of fit.

Often times, a major obstacle in using a desired weighted model is the difficulty of drawing random samples from it due to the tremendous amount of computation involved. In this paper, we present several weighted sampling schemes that can be used for different purposes. With our special strategy, these weighted sampling schemes can be implemented efficiently.

The rest of this paper is organized as follows.

In Section 2, we present four *exact* weighted sampling schemes for drawing random samples from any weighted model. These schemes serve general sampling purposes, taking into account a wide variety of requirements such as order/unordered samples, with/without replacement, and fixed/nonfixed sample size. The use of these schemes is by no means limited to the "weighted polynomial models" described in Sections 3 and 4, where the main focus is placed upon the flexibility of these models and the efficiency of the corresponding sampling schemes.

In Section 3, we introduce a new class of weighted models called *weighted polynomial models* or, in short, WPM. The probability density of a WPM is determined, up to a normalizing constant, by a symmetric polynomial of the

weights of the units in the sample. Several models found in statistical analyses including survey sampling, logistic regression, case-control studies, lottery, biological sequence analysis and MCMC simulations can be identified as members of this class. We use these applications to show the flexibility and potential of the WPM in finite population studies.

In Section 4, we provide general strategies that can help improve the efficiency of the exact weighted sampling schemes described in Section 2 for drawing samples from the seemingly complex WPM. The use of these strategies is illustrated with two examples from Section 3. We show that under a mild condition, sampling n units from N population units for any given WPM can be implemented within $O(nN^2)$ operations, as supposed to $O(N^n)$ operations needed for a whole-sample scheme. An even higher efficiency can be achieved when the degree of the polynomial in a WPM is less than the sample size.

In Section 5, we propose an *approximate* weighted sampling scheme for drawing samples from any weighted model (not limited to the WPM). This scheme is suitable when the probability density is explicitly defined (sometimes only up to a normalizing constant) but does not allow for an efficient implementation of any exact sampling scheme, especially when the sample space is too large to enumerate. The scheme we suggest is in fact a direct application of the Metropolis–Hasting algorithm. The convergence rate of a Metropolis–Hasting type algorithm is closely related to the average acceptance rate of the algorithm. We derive a closed form formula for the average acceptance rate based only on the inclusion probabilities when the finite population is "monotone." We show that under a mild condition, the average acceptance rate for any WPM can be expressed in the same closed form formula.

2. Exact weighted sampling schemes. In this section, we present four exact sampling schemes that can be used in general finite population situations. For convenience, we present the schemes for the case when the sample units are unordered and drawn without replacement, and the sample size is fixed (except for Scheme 4, which is particularly designed for the case of nonfixed sample size). With minor modifications, these schemes can be easily generalized to situations in which certain combinations of ordered/unordered samples, with/without replacement, and fixed/nonfixed sample size are required. These schemes are also applicable in any single stage of a complex sampling design such as multistage sampling and stratified sampling.

Throughout this paper, we use N and n to denote the population size and the *fixed* sample size, respectively. Let $\mathscr{S} = \{1, \ldots, N\}$ be the index set of the population units. For any subset $A \subset \mathscr{S}$, its size is denoted by $|A|$ and its complement, A^c. An unordered sample $s = \{i_1, \ldots, i_n\}$ is essentially a subset of $\mathscr{S}$. The sample space, denoted by Ω, is the collection of all possible samples. The formal definitions of a "sampling design" and a "sampling scheme" are given as follows.

DEFINITION 1. A *sampling design* is a probability measure p defined on Ω that satisfies $p(s) \geq 0$ for all $s \in \Omega$ and $\sum_{s \in \Omega} p(s) = 1$.

DEFINITION 2. A *sampling scheme* is a process of selecting samples from Ω according to a sampling design p.

The first scheme is based on the enumeration of the entire sample space, and is often referred to as the "whole-sample" scheme.

SCHEME 1. Line up all possible samples in any arbitrary order to form a queue, say $s_1, s_2, \ldots, s_{|\Omega|}$. Draw a random number u uniformly from $[0, 1)$. Starting from the first sample in the queue, check one sample at a time and take the first sample, say s_k, that satisfies $\sum_{i=1}^{k} p(s_i) \geq u$.

In general, when the population size and/or the sample size get moderately large, the whole-sample scheme can quickly become intractable. For example, when $N = 100$ and $n = 10$, a whole-sample scheme needs to deal with $N!/[n!(N-n)!] = 1.73 \times 10^{13}$ possible samples. It would take a one-billion-operations-per-second computer 55 years to draw one sample! A good way to get around this problem is to use a draw-by-draw scheme, which draws one unit at a step and only deals with a maximum of N units at each step.

For the following two draw-by-draw schemes (Schemes 2 and 3), let s_k denote the set of units selected into the sample after k steps for $k = 1, \ldots, N$. The first draw-by-draw scheme repeatedly selects one unit at a time from the unselected units with an appropriate probability until n units are obtained.

SCHEME 2. Start with $s_0 = \varnothing$. At step k ($k = 1, \ldots, n$), a unit $j \in \mathscr{S} \setminus s_{k-1}$ is selected into the sample (i.e., $s_k \leftarrow s_{k-1} \cup \{j\}$) with probability

$$P_1(j \mid s_{k-1}) = \sum_{s \in \Omega,\, s_{k-1} \subset s,\, j \in s} p(s) \Bigg/ \left[(n-k+1) \sum_{s \in \Omega,\, s_{k-1} \subset s} p(s) \right].$$

The process stops after n units are selected into the sample.

Scheme 2 indeed draws a sample from p. The proof is straightforward by the "telescope" law,

$$\begin{aligned} &P\{s = \{i_1, \ldots, i_n\}\} \\ &\quad = P\{s_1 = \{i_1\}, s_2 \setminus s_1 = \{i_2\}, \ldots, s_n \setminus s_{n-1} = \{i_n\}\} \\ &\quad = \left[nP\{s_1 = \{i_1\}\} \right] \left[(n-1) P\{s_2 \setminus s_1 = \{i_2\} \mid s_1 = \{i_1\}\} \right] \\ &\qquad \cdots \left[P\{s_n \setminus s_{n-1} = \{i_n\} \mid s_1 = \{i_1\}, \ldots, s_{n-1} \setminus s_{n-2} = \{i_{n-1}\}\} \right]. \end{aligned}$$

In the second draw-by-draw scheme, the population units are considered one at a time sequentially from unit 1 to unit N, and each time the unit being considered is selected into the sample with an appropriate probability. Let $A_k = \{1, \ldots, k\}$ for $k = 1, \ldots, N$.

SCHEME 3. Start with $s_0 = \varnothing$. At step k $(k = 1, \ldots, N)$, the kth unit is selected into the sample (i.e., $s_k \leftarrow s_{k-1} \cup \{k\}$) with probability

$$P_2(k \mid s_{k-1}) = \sum_{\substack{s \in \Omega,\, s_{k-1} \subset s,\, k \in s \\ (A_{k-1} \setminus s_{k-1}) \subset (\mathscr{S} \setminus s)}} p(s) \Bigg/ \sum_{\substack{s \in \Omega,\, s_{k-1} \subset s \\ (A_{k-1} \setminus s_{k-1}) \subset (\mathscr{S} \setminus s)}} p(s).$$

The process stops after n units are selected into the sample.

