Weighted Finite Population Sampling to Maximize Entropy

Article *in* Biometrika · August 1994

DOI: 10.1093/biomet/81.3.457 · Source: OAI

3 authors, including:

Arthur Dempster
Harvard University
80 PUBLICATIONS **76,173** CITATIONS

SEE PROFILE

Jun S Liu
Harvard University
530 PUBLICATIONS **39,198** CITATIONS

SEE PROFILE

Biometrika (1994), **81**, 3, *pp.* 457–69
Printed in Great Britain

Weighted finite population sampling to maximize entropy

BY XIANG-HUI CHEN, ARTHUR P. DEMPSTER AND JUN S. LIU
Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.

SUMMARY

Attention is drawn to a method of sampling a finite population of N units with unequal probabilities and without replacement. The method was originally proposed by Stern & Cover (1989) as a model for lotteries. The method can be characterized as maximizing entropy given coverage probabilities π_i, or equivalently as having the probability of a selected sample proportional to the product of a set of 'weights' w_i. We show the essential uniqueness of the w_i given the π_i, and describe practical, geometrically convergent algorithms for computing the w_i from the π_i. We present two methods for stepwise selection of sampling units, and corresponding schemes for removal of units that can be used in connection with sample rotation. Inclusion probabilities of any order can be written explicitly in closed form. Second-order inclusion probabilities π_{ij} satisfy the condition $0 < \pi_{ij} < \pi_i \pi_j$, which guarantees Yates & Grundy's variance estimator to be unbiased, definable for all samples and always nonnegative for any sample size.

Some key words: Exponential family; Independent Bernoulli trials; Iterative proportional fitting; Maximum entropy; Rotatability; Sampling with unequal probabilities and without replacement; Survey sampling; Weighted sampling.

1. INTRODUCTION

Random sampling of n distinct units from a population of N units may be called weighted sampling when the probabilities associated with the $N!/\{(N-n)!n!\}$ possible choices are not all equal. A problem often considered in the literature, e.g. Hanif & Brewer (1980), Chaudhuri & Vos (1988, pp. 143–346), is to define a particular weighted sampling scheme subject to prespecification of the marginal probabilities π_i that the sample includes the ith population unit, where

$$0 < \pi_i < 1 \quad (i = 1, \ldots, N), \quad \sum_{i=1}^{N} \pi_i = n. \tag{1}$$

We propose and study a simple way to do this that appears to have been overlooked in sample survey literature.

We find it convenient to use indicator variables to represent samples. Thus the random sample may be denoted by X where

$$X = (X_1, \ldots, X_N)$$

and the random variable X_i takes the values 1 or 0 according as the ith unit is in or out of the sample, for $i = 1, \ldots, N$. Let

$$D^n = \{x = (x_1, \ldots, x_N) : x_i = 0 \text{ or } 1, \text{ and } x_1 + \ldots + x_N = n\}.$$

The random vector X takes values in D^n. We denote the probability density function of

a typical sampling scheme by $p(x)$ for any vector $x \in D^n$, where $p(x) > 0$ and $\sum p(x) = 1$. The associated probability that the sample includes the ith unit is

$$\pi_i = E(X_i) = \sum_{x \in D^n} x_i p(x), \tag{2}$$

where the π_i satisfy (1).

The particular family of sampling schemes that we propose can be defined in any of three ways that we show to be equivalent.

Method 1. Pick any vector of weights, $w = (w_1, \ldots, w_N)$, where $w_i > 0$ for $i = 1, \ldots, N$, and define

$$p(x) \propto \prod_{i=1}^{N} w_i^{x_i}. \tag{3}$$

It is obvious that rescaling the w_i by a positive constant multiplier determines the same $p(x)$ but that modulo rescaling the $p(x)$ corresponding to distinct w are distinct. It is less obvious that the coverage probabilities determined by (2) are in one–one correspondence with the weights w_i modulo scaling, but we show in § 2 that this correspondence is a direct consequence of standard exponential family theory. The notation $w_i = e^{\theta_i}$ makes the exponential family connection obvious because (3) can be rewritten as

$$p(x) \propto \exp\left(\sum_{i=1}^{N} \theta_i x_i\right)$$

in terms of the parameters $\theta = (\theta_1, \ldots, \theta_N)$ which are determined up to an additive constant.