It is easy to see that Scheme 3 indeed generates a sample from p by the "telescope" law:

$$\begin{aligned} &P\{s = \{i_1, \ldots, i_n\}\} \\ &\quad = P\{1 \in s, 2 \in s, \ldots, N \in s\} \\ &\quad = P\{1 \in s\} P\{2 \in s \mid 1 \in s\} \cdots P\{N \in s \mid 1 \in s, \ldots, N-1 \in s\}. \end{aligned}$$

Between the two draw-by-draw schemes described above, Scheme 3 is usually more efficient than Scheme 2 because in Scheme 3, each population unit is only considered once throughout the procedure, whereas in Scheme 2, each population unit is considered at almost every step of the procedure.

The last scheme we present is particularly designed for the case when samples of different sizes are required. Such an example can be found in Chen (1992), where the task is to pick out from an unstructured list of, say 200 observations, a subset of, say 15 to 25, positive true signals that rise above a certain level of random errors. To serve this purpose, one can first break the entire sample space into smaller subsample spaces, each with a fixed sample size, and then use any of the previously described schemes to draw samples from a selected subsample space.

SCHEME 4. The sampling is done in two stages.

Stage 1. Determine the sample size k from $\{|s|\colon s \in \Omega\}$ according to the probability $\sum_{|s|=k} p(s)$.

Stage 2. Use any of Schemes 1–3 to draw a sample s_0 of size k from the subsample space $\{s\colon |s| = k, s \in \Omega\}$ with probability $p(s_0)/\sum_{s \in \Omega, |s|=k} p(s)$.

3. Weighted polynomial models. In this section, we introduce a new class of models called weighted polynomial models (in short, WPM) and show several statistical applications in which these models have been used.

It is quite general and realistic to think that the part any unit plays in a sampling design is determined by two factors: (1) whether this unit is in or out of the sample and (2) all available auxiliary information about this unit. The first factor is readily represented by s. As for the second factor, we assume that all auxiliary information can be summarized by a single variable, which we call the "weight". For examples, the "sizes" used in PPS may be interpreted as the weights of the population units [Chen, Dempster and Liu (1994)], and the odds ratios in case-control studies also play the role of the weights [Chen and Liu (1997)].

Denote the weights for the population units by $\mathbf{w} = (w_1, \ldots, w_N)$, where w_i is the weight for the ith unit, and the weights for the units in a sample $s = \{i_1, \ldots, i_n\}$ by $\mathbf{w}_s = (w_{i_1}, \ldots, w_{i_n})$. Without further specifications, we assume that the two factors s and $\mathbf{w}$ enter the probability density through a "link function" F. The class of the WPM is formally defined as follows.

DEFINITION 3. A probabilistic model is called a *weighted polynomial model*, or, in short, WPM, if there exists an n-variate function F such that for any $s \in \Omega$, $F(\mathbf{w}_s)$ is:

(i) proportional to $p(s)$, that is, $p(s) = F(\mathbf{w}_s)/\sum_{t \in \Omega} F(\mathbf{w}_t)$;
(ii) symmetric in the w_i for all $i \in s$;
(iii) a polynomial of the w_i for all $i \in s$.

The first assumption in Definition 3 requires that the probability of a sample depends only on the weights of the units in the sample, except for a normalizing constant. The second assumption is equivalent to claiming that the sample units are exchangeable, or equivalently, unordered. The third assumption can help make sampling from WPM efficient.

The following lemma shows that the specification of a WPM is unique up to a constant.

LEMMA 1. *For any WPM, the link function F is unique up to a constant.*

PROOF. Suppose there exist two link functions F_1 and F_2 that are not proportional to each other for at least one sample, say $t \in \Omega$. Since

$$p(t) = \frac{F_1(\mathbf{w}_t)}{\sum_{s \in \Omega} F_1(\mathbf{w}_s)} = \frac{F_2(\mathbf{w}_t)}{\sum_{s \in \Omega} F_2(\mathbf{w}_s)},$$

we get

$$\frac{F_1(\mathbf{w}_t)}{F_2(\mathbf{w}_t)} = \frac{\sum_{s \in \Omega} F_1(\mathbf{w}_s)}{\sum_{s \in \Omega} F_2(\mathbf{w}_s)}.$$

By the first assumption of Definition 3, the left-hand side of the equation above depends only on the weights of the units in the sample t, while the right-hand side depends also on the weights of the units that are not in t. This is obviously a contradiction. This completes the proof. □

DEFINITION 4. A WPM is called a k-degree WPM if the degree of its polynomial link function F is k.

It is easy to see by Lemma 1 that the degree of any given WPM is unique.

Not every model resulting from PPS [see, e.g., Hanif and Brewer (1980)] satisfies all three assumptions of the WPM. For example, Sampford's model

(1967) has the sampling design

$$p(s) = \Big(1 - \sum_{i \in s} w_i\Big) \prod_{i \in s} w_i \Big/ \prod_{i \in s} (1 - n w_i)$$
$$\propto \Big(1 - \sum_{i \in s} w_i\Big)\Big(\prod_{i \in s} w_i\Big)\Big[\prod_{i \in \mathscr{S} \setminus s} (1 - n w_i)\Big],$$

where $nw_i < 1$ for all $i \in \mathscr{S}$ and $\sum_{i=1}^{N} w_i = 1$. By Definition 3 and Lemma 1, the link function for a WPM can only depend on the weights in s and is unique up to a constant. One cannot find such a link function because the last term in the second expression depends also on the weights outside s.

Many existing weighted models that have been used in various research areas are in fact members of the WPM. The five examples we give next are by no means exhaustive.

EXAMPLE 1. Recently, a finite population model, *conditional Bernoulli model*, has been extensively studied and applied in various areas including survey sampling [Chen, Dempster and Liu (1994)], logistic regression and case-control studies [Chen and Liu (1996)] and lottery [Stern and Cover (1989)]. This model was first introduced by Stern and Cover (1989) as the *maximum entropy model*.

The conditional Bernoulli model is obtained by choosing a sampling design p to maximize the entropy $-\sum_{s \in \Omega} p(s) \log p(s)$ subject to the marginal constraints $\sum_{s \in \Omega, i \in s} p(s) = \pi_i$. The resulting density of the conditional Bernoulli model is

$$p(s) \propto F(\mathbf{w}_s) = \prod_{i \in s} w_i, \tag{1}$$

where the w_i need to satisfy the marginal constraints and can be found via an iterative procedure by Chen, Dempster and Liu (1994). Obviously, the conditional Bernoulli model is an n-degree WPM.

EXAMPLE 2. In sample surveys literature, Lahiri's model (1951) is used to obtain unbiased ratio-type estimates. Suppose y_i is the unknown characteristic associated with the ith unit, and the total $Y = \sum_{i \in \mathscr{S}} y_i$ is to be estimated. For all population units, the values of an auxiliary variable z_i are completely known and supposedly have a high positive correlation with the y_i. The ordinary ratio-type estimate of Y, based on a sample s, is $\hat{Y}_R = (\sum_{i \in s} y_i / \sum_{i \in s} z_i) \sum_{i \in \mathscr{S}} z_i$. When the sample is drawn with SRS, $\hat{Y}_R$ is biased. The idea of Lahiri's sampling model is that if the sample s is drawn with the probability proportional to $\sum_{i \in s} z_i$, $\hat{Y}_R$ is guaranteed to be unbiased. Treating the z_i as the weights w_i, the sampling design can be written as

$$p(s) \propto F(\mathbf{w}_s) = \sum_{i \in s} w_i, \tag{2}$$

which is a one-degree WPM. □

EXAMPLE 3. Given the same set-up as in Example 2, one can also use a linear regression estimate of Y given by $\overline{Y}_{lr} = N[\bar{y} + b(\overline{Z} - \bar{z})]$, where $\bar{y}$ is the sample mean for the y variable, $\overline{Z}$ and $\bar{z}$ are the population mean and the sample mean of the z variable, respectively, and

$$b = \frac{\sum_{i \in s}(y_i - \bar{y})(z_i - \bar{z})}{\sum_{i \in s}(z_i - \bar{z})^2}.$$

Like all ratio-type estimates, the linear regression estimate is biased if SRS is used to draw the sample. Singh and Srivastava (1980) proposed two sampling designs to obtain unbiased linear regression estimates. In the first sampling design, a sample is drawn with probability proportional to $\sum_{i \in s}(z_i - \bar{z})^2$, and the usual regression estimate $\bar{y}_{lr}$ becomes unbiased. Treating the z_i as the weights w_i, the sampling design can be expressed as (after some algebra),

$$p(s) \propto F(\mathbf{w}_s) = (n-1)\sum_{i \in s} w_i^2 - \sum_{i \in s}\sum_{j \in s, j \neq i} w_i w_j, \tag{3}$$

which is a two-degree WPM. Their second sampling design is also a two-degree WPM, though it does not make $\bar{y}_{lr}$ (but another regression estimate instead) unbiased. □

EXAMPLE 4. In fitting lottery data, Joe (1987) suggested various distance measures to generate probability models. A particular class of distance measures he considered is of the form $\sum_{s \in \Omega}\phi(p(s))$, where ϕ is strictly convex so that a minimum is guaranteed. Chen (1995) showed that if ϕ is invertible, minimizing Joe's distance measure subject to the marginal constraints $\sum_{s \in \Omega, i \in s} p(s) = \pi_i$ (where the π_i are observed frequencies of the lottery numbers) yields

$$p(s) \propto \left[h\left(\sum_{i \in s} w_i\right)\right]_+, \tag{4}$$

where $h = [\phi']^{-1}$, and the notation $[y]_+ = \max\{0, y\}$. The parameters w_i can be determined by the marginal constraints. For convenience, we only consider the cases where $h(\mathbf{w}_s)$ is always nonnegative for any s so that the subscript "+" in (4) can be ignored.

It can be easily checked that the model in (4) satisfies the first two assumptions of Definition 3 for any h; and when h is an m-degree polynomial, the model becomes an m-degree WPM.

When fitting the Canada's Lotto 6/49 data, Joe (1987) chose the convex functions $\phi_\alpha(u) = (u^{1+\alpha} - u)/\alpha$, $0 \leq \alpha \leq 1$, where the limit $\phi_0(u) = u \log u$ is obtained with $\alpha = 0$. This results in the probability model

$$p(s) \propto F(\mathbf{w}_s) = \left(\sum_{i \in s} w_i\right)^{1/\alpha} \tag{5}$$

for $\alpha > 0$, and the conditional Bernoulli model in (1) for $\alpha = 0$. Notice that when $\alpha = 1$, the model in (5) becomes Lahiri's model in (2). In general, when $\alpha = 1/m$ with m being any positive integer, the model in (5) becomes an m-degree WPM.

EXAMPLE 5. Liu, Neuwald and Lawrence (1995) used a Bayesian missing-data methodology for multiple local DNA sequence alignment. To sample from the posterior distribution, they adopted a Gibbs sampler, part of which is performed through what they call a doubly proportional sampling chain. The sampling procedure can be best illustrated by urn models as follows.

At each step of the procedure, the urn contains n balls and $N - n$ balls are outside. At the next step, a ball, say i, is drawn from the n balls in the urn with probability proportional to w_i^τ and is taken out of the urn; then a ball, say j, is picked from the $N - n$ balls outside the urn with probability proportional to w_j^β and is put into the urn. This process is repeated until the distribution of the n balls inside the urn converges. Liu, Neuwald and Lawrence (1995) showed that this procedure produces a reversible Markov chain with the equilibrium distribution,

$$p(s) \propto F(\mathbf{w}_s) = \Big(\prod_{i \in s} w_i^{\beta - \tau}\Big)\Big(\sum_{i \in s} w_i^\tau\Big). \tag{6}$$

It is easy to see from above that, when $\tau = 0$ and $\beta = 1$, their procedure converges to a conditional Bernoulli model in (1) and, when $\tau = \beta = 1$, to Lahiri's model in (2). In general, whenever τ, β are integers and $(\beta - \tau)\tau \geq 0$, the equilibrium distribution is a $[n\beta - (n - 1)\tau]$-degree WPM.

Using the techniques described in Section 4, sampling directly from the equilibrium distribution in (6) can be done exactly, rather than asymptotically as in Liu, Neuwald and Lawrence (1995). They did not use an exact sampling scheme because the sampling chain is only a part of the Gibbs sampler and accurate sampling at that stage is not necessary. Notice that their method is similar to the Metropolis–Hasting algorithm, and the population is "monotone." As will be shown in Section 5, the average acceptance rate of their scheme can be explicitly expressed in closed form as a function of the inclusion probabilities.

Last, we would like to point out that the WPM also provides a convenient platform for building hierarchical models. In particular, models on the lower level can be built directly upon the weights w_i regardless of the form of the WPM. For example, we can specify a generalized linear model for the w_i as follows:

$$w_i = g\big(\mathbf{z}_i^T \boldsymbol{\beta}\big) \tag{7}$$

where $\mathbf{z}_i$ is the covariate vector for the ith unit, which supposedly contains all auxiliary information, and $\boldsymbol{\beta}$ is the parameter vector. Thus, complex analysis such as Bayesian inference [treating (7) as the prior distribution for the w_i] is readily applicable to the WPM.

4. Exact sampling from weighted polynomial models. In this section, we discuss the issue of drawing random samples from WPM using exact weighted sampling schemes and provide general strategies that can help improve the efficiency of the schemes. As pointed out in Section 2, Scheme 3 is the most efficient among the four exact weighted sampling schemes. Therefore we will focus on the use of Scheme 3 in sampling from WPM.

Define the "normalizing functions" as follows:

$$L(B, D) = \sum_{s \in \Omega,\, B \subset s,\, (D \setminus B) \subset (\mathscr{S} \setminus s)} F(\mathbf{w}_s)$$

for any $B \subset D \subset \mathscr{S}$ and $|B| \le n$. The summation in the definition above is over the samples that include the units in B but exclude the units in $D \setminus B$. Then the selection probability at the kth step in Scheme 3 can be expressed as

$$P_2(k \mid s_{k-1}) = \sum_{\substack{s \in \Omega,\, s_{k-1} \subset s,\, k \in s \\ (A_{k-1} \setminus s_{k-1}) \subset (\mathscr{S} \setminus s),}} F(\mathbf{w}_s) \Bigg/ \sum_{\substack{s \in \Omega,\, s_{k-1} \subset s \\ (A_{k-1} \setminus s_{k-1}) \subset (\mathscr{S} \setminus s)}} F(\mathbf{w}_s)$$

$$= \frac{L(s_{k-1} \cup \{k\}, A_k)}{L(s_{k-1}, A_{k-1})}.$$

In particular, $L(\varnothing, \varnothing) = \sum_{s \in \Omega} F(\mathbf{w}_s)$ is the normalizing constant for the sampling design p.

At the first step of Scheme 3, we need to compute $L(\varnothing, \varnothing)$ and $L(\{1\}, \{1\})$. At each subsequent step k ($k \ge 2$), we only need to compute $L(s_{k-1} \cup \{k\}, A_k)$ for $P_2(k \mid s_{k-1})$ because $L(s_{k-1}, A_{k-1})$ is the same as the numerator in either $P_2(k-1 \mid s_{k-2})$ or $1 - P_2(k-1 \mid s_{k-2})$. Therefore we only need to compute a maximum of $N + 1$ normalizing functions for the entire procedure.