Method 2. Pick any vector of coverage probabilities $\pi = (\pi_1, \ldots, \pi_N)$, where the π_i satisfy (1), and choose $p(x)$ subject to the constraints (2) to maximize the entropy $-\sum p(x) \log p(x)$.

It is easily proved that if a weight vector w or its corresponding exponent vector θ can be determined such that $p(x)$ defined by Method 1 matches the π given for Method 2, then this Method 1 choice is the unique maximum entropy scheme proposed as Method 2. A more general form of this result is proved by Darroch & Ratcliff (1972).

This model was first proposed by Stern & Cover (1989), and further generalized by Joe (1990), for determining optimal lottery strategies. We address in § 2 the existence and uniqueness of w, modulo scaling, given π.

The third characterization that we suggest is a trivial variant of Method 1.

Method 3. Pick any vector of probabilities $p = (p_1, \ldots, p_N)$, where $0 < p_i < 1$ for $i = 1, \ldots, N$, and define $Z = (Z_1, \ldots, Z_N)$ to be independent Bernoulli trials with probabilities $p_1, \ldots, p_N$. Then define the sampling distribution of X to be the conditional distribution of Z given $\sum Z_i = n$. It is evident that Method 3 gives the same sampling scheme as Method 1 if and only if the w_i are proportional to $p_i/(1 - p_i)$.

Thus the base model for all three methods described above is

$$p(x) = \prod_{i=1}^{N} w_i^{x_i} \Big/ \sum_{y \in D^n} \left(\prod_{i=1}^{N} w_i^{y_i}\right) \propto \exp\left(\sum_{i=1}^{N} \theta_i x_i\right) \quad (x \in D^n), \tag{4}$$

where w, or equivalently θ, may be determined by π through (2). We refer to (4) as the maximum entropy model.

In § 2, we show that, for any proper vector π satisfying (1), the corresponding w exists and is determined up to a scaling factor. Moreover, the vector w can be easily found via a modification of the iterative proportional fitting procedure (Deming & Stephan, 1940) which converges monotonically and geometrically with a rate bounded by $\max \pi_i$. Each iteration requires $O(n^2N)$ operations. We demonstrate several other properties of the mapping between w and π.

In § 3, we describe several procedures for selecting a sample of n distinct units from a population of N units under the maximum entropy model. Each procedure selects or removes population units one by one until a sample of n distinct units is obtained. The simplest of these procedures requires $O(nN)$ operations.

In § 4, we discuss the application of the maximum entropy model in survey sampling, considering in particular estimation of the population total $Y = \sum y_i$, where y_i is a characteristic associated with the ith individual in the population. We exhibit desirable properties of the Horvitz & Thompson (1952) estimator of Y guaranteed by the maximum entropy model. We also give two schemes for sample rotation.

In § 5, we give two numerical examples of simulation to illustrate the properties of the mapping between w and π.

The proofs for some important results are sketched in the Appendix. For details a technical report is available from X.-H. Chen upon request.

2. Relation between weights and coverage probabilities

The relation between w and π is a special case of that between natural and mean-value parameterizations for an exponential family. The following result can be proved by using Theorem 3.6 of Brown (1986, p. 74).

Theorem 1. *For any vector π satisfying* (1), *there exists a vector w for the maximum entropy model subject to the constraint* (2), *and w is unique up to rescaling.*

To compute w from π, we recast (2) in the form of a set of equations (5) below, and solve these iteratively as in (7). Throughout the paper we use the following notation: $S = \{1, \ldots, N\}$, capital letters such as A, B or C for subsets of S, $A^c = S \backslash A$ for the complement of A in S, and $|A|$ for the number of elements of A. Also

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

for any nonempty set $C \subset S$ and $1 \leqslant k \leqslant |C|$, $R(0, C) = 1$ and $R(k, C) = 0$ for any $k > |C|$. The following Proposition 1 follows immediately from the definitions.

Proposition 1. *For any nonempty set $C \subset S$ and $1 \leqslant k \leqslant |C|$:*

(*a*) $\sum_{j \in C} w_j R(k-1, C \backslash \{j\}) = kR(k, C)$,

(*b*) $\sum_{j \in C} R(k, C \backslash \{j\}) = (|C| - k) R(k, C)$,

(*c*) $\sum_{i=0}^{k} R(i, C) R(k-i, C^c) = R(k, C)$.