The following are two useful properties of the normalizing functions.

LEMMA 2. *For any $B \subset D \subset \mathscr{S}$ and $|B| \le n$, $L(B, D)$ is symmetric*:

(i) *in the w_i for all $i \in D^c$*;
(ii) *in the w_i for all $i \in B$.*

PROOF. (i) Let $L(B, D)_{j \mid i}$ denote the same polynomial as $L(B, D)$ except that w_i is exchanged with w_j for distinct $i, j \in D^c$. Let $s_{j \mid i}$ denote the same sample as s except that unit i is exchanged with unit j if $i, j \in s$, and unit i is replaced by unit j if $i \in s$, $j \notin s$. Let $\Omega(B, D) = \{s\colon s \in \Omega, B \subset s, (D \setminus B) \subset (\mathscr{S} \setminus s)\}$. Then, for any distinct units $i, j \in D^c$,

$$L(B, D)_{j \mid i} = \sum_{\substack{s \in \Omega(B, D) \\ i, j \in s}} F(\mathbf{w}_{s_{j \mid i}}) + \sum_{\substack{s \in \Omega(B, D) \\ i \in s,\, j \notin s}} F(\mathbf{w}_{s_{j \mid i}})$$

$$+ \sum_{\substack{s \in \Omega(B, D) \\ i \notin s,\, j \in s}} F(\mathbf{w}_{s_{i \mid j}}) + \sum_{\substack{s \in \Omega(B, D) \\ i \notin s,\, j \notin s}} F(\mathbf{w}_s)$$

$$= \sum_{\substack{s \in \Omega(B,D) \\ i,j \in s}} F(\mathbf{w}_s) + \sum_{\substack{s \in \Omega(B,D) \\ i \notin s,\, j \in s}} F(\mathbf{w}_s)$$
$$+ \sum_{\substack{s \in \Omega(B,D) \\ i \in s,\, j \notin s}} F(\mathbf{w}_s) + \sum_{\substack{s \in \Omega(B,D) \\ i \notin s,\, j \notin s}} F(\mathbf{w}_s)$$
$$= L(B,D).$$

(ii) The proof is similar to that for (i). □

In order to best facilitate Scheme 3, we need to find an efficient way to calculate the normalizing functions. The idea is to express the L functions in terms of the R functions (also known as the "elementary symmetric functions" in polynomial theory) defined as follows:

$$R(k,C) = \sum_{B \subset C,\, |B|=k} \left(\prod_{i \in B} w_i \right)$$

for any nonempty set $C \subset \mathscr{S}$ and $1 \le k \le |C|$, $R(0,C) = 1$, and $R(k,C) = 0$ for any $k > |C|$.

Computing $R(k,C)$ by definition involves $|C|!/[(k-1)!(|C|-k)!]$ multiplications and additions, which becomes intractable even when k and $|C|$ are moderately large. The following recursive formula known as "Newton's identity" helps reduce the computational complexity of $R(k,C)$ to only $O(k|C|)$ operations [Chen and Liu (1997)]:

$$R(k,C) = \frac{1}{k} \sum_{i=1}^{k} (-1)^{i+1} T(i,C) R(k-i,C), \tag{8}$$

where $T(i,C) = \sum_{j \in C} w_j^i$ for any $i \ge 1$ and $C \subset \mathscr{S}$. The formula (8) is in fact a natural generalization of the well-known "inclusion-exclusion" formula

$$\binom{c-1}{k-1} = \binom{c}{k-1} - \binom{c}{k-2} + \cdots + (-1)^{k+1} \binom{c}{0}.$$

The following is a well-known result in polynomial theory, often referred to as the "fundamental theorem on symmetric functions" [e.g., Uspensky (1948)].

THEOREM 1. *Any k-degree ($k \ge 1$) symmetric polynomial of the variables $w_1, \ldots, w_N$ can be expressed as a polynomial in the elementary symmetric functions $R(1,\mathscr{S}), \ldots, R(k,\mathscr{S})$. Furthermore, coefficients of the latter polynomial are built up by additions and subtractions of the coefficients of the former symmetric polynomial.*

The following result follows immediately from Lemma 2 and Theorem 1.

COROLLARY 1. *For any $B \subset D \subset \mathscr{S}$ and $|B| \le n$, the normalizing function $L(B,D)$ can be expressed as a polynomial G in $R(1,D^c), \ldots, R(n,D^c)$, and $R(1,B), \ldots, R(|B|,B)$.*

Since each L function can be expressed by a polynomial of the R functions and the R functions can be computed efficiently by (8), we should expect Scheme 3 to be quite efficient for the WPM.

THEOREM 2. *For any given WPM, if the number of nonzero coefficients in the polynomial expansion G of each normalizing function does not depend on N or n, then sampling from this WPM using Scheme* 3 *requires* $O(nN^2)$ *operations.*

PROOF. As pointed out earlier, we only need to compute $L(s_{k-1} \cup \{k\}, A_k)$ at the kth step. By Corollary 1, $L(s_{k-1} \cup \{k\}, A_k)$ can be expressed as a polynomial G of $R(1, A_k^c), \ldots, R(n, A_k^c)$, and $R(1, s_k), \ldots, R(|s_k|, s_k)$. If the number of nonzero coefficients in G does not depend on N or n, then the number of operations needed to evaluate G is also independent of N and n, given the values of the R functions. Then most of the computing time will be spent on calculating the R functions. According to Chen and Liu (1997), it needs only about $2n(N-k)$ operations to get all $R(1, A_k^c), \ldots, R(n, A_k^c)$, and about $2|s_k|^2$ operations to get all $R(1, s_k), \ldots, R(|s_k|, s_k)$. Since Scheme 3 takes no more than N steps to complete, it requires $O(nN^2)$ operations in total to get all R functions needed for the entire procedure. □

COROLLARY 2. *If the degree of the link function F for a given WPM does not depend on N or n, then sampling from this WPM using Scheme* 3 *requires* $O(nN^2)$ *operations.*

PROOF. By polynomial theory, the number of terms in G is in fact equal to the number of ways in which the degree of F can be written as the sum of smaller integers. If the degree of F does not depend on N or n, nor does the number of terms in G, then the result follows immediately from Theorem 2. □

All of the sampling designs in the five examples described in Section 3 satisfy the condition in Theorem 2. For example, as will be shown next, the number of nonzero coefficients in G is 7 for Example 4 when $\alpha = 1/3$, and 3 for Example 5 when $\beta = 2$ and $\tau = 1$. However, only those sampling designs in Examples 2, 3 and 4 satisfy the condition in Corollary 2. Notice that the condition in Corollary 2 is stronger than that in Theorem 2. Thus the result in Theorem 2 is still applicable even if the degree of the link function does depend on the sample size n, as in the conditional Bernoulli model (1) and Liu, Neuwald and Lawrence (1995).

We now provide general strategies for transforming an L function into a polynomial G of the R functions. There are two cases.

Case 1. The condition in Corollary 2 is satisfied. In such cases, the number of all possible terms in the polynomial G does not depend on N or n. We can first identify all possible terms in G and then solve for their coefficients. If there are m terms in G, this should not take more than $O(m^2)$

operations—the same amount as required for solving m linear equations with m unknowns.

Case 2. The condition in Corollary 2 is not satisfied, but the condition in Theorem 2 is. In such cases, the number of nonzero coefficients in the polynomial G does not depend on N or n, but the number of all possible terms in G does. It is not so easy to identify the few terms with nonzero coefficients out of a large number of all possible terms. Therefore some type of "inclusion-exclusion" trick has to be played to obtain the expansion G of the R functions.

We illustrate the general strategies described above using one example from Section 3 for each case.

EXAMPLE 4 (Continued). Consider the case when $\alpha = 1/3$. The sampling design is

$$p(s) \propto F(\mathbf{w}_s) = \Big(\sum_{i\in s} w_i\Big)^3.$$

Obviously, the condition in Corollary 2 is satisfied since F has a degree of 3. At the kth step of Scheme 3, there are in total ten possible terms in G,

$$\begin{aligned}&\big[R(1,s_k)\big]^3,\ R(1,s_k)R(2,s_k),\ R(3,s_k),\ \big[R(1,s_k)\big]^2R(1,A_k^c),\\ &\quad R(2,s_k)R(1,A_k^c),R(1,s_k)\big[R(1,A_k^c)\big]^2,\ R(1,s_k)R(2,A_k^c),\\ &\quad \big[R(1,A_k^c)\big]^3,\ R(1,A_k^c)R(2,A_k^c),\ R(3,A_k^c).\end{aligned}$$

It is easy to check that the second, third and fifth terms have zero coefficients. The coefficients for the other seven terms will be determined subsequently. To simplify our derivation, define

$$l_m(k,C) = \sum_{B\subset C,|B|=k}\Big(\sum_{i\in B} w_i\Big)^m$$

for any nonempty set $C\subset\mathscr{S}$, $1\le k\le|C|$, and $m = 1,2,3$. Just as the R functions, $l_m(k,C)$ is 1 when $k=0$, and 0 for any $k>|C|$.

By Corollary 1, the normalizing function at the first step can be written as

$$l_3(n,\mathscr{S}) = L(\varnothing,\varnothing) = a\big[R(1,\mathscr{S})\big]^3 + bR(1,\mathscr{S})R(2,\mathscr{S}) + cR(3,\mathscr{S}). \tag{9}$$

Since $l_3(n,\mathscr{S})$ is a homogeneous polynomial with degree 3, only three coefficients in (9) are nonzero. We now determine the three coefficients in (9).

Setting $w_1 = 1, w_2 = \cdots = w_N = 0$, we get

$$1^3\times\binom{N-1}{n-1} = a\times 1^3 + b\times 0\times 0 + c\times 0.$$

So $a = \binom{N-1}{n-1}$. Setting $w_1 = w_2 = 1, w_3 = \cdots = w_N = 0$, we get

$$2^3\times\binom{N-2}{n-2} + \binom{2}{1}\times 1^3\times\binom{N-2}{n-1} = a\times 2^3 + b\times 2\times 1 + c\times 0.$$

So $b = -3\binom{N-2}{n-1}$. Setting $w_1 = w_2 = w_3 = 1$, $w_4 = \cdots = w_N = 0$, we get

$$3^3 \times \binom{N-3}{n-3} + \binom{3}{2} \times 2^3 \times \binom{N-3}{n-2} + \binom{3}{1} \times 1^3 \times \binom{N-3}{n-1} = a \times 3^3 + b \times 3 \times 3 + c \times 1.$$

So $c = 3((N-2n)/(N-n-1))\binom{N-3}{n-1}$. Thus (9) becomes

$$l_3(n,\mathscr{S}) = \binom{N-1}{n-1}[R(1,\mathscr{S})]^3 - 3\binom{N-2}{n-1}R(1,\mathscr{S})R(2,\mathscr{S}) + 3\frac{N-2n}{N-n-1}\binom{N-3}{n-1}R(3,\mathscr{S}). \tag{10}$$

Unit 1 is considered at the first step. We include this unit with probability

$$P_2(1 \mid \varnothing) = \frac{L(\{1\},\{1\})}{L(\varnothing,\varnothing)} = \sum_{s\in\Omega,\{1\}\subset s}\Big(\sum_{i\in s} w_i\Big)^3 \Big/ l_3(n,\mathscr{S}).$$

We can expand each term in the summation by

$$\Big(\sum_{i\in s} w_i\Big)^3 = w_1^3 + 3w_1^2\Big(\sum_{i\in s, i\neq 1} w_i\Big) + 3w_1\Big(\sum_{i\in s, i\neq 1} w_i\Big)^2 + \Big(\sum_{i\in s, i\neq 1} w_i\Big)^3.$$

Thus the numerator in $P_2(1 \mid \varnothing)$ becomes

$$\binom{N-1}{n-1}w_1^3 + 3w_1^2 l_1\big(n-1,\{1\}^c\big) + 3w_1 l_2\big(n-1,\{1\}^c\big) + l_3\big(n-1,\{1\}^c\big),$$

where l_1 and l_2 can be expressed as (obtained in a similar way as for l_3),

$$l_1(n,\mathscr{S}) = \binom{N-1}{n-1}R(1,\mathscr{S}) \quad \text{and} \quad l_2(n,\mathscr{S}) = \binom{N-1}{n-1}[R(1,\mathscr{S})]^2 - 2\binom{N-2}{n-1}R(2,\mathscr{S}). \tag{11}$$

If unit 1 is selected in the first step, then at the second step, we include unit 2 with probability

$$P_2(2 \mid \{1\}) = \frac{L(\{1,2\},\{1,2\})}{L(\{1\},\{1\})} = \sum_{s\in\Omega,\{1,2\}\subset s}\Big(\sum_{i\in s} w_i\Big)^3 \Big/ \sum_{s\in\Omega,\{1\}\subset s}\Big(\sum_{i\in s} w_i\Big)^3.$$

The denominator in $P_2(2 \mid \{1\})$ is the same as the numerator in $P_2(1 \mid \varnothing)$, and the numerator of $P_2(2 \mid \{1\})$ can be expanded in a similar fashion as for $P_2(1 \mid \varnothing)$,

$$\binom{N-2}{n-2}(w_1+w_2)^3 + 3(w_1+w_2)^2 l_1\big(n-2,\{1,2\}^c\big) + 3(w_1+w_2)l_2\big(n-2,\{1,2\}^c\big) + l_3\big(n-2,\{1,2\}^c\big).$$

If unit 1 is not selected in the first step, then the inclusion probability for unit 2 becomes

$$P_2(2 \mid \varnothing) = \frac{L(\{2\},\{1,2\})}{L(\varnothing,\{1\})} = \sum_{s\in\Omega,\{2\}\subset s,\{1\}\subset\mathscr{S}\setminus s} \Big(\sum_{i\in s} w_i\Big)^3 \Big/ l_3\big(n,\{1\}^c\big),$$

where the numerator can be expanded into

$$\binom{N-2}{n-1} w_2^3 + 3w_2^2 l_1\big(n-1,\{1,2\}^c\big) + 3w_2 l_2\big(n-1,\{1,2\}^c\big) + l_3\big(n-1,\{1,2\}^c\big).$$

In general, the normalizing function at the kth step is

$$\begin{aligned}
&L(s_k, A_k)\\
&\quad= \binom{N-k}{n-|s_k|}\big[R(1,s_k)\big]^3 + 3\big[R(1,s_k)\big]^2 l_1(n-|s_k|, A_k^c)\\
&\qquad + 3R(1,s_k) l_2(n-|s_k|, A_k^c) + l_3(n-|s_k|, A_k^c)\\
&\quad= \binom{N-k}{n-|s_k|}\big[R(1,s_k)\big]^3 + 3\binom{N-k-1}{n-|s_k|-1}\big[R(1,s_k)\big]^2 R(1,A_k^c)\\
&\qquad + 3\binom{N-k-1}{n-|s_k|-1} R(1,s_k)\big[R(1,A_k^c)\big]^2\\
&\qquad - 6\binom{N-k-2}{n-|s_k|-1} R(1,s_k) R(2,A_k^c)\\
&\qquad + \binom{N-k-1}{n-|s_k|-1}\big[R(1,A_k^c)\big]^3 - 3\binom{N-k-2}{n-|s_k|-1} R(1,A_k^c) R(2,A_k^c)\\
&\qquad + 3\frac{N-k-2n+2|s_k|}{N-k-n+|s_k|-1}\binom{N-k-3}{n-|s_k|-1} R(3,A_k^c),
\end{aligned}$$

where the second expression is obtained from (10) and (11) with n and $\mathscr{S}$ replaced by $n-|s_k|$ and A_k^c, respectively. The second expression is derived to show that each normalizing function is indeed a polynomial of the R functions.