Using this notation, we may rewrite (2) as

$$\pi_i = \frac{w_i R(n-1, \{i\}^c)}{R(n, S)} \quad (i = 1, \ldots, N). \tag{5}$$

Note that an application of (*a*) in Proposition 1 proves that the right-hand sides of (5)

sum to n, consistent with the sum of the π_i as in (1). Thus for fixed n, there are $N-1$ linearly independent relations among the N relations of (5). Without loss of generality, we assume that $\pi_1 \leqslant \pi_2 \leqslant \ldots \leqslant \pi_N$ and let $w_N = \pi_N$. Dividing each of the first $N-1$ equations of (5) by the Nth equation on both sides and rearranging the terms for each equation, we get the following set of restricted equations,

$$w_i = \frac{\pi_i R(n-1, \{N\}^c)}{R(n-1, \{i\}^c)} \quad (i = 1, \ldots, N-1), \quad w_N = \pi_N. \tag{6}$$

Although a closed-form solution to (6) seems impossible, the equations can be solved as a fixed-point problem by using an iterative procedure. Specifically, the following updating scheme provides a solution of (6):

$$w_i^{(t+1)} = \left.\frac{\pi_i R(n-1, \{N\}^c)}{R(n-1, \{i\}^c)}\right|_{w=w^{(t)}} \quad (i = 1, \ldots, N-1), \quad w_N^{(t+1)} = w_N^{(t)} = \pi_N, \tag{7}$$

where $w^{(t)} = (w_1^{(t)}, \ldots, w_N^{(t)})$.

THEOREM 2. *Define* $\mathscr{W} = \{w : 0 < w_i \leqslant \pi_i, i = 1, \ldots, N-1; w_N = \pi_N\}$. *Then:*

(*a*) *the set of equations in* (6) *has a unique solution,* w^*, *in* $\mathscr{W}$;

(*b*) *starting from* $w^{(0)} = \pi$, *the sequence* (7) *of vectors* $w^{(t)}$ $(t = 1, 2, \ldots)$ *converges monotonically and geometrically to* w^* *with a rate bounded by* π_N.

Remark. The iterative procedure defined by (7) is a variant of the iterative proportional fitting algorithm first proposed by Deming & Stephan (1940). Imagine a 2^N contingency table initialized with a uniform distribution over the subset of cells such that n of the N binary variables take the value 1 and the remaining $N-n$ take the value 0. As originally proposed, the Deming–Stephan procedure would use the formula (7) for $i = 1, \ldots, N$ to define a cycle, but updating the table N times within each cycle, after each $w_i^{(t)}$ is modified into $w_i^{(t+1)}$, whereas we fix the normalization of the w_i by holding $w_N = \pi_N$ where π_N is the largest π_i, and hence we update the w_i in each cycle only for $i = 1, \ldots, N-1$. While Deming–Stephan is known to increase entropy monotonely at each step (Darroch & Ratcliff, 1972), further properties such as those in Theorem 2 do not apply to Deming–Stephan. In practice, the procedure in (7) takes far fewer iterations to converge than Deming–Stephan.

Our algorithm is also more direct and much faster than the one proposed by Stern & Cover (1989) which uses a generalized iterative scaling algorithm of Darroch & Ratcliff (1972).

LEMMA 1. *For any* $i, j \in S$, *the following properties hold:*

(*a*) $\pi_i = \pi_j \Leftrightarrow w_i = w_j$;

(*b*) $\pi_i > \pi_j \Leftrightarrow w_i > w_j$;

(*c*) $\pi_i > \pi_j \Leftrightarrow w_i/w_j > \pi_i/\pi_j$;

(*d*) *if* $c_1 \leqslant \pi_k \leqslant c_2$, *for all* π_k *and some* $c_1, c_2 \in (0, 1)$, *then* $w_i/w_j \to \pi_i/\pi_j$ *as* $N/n \to \infty$.

The naive approach to computing R by adding all the terms in its definition is inefficient and often prohibitively expensive. The following result enables us to calculate the function R recursively and relatively cheaply.