EXAMPLE 5 (Continued). Consider the case when $\beta = 2$ and $\tau = 1$. The sampling design is

$$p(s) \propto F(\mathbf{w}_s) = \Big(\prod_{i\in s} w_i\Big)\Big(\sum_{i\in s} w_i\Big).$$

The condition in Corollary 2 is not satisfied since F has a degree of $n+1$. Therefore we have to play some type of "inclusion-exclusion" trick to obtain

the polynomial G. For the kth step of Scheme 3, the normalizing function is

$$\begin{aligned}
&L(s_k, A_k)\\
&\quad= \sum_{\substack{s\in\Omega,\, s_k\subset s\\ (A_k\setminus s_k)\subset(\mathscr{S}\setminus s)}} \Big(\sum_{i\in s} w_i\Big)\Big(\sum_{i\in s} w_i\Big)\\
&\quad= \sum_{t\in A_k^c,\,|t|=n-|s_k|} \Big(\prod_{i\in s_k\cup t} w_i\Big)\Big(\sum_{i\in s_k\cup t} w_i\Big)\\
&\quad= \Big(\prod_{i\in s_k} w_i\Big) \sum_{t\in A_k^c,\,|t|=n-|s_k|} \Big(\prod_{i\in t} w_i\Big)\Big(\sum_{i\in s_k} w_i + \sum_{i\in A_k^c} w_i - \sum_{i\in A_k^c\setminus t} w_i\Big)\\
&\quad= R(|s_k|, s_k)\big[R(1, s_k)R(n-|s_k|, A_k^c) + R(1, A_k^c)R(n-|s_k|, A_k^c)\\
&\qquad\qquad -(n-|s_k|+1)R(n-|s_k|+1, A_k^c)\big],
\end{aligned}$$

which is indeed a polynomial of the R functions.

Finally, we show that under certain circumstances, an even greater efficiency than $O(nN^2)$ can be achieved. For any $1 \le k \le n$, the link function F of a WPM is called k-dependent if each term in F involves at most k distinct w_i, and at least one term involves exactly k distinct w_i. If a link function is k-dependent, we can first use a properly designed weighted sampling scheme to draw a subsample of size k and then draw the rest of the sample using SRS. Since SRS is much more efficient than any weighted sampling scheme, we can expect to reduce computational cost significantly. To construct such a scheme, we need to decompose the link function into a series of functions, each involving at most k distinct w_i.

LEMMA 3. *If the link function F of a WPM is k-dependent where $1 \le k \le n-1$, then there exists a k-variate function, f, such that $F(\mathbf{w}_s) = \sum_{t\subset s,\,|t|=k} f(\mathbf{w}_t)$.*

PROOF. For any $1 \le j \le k$, consider the terms in $F(\mathbf{w}_s)$ that involve exactly j distinct w_i. Arbitrarily pick a subsample t_j^* from s, where the subscript j denotes the size of the subset. Denote all terms in $F(\mathbf{w}_s)$ that involve exactly those units in t_j^* by $f_j(\mathbf{w}_{t_j^*})$. Then the terms in $F(\mathbf{w}_s)$ that involve exactly j distinct w_i can be written as $\sum_{t_j\subset s} f_j(\mathbf{w}_{t_j})$ due to the symmetry of all $w_i \in s$. Thus the link function can be written as

$$F(\mathbf{w}_s) = \sum_{j=1}^{k}\Big[\sum_{t_j\subset s} f_j(\mathbf{w}_{t_j})\Big]. \tag{12}$$

We now need to transform each $\sum_{t_j\subset s} f_j(\mathbf{w}_{t_j})$ into a function, say $\sum_{t_k\subset s} f_j^*(\mathbf{w}_{t_k})$, that sums over all subsamples of size k. We know that there are $\binom{n}{j} f_j(\mathbf{w}_{t_j})$'s, each involving exactly j w_i's. On the other hand, each of the $\binom{n}{k} f_j^*(\mathbf{w}_{t_k})$'s has

$\binom{k}{j}$ terms, and each term involves exactly j w_i's. Thus,

$$\binom{n}{k}\binom{k}{j}\sum_{t_j\subset s} f_j(\mathbf{w}_{t_j}) = \binom{n}{j}\sum_{t_k\subset s} f_j^*(\mathbf{w}_{t_k}). \tag{13}$$

Plug (13) into (12),

$$F(\mathbf{w}_s) = \sum_{j=1}^{k} \frac{\binom{n}{j}}{\binom{n}{k}\binom{k}{j}} \sum_{t_k\subset s} f_j^*(\mathbf{w}_{t_k}) = \sum_{t\subset s,\,|t|=k} \left[\sum_{j=1}^{k} \frac{\binom{n}{j}}{\binom{n}{k}\binom{k}{j}} f_j^*(\mathbf{w}_t)\right].$$

We can take everything inside "[ ]" as $f(\mathbf{w}_t)$ and thus complete the proof. □

Notice that the proof of Lemma 3 also provides a procedure for obtaining the decomposition of a link function. The following is an example.

EXAMPLE 4 (Continued). Letting $\alpha = 1/3$ in Example 4 gives a three-dependent link function. Following the procedure described in the proof of Lemma 3, we get, for any $t_3 \subset s$,

$$f_1^*(\mathbf{w}_{t_3}) = \sum_{i\in t_3} w_i^3, \qquad f_2^*(\mathbf{w}_{t_3}) = 3\sum_{i,\,j\in t_3} w_i^2 w_j \quad \text{and} \quad f_3^*(\mathbf{w}_{t_3}) = 6\prod_{i\in t_3} w_i.$$

Thus for any $t \subset s, |t| = 3$, we have

$$\begin{aligned} f(\mathbf{w}_t) &= \sum_{j=1}^{3} \frac{\binom{n}{j}}{\binom{n}{3}\binom{3}{j}} f_j^*(\mathbf{w}_t) \\ &= \frac{2}{(n-1)(n-2)}\sum_{i\in t} w_i^3 + \frac{3}{n-2}\sum_{i,\,j\in t} w_i^2 w_j + 6\prod_{i\in t} w_i. \end{aligned}$$

Clearly, each term in the decomposition above involves only three weights.

We now use the result of Lemma 3 to construct the following sampling scheme.

SCHEME 5. Suppose the link function F is k-dependent with the decomposition $\Sigma_{t\subset s,\,|t|=k} f(\mathbf{w}_t)$.

Stage 1. Use Scheme 1 or 2 to draw k units, $i_1, \ldots, i_k$, from Ω with probability proportional to $f(w_{i_1}, \ldots, w_{i_k})$.