THEOREM 3. *Define* $T(i, C) = \sum_{j \in C} w_j^i$ *for any* $i \geqslant 1$ *and* $C \subset S$. *Then for any* $1 \leqslant k \leqslant |C|$,

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

When $w_i = 1$ for all i,

$$R(k, C) = \binom{|C|}{k},$$

and (8) gives the following combinatorial formula:

$$\binom{m}{k} = \frac{m}{k} \sum_{i=1}^{k} (-1)^{i+1} \binom{m}{k-i} = \frac{m}{k} \left\{ \binom{m}{k-1} - \binom{m}{k-2} + \ldots + (-1)^{k+1} \binom{m}{0} \right\}.$$

If $T(i, C)$ and $R(k-i, C)$ $(i = 1, \ldots, k)$ are all available, it requires $O(k)$ operations to calculate $R(k, C)$ using (8). By induction, if $T(i, C)$ $(i = 1, \ldots, k)$ are all available, it requires $O(k^2)$ operations to get $R(k, C)$. For each iteration modifying $w^{(t)}$ into $w^{(t+1)}$ in (7), we need to compute $R(n-1, \{j\}^c)$ $(j = 1, \ldots, N)$, which requires $O(n^2 N)$ operations if $T(i, \{j\}^c)$ $(i = 1, \ldots, n-1, j = 1, \ldots, N)$ are all available. However, it takes only $O(nN)$ operations to get all $T(i, \{j\}^c)$ using the formula $T(i, \{j\}^c) = T(i, S) - w_j^i$. In total, each iteration in (7) needs $O(n^2 N)$ operations.

3. Draw-by-draw selection procedures

In this section, we discuss procedures for drawing a sample from the maximum entropy model. Given the w_i, there is an explicit formula for each possible sample in D^n, but the large number, $N!/\{(N-n)!n!\}$, of choices renders impractical a naive approach of a single multinomial draw from $N!/\{(N-n)!n!\}$ cells. Therefore, we consider draw-by-draw selection procedures, where we draw one unit at a time until n units are obtained.

We call a selection procedure 'forward' if it selects n units from the population as the sample, or 'backward' if it removes $N-n$ units from the population and takes the remaining n units as the sample. Noticing that, for any $x \in D^n$,

$$p(x) \propto \prod_{i \in A_x} w_i \propto \prod_{i \in A_x} w_i \Big/ \prod_{i \in S} w_i = \prod_{i \in A_x^c} w_i^{-1},$$

where $A_x = \{i : x_i = 1\}$, it is obvious that for any 'forward' procedure, there is a corresponding 'backward' procedure, which selects unsampled units in the same way using w_i^{-1} instead of w_i.

We also distinguish among procedures by whether or not a procedure requires n fixed in advance. In the context of sample surveys, the π_i are usually prespecified. Thus the sample size n and the w_i are fixed in advance. However, it is possible in some applications that the w_i are prespecified and different sample sizes are to be experimented with. In this case, a selection procedure that does not depend on n is desired.

The output of a draw-by-draw procedure is represented by $A_0, A_1, \ldots, A_n$ where $A_0 = \varnothing$, and $A_k \subset S$ denotes the set of selected indices after k draws. The following are two 'forward' procedures, one for fixed n and the other for nonfixed n. The 'backward' version of these procedures can be defined accordingly.

Procedure 1 (*forward, n fixed*). At the kth draw $(k = 1, \ldots, n)$, a unit $j \in A_{k-1}^c$ is selected with probability

$$P_1(j, A_{k-1}^c) := \frac{w_j R(n-k, A_{k-1}^c \backslash \{j\})}{(n-k+1) R(n-k+1, A_{k-1}^c)}.$$

Using the relation $w_i \propto p_i/(1-p_i)$, the function R can be written as

$$R(k, C) = \text{pr}\,(\textstyle\sum_{t\in C} Z_t = k) \prod_{i\in C}(1+w_i)$$

for any nonempty set $C \subset S$ and $0 \leqslant k \leqslant |C|$. Thus P_1 has the interpretation

$$P_1(j, A_{k-1}^c) = \frac{1}{n-k+1}\,\text{pr}\,(Z_j = 1 \mid \textstyle\sum_{t\in A_{k-1}^c} Z_t = n-k+1),$$

where $Z_1, \ldots, Z_N$ are independent Bernoulli trials as defined in Method 3 in § 1.