Stage 2. Draw $n - k$ units, $i_{k+1}, \ldots, i_n$, with equal probability. Then the units $i_1, \ldots, i_n$ form a sample from p.

It is easy to see that Scheme 5 will correctly generate a random sample from p because

$$\begin{aligned} p(s) &= \sum_{t\subset s,\,|t|=k} P\{\text{Stage 1} = t\}P\{\text{Stage 2} = s\setminus t \mid \text{Stage 1} = t\} \\ &\propto \sum_{t\subset s,\,|t|=k} f(\mathbf{w}_t) = F(\mathbf{w}_s). \end{aligned}$$

The sampling schemes used in Lahiri (1951) and Singh and Srivastava (1980) are both special cases of Scheme 5, being one-dependent and two-dependent, respectively. The first stage of their sampling schemes are "draw one unit i with probability proportional to w_i" (Lahiri) and "draw two units i and j with probability proportional to $(w_i - w_j)^2$" (Singh and Srivastava).

For Lahiri's model, Schemes 2 and 3 are the same at Stage 1. Singh and Srivastava, however, used Scheme 2 at Stage 1, instead of the more efficient Scheme 3. Obviously, Scheme 5 cannot be applied to the conditional Bernoulli model or the model in Liu, Neuwald and Lawrence (1995) since both models are at least n-dependent.

5. Approximate weighted sampling scheme. For a general finite population model that is not of the WPM type, exact sampling schemes may not benefit from the efficiency-improving strategies described in Section 4 and can quickly become intractable as the population size and/or sample size get moderately large. In such situations, an approximate sampling scheme is often desired. The approximate sampling scheme we propose in this section is a direct application of the Metropolis–Hasting (in short, M–H) algorithm. See, for example, Smith and Roberts (1993) for a good review on this algorithm.

The M–H algorithm is essentially a Markov chain process. In the context of finite population, the algorithm will start from one of the possible samples and keep transiting the current sample to a new sample. The transition probabilities are designed in such a way that the process will eventually converge to the correct sampling design p.

The transition at each step involves two stages: "proposal" and "decision." First comes the "proposal" stage, in which a prospective sample, say s^*, is selected using any conditional distribution $T(s^* \mid s)$, where s is the current sample. Second is the "decision" stage, in which the transition from s to s^* is accepted with probability $\min\{1, T(s \mid s^*)p(s^*)/[T(s^* \mid s)p(s)]\}$. If the transition is rejected, the new sample will be the same as the current sample. Notice that the choice for the acceptance probability at the "decision" stage is not unique. Nevertheless, the above choice is convenient for the finite population setting, especially with fixed sample size. The advantage of having two stages in the M–H algorithm is that we are free to choose any distribution T at the first stage, and with a good choice of T, a relatively fast convergence can be achieved.

For simplicity, we will choose the uniform distribution for T. A prospective sample is constructed by swapping a number of units in the current sample with the same number of units from outside. Let k denote the number of units swapped at each step. Then k can be any integer from 1 up to $n - 1$. We give the general "swap-k" scheme as follows.

SCHEME 6 (swap-k). Repeat the following two stages at each step of the M–H algorithm.

Proposal. Draw k units $i_1, \ldots, i_k$ from the current sample s with uniform probabilities, and draw k units $j_1, \ldots, j_k$ from $\mathscr{S} \setminus s$ with uniform probabili-

ties. Let $s^* = [s \cup \{j_1, \ldots, j_k\}] \setminus \{i_1, \ldots, i_k\}$, that is, the same as s except that the i's are replaced by the j's.

Decision. Accept the transition from s to s^* with probability

$$H(s^* \mid s) = \begin{cases} 1, & \text{if } p(s^*) \geq p(s), \\ p(s^*)/p(s), & \text{if } p(s^*) < p(s). \end{cases}$$

If the transition is accepted, take s^* as the new sample; otherwise take s as the new sample.

One way to estimate the speed of convergence of the M–H algorithm is to look at the expected acceptance rate. Intuitively, for a given sampling design, the higher the acceptance rate is, the larger portion of the sample space the M–H algorithm runs through, and therefore the faster we expect the algorithm to converge. However, there is no simple theoretical result on the relation between the acceptance rate and the convergence rate of the general M–H algorithm. Thus the use of acceptance rate should be cautioned. In particular, the acceptance rates for two different sampling designs should not be compared. For example, a sampling design dominated by a few samples with extremely large probabilities compared to the rest of the samples may have an acceptance rate near zero, but it only takes a few steps for the M–H algorithm to converge; whereas the M–H scheme for a uniform design will take a lot more steps to converge, even with an acceptance rate one.

In general, there is no simple closed form formula for evaluating the average acceptance rate of the general M–H algorithm. For finite populations, however, we will show that under mild conditions, the average acceptance rate under a swap-k scheme can be expressed in closed form using only the kth-order inclusion probabilities.

As in Section 4, we attach a subscript, say k $(1 \leq k \leq n-1)$, to a sample, say s, to denote a subset s_k of $\mathscr{S}$ with the size k. Because the size $k < n$, we will call s_k a "subsample" of $\mathscr{S}$.

DEFINITION 5. Suppose s_k and s_k^* are two disjoint subsamples of $\mathscr{S}$ $(1 \leq k \leq n-1)$. The subsample s_k is said to be "never smaller" than s_k^* (or equivalently s_k^* is "never larger" than s_k) if $p(s_k \cup s'_{n-k}) \geq p(s_k^* \cup s'_{n-k})$ for all possible $s'_{n-k} \subset \mathscr{S}$ with $s_k \cap s'_{n-k} = s_k^* \cap s'_{n-k} = \varnothing$. This relation is denoted by $s_k \succcurlyeq s_k^*$ (or equivalently, $s_k^* \preccurlyeq s_k$).

DEFINITION 6. A finite population is said to be "kth-order monotone" if for all possible disjoint $s_k, s_k^* \subset \mathscr{S}$, either $s_k \succcurlyeq s_k^*$ or $s_k \preccurlyeq s_k^*$.

DEFINITION 7. The kth-order inclusion probability for the subsample s_k is defined as $\pi(s_k) = \sum_{s \in \Omega,\, s_k \subset s} p(s)$, that is, the total probabilities of the samples that contain the subsample s_k. In particular, the term "marginal probabilities" refers to the first-order inclusion probabilities.

THEOREM 3. *Suppose that under a sampling design p, the finite population is kth-order monotone. Without loss of generality, suppose all subsamples of size k are arranged as $s_k^{(1)} \preccurlyeq s_k^{(2)} \preccurlyeq \cdots \preccurlyeq s_k^{(c)}$, where $c = N!/[k!(N-k)!]$.*

Then the average acceptance rate of the swap-k scheme is given by

$$P\{\text{accept}\} = \left[\binom{n}{k}\binom{N-n}{k}\right]^{-1}\left[2\binom{N}{k}\binom{n}{k} - \frac{n!}{(n-2k)!} - 2\sum_{i=1}^{c} i\pi\left(s_k^{(i)}\right)\right] \tag{14}$$

where $(n-2k)! \stackrel{\text{def}}{=} 1$ *if* $k > n/2$.

PROOF. For any two samples $s^1, s^2 \in \Omega$, the difference between s^1 and s^2 is defined as $s^1 - s^2 = \{i\colon i \in s^1,\ i \notin s^2\}$. Note that $s^1 - s^2$ and $s^2 - s^1$ are always disjoint. Since the conditional distribution at the "proposal" stage is uniform, each prospective sample is selected with probability

$$T = T(s^* \mid s) = \left[\binom{n}{k}\binom{N-n}{k}\right]^{-1}.$$

Then by Scheme 6 and the definition $P\{\text{accept}\} = E_{s,s^*}[H(s^* \mid s)]$, we have

$$P\{\text{accept}\} = \sum_{s\in\Omega}\left[\sum_{\substack{s-s^*\subset\mathscr{S}\\ |s-s^*|=k}}\ \sum_{\substack{s^*-s\subset\mathscr{S}\\ |s^*-s|=k}} H(s^* \mid s)T(s^* \mid s)\right]p(s).$$

Let $t = s - s^*$, $t^* = s^* - s$ and $\Omega_{t\mid t^*} = \{s\colon t \subset s,\ t^* \subset \mathscr{S}\setminus s\}$. Interchange the three summations,

$$P\{\text{accept}\} = T\sum_{t\subset\mathscr{S},|t|=k}\ \sum_{\substack{t^*\subset\mathscr{S},|t^*|=k\\ t^*\cap t=\varnothing}}\left[\sum_{s\in\Omega_{t\mid t^*}} H(s\cup t^*\setminus t \mid s)\,p(s)\right].$$

Since t and t^* run through all subsamples of size k in Ω, they can be replaced by the ordered subsamples $s_k^{(i)}$. Let $\Omega_{i\mid j} = \{s\colon s_k^{(i)} \subset s,\ s_k^{(j)} \subset \mathscr{S}\setminus s\}$. We have

$$\begin{aligned}
&P\{\text{accept}\}\\
&= T\sum_{i=1}^{c}\sum_{j\neq i}\sum_{s\in\Omega_{i\mid j}} \min\left\{1, \frac{p\left(s\cup s_k^{(j)}\setminus s_k^{(i)}\right)}{p(s)}\right\}p(s)\\
&= T\sum_{i=1}^{c}\left[\sum_{j<i}\sum_{s\in\Omega_{i\mid j}} \frac{p\left(s\cup s_k^{(j)}\setminus s_k^{(i)}\right)}{p(s)}p(s) + \sum_{j>i}\sum_{s\in\Omega_{i\mid j}} p(s)\right]\\
&= T\sum_{i=1}^{c}\left\{\sum_{j<i}\left[\pi\left(s_k^{(j)}\right) - \pi\left(s_k^{(i)}\cup s_k^{(j)}\right)\right] + \sum_{j>i}\left[\pi\left(s_k^{(i)}\right) - \pi\left(s_k^{(j)}\cup s_k^{(i)}\right)\right]\right\} \qquad (15)\\
&= 2T\sum_{i=1}^{c}\sum_{j>i}\pi\left(s_k^{(i)}\right) - T\sum_{i=1}^{c}\sum_{j\neq i}\pi\left(s_k^{(j)}\cup s_k^{(i)}\right)\\
&= 2T\sum_{i=1}^{c}(c-i)\pi\left(s_k^{(i)}\right) - T(2k)!\sum_{s_{2k}\in\Omega}\pi(s_{2k})\\
&= 2T\left[\binom{N}{k}\binom{n}{k} - \sum_{i=1}^{c} i\pi\left(s_k^{(i)}\right)\right] - T(2k)!\binom{n}{2k},
\end{aligned}$$

where in (15) we assume $k \le n/2$; otherwise, (15) becomes

$$P\{\text{accept}\} = 2T \sum_{i=1}^{c} (c - i)\pi\left(s_k^{(i)}\right) - Tn! \sum_{s_n \in \Omega} \pi(s_n).$$

The result follows immediately. □

It is interesting to see that if the population is kth-order monotone and the swap-k scheme is used, the acceptance rate depends only on the kth-order inclusion probabilities, regardless of the inclusion probabilities of other orders.

From the proof of Theorem 3, we also find an interpretation for the formula in (14):

$$P\{\text{accept}\} = 2P\{s_k^* \succcurlyeq s_k\},$$

where s_k is the subsample drawn from the current sample, and s_k^* is the subsample drawn from the units outside the sample.

When the population is first-order monotone, the result in (14) simplifies to

$$P\{\text{accept}\} = \frac{1}{N-n}\left(2N - n + 1 - \frac{2}{n}\sum_{i=1}^{N} i\pi(i)\right), \tag{16}$$

where $\pi(i)$ is the marginal probability for the ith smallest unit.

The following lemma shows that under a mild condition, a WPM is first-order monotone, and hence the average acceptance rate of the corresponding swap-1 scheme can be explicitly evaluated from (16).

LEMMA 4. *The finite population is first-order monotone under a given WPM if and only if the link function $F(\mathbf{w}_s)$ is monotone in all w_i for $i \in s$.*

PROOF. By the second assumption in Definition 3, the population units are exchangeable. Thus we only need to prove the lemma for units 1 and 2. Without loss of generality, assume $w_1 \le w_2$. We have

F is always nondecreasing in the first argument

$$\begin{aligned}
&\Leftrightarrow F(w_1, w_{i_1}, \ldots, w_{i_{n-1}}) \\
&\quad \le F(w_2, w_{i_1}, \ldots, w_{i_{n-1}}) \text{ for any distinct } i_1, \ldots, i_{n-1} \in \mathscr{S} \setminus \{1, 2\} \\
&\Leftrightarrow \{1\} \preccurlyeq \{2\}.
\end{aligned}$$

This completes the proof. □

The result in Lemma 4 provides an easy way to check first-order monotonicity for any WPM by checking the sign of $\partial F(\mathbf{w}_s)/\partial w_i$ for all $i \in s$. In fact all sampling designs described in Section 3 except the one by Singh and Srivastava (1980) are kth-order monotone for any $1 \le k \le n - 1$, and therefore the average acceptance rate of any swap-k scheme can be evaluated for these models as in (14).

Acknowledgments. The author thanks Professor A. P. Dempster for insightful suggestions and Ms. Yonghong Mao for useful comments and technical support.

REFERENCES

CHEN, S. X. (1992). Metropolis algorithm and the nearly black object. Technical report, Dept. Statistics, Harvard Univ.

CHEN, S. X., DEMPSTER, A. P. and LIU, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81** 457–469.

CHEN, S. X. and LIU, J. S. (1997). Statistical applications of the Poisson-Binomial and conditional Bernoulli distributions. *Statist. Sinica* **7** 875–892.

HANIF, M. and BREWER, K. R. W. (1980). Sampling with unequal probabilities without replacement: a review. *Internat. Statist. Rev.* **48** 317–335.

JOE, H. (1990). A winning strategy for lotto games? *Canad. J. Statist.* **18** 233–244.

LAHIRI, D. B. (1951). A method for sample selection providing unbiased ratio estimates. *Bull. Internat. Statist. Inst.* **33** 133–140.

LIU, J. S., NEUWALD, A. F. and LAWRENCE C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc.* **90** 1156–1170.

SAMPFORD, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54** 499–513.

SINGH, P. and SRIVASTAVA, A. K. (1980). Sampling schemes providing unbiased regression estimators. *Biometrika* **67** 205–209.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. B* **55** 3–23.

STERN, H. and COVER, T. M. (1989). Maximum entropy and the lottery. *J. Amer. Statist. Assoc.* **84** 980–985.

USPENSKY, J. V. (1948). *Theory of Equations*. McGraw-Hill, New York.

STERN SCHOOL OF BUSINESS
NEW YORK UNIVERSITY
NEW YORK, NEW YORK 10012
E-MAIL: schen3@stern.nyu.edu