It is easy to see by (*a*) in Proposition 1 that $P_1(., A_{k-1}^c)$ is a probability density on A_{k-1}^c. To see that a random sample of n units selected by Procedure 1 is a sample from the maximum entropy model, we first compute the probability of choosing an ordered set of indices $i_1, \ldots, i_n$ using Procedure 1, where in this case $A_k = \{i_1, \ldots, i_k\}$ for $k = 1, \ldots, n$:

$$\prod_{k=1}^{n} P_1(i_k, A_{k-1}^c) = \prod_{k=1}^{n} \frac{w_{i_k} R(n-k, A_k^c)}{(n-k+1)R(n-k+1, A_{k-1}^c)}$$

$$= \frac{1}{n!}\left(\prod_{k=1}^{n} w_{i_k}\right)\frac{R(0, A_n^c)}{R(n, S)} = \frac{1}{n!}\,\text{pr}\,(X_t = 1, t \in A_n).$$

Since there are $n!$ different ways of ordering the indices $i_1, \ldots, i_n$, the probability of obtaining the units $i_1, \ldots, i_n$ without regard to order is exactly $\text{pr}\,(X_t = 1, t \in A_n)$.

Suppose that $\pi_{i_1 \ldots i_k}$ is the kth-order inclusion probability for the units $i_1, \ldots, i_k$ to be in a sample of size n from the maximum entropy model. Then a property of Procedure 1 is that

$$\text{pr}\,(A_k = \{i_1, \ldots, i_k\}) \propto \left(\prod_{t=1}^{k} w_{i_t}\right)\frac{R(n-k, \{i_1, \ldots, i_k\}^c)}{R(n, S)} = \pi_{i_1 \ldots i_k}. \tag{9}$$

Although P_1 can be calculated directly from R functions, the computation can be much simplified by noticing that $P_1(j, A_0^c) = \pi_j/n$ and using the following formula recursively for the consecutive draws.

LEMMA 2. *For any* $1 \leqslant k \leqslant n-1$ *and* $j \in A_k^c$,

$$P_1(j, A_k^c) = \frac{w_{i_k} P_1(j, A_{k-1}^c) - w_j P_1(i_k, A_{k-1}^c)}{(n-k)(w_{i_k} - w_j)P_1(i_k, A_{k-1}^c)}. \tag{10}$$

Procedure 1 can be realized using the following algorithm, which requires $O(nN)$ operations.

1. For $j = 1, \ldots, N$, calculate $P_1(j, S)$, which is given by π_j/n. Then draw unit i_1 according to the probability $P_1(i_1, S)$.
2. If $n > 1$, then $A_0 \leftarrow \varnothing$, $A_1 \leftarrow \{i_1\}$, $k \leftarrow 2$, go to 3; otherwise stop.
3. For all $j \in A_{k-1}^c$, calculate $P_1(j, A_{k-1}^c)$ from $P_1(j, A_{k-2}^c)$ and $P_1(i_{k-1}, A_{k-2}^c)$ using (10). Then draw unit i_k according to the probability $P_1(i_k, A_{k-1}^c)$.
4. If $k < n$, then $A_k \leftarrow A_{k-1} \cup \{i_k\}$, $k \leftarrow k+1$, go to 3; otherwise stop.

Procedure 2 (*forward, n nonfixed*). At the kth draw ($k = 1, \ldots, n$), a unit $j \in A_{k-1}^c$ is selected with probability

$$P_2(j, A_{k-1}^c) := \sum_{i=0}^{k-1} \frac{w_j R(k-i-1, A_{k-1}^c \backslash \{j\}) R(i, A_{k-1})}{(k-i)R(k, S)}.$$

By (*a*) and (*c*) in Proposition 1,

$$\sum_{j\in A_{k-1}^c} P_2(j, A_{k-1}^c) = \sum_{i=0}^{k-1} \frac{R(i, A_{k-1})}{(k-i)R(k, S)} \left\{\sum_{j\in A_{k-1}^c} w_j R(k-i-1, A_{k-1}^c\backslash\{j\})\right\}$$

$$= \sum_{i=0}^{k-1} \frac{R(i, A_{k-1})}{(k-i)R(k, S)} (k-i)R(k-i, A_{k-1}^c) = 1.$$

Thus $P_2(., A_{k-1}^c)$ is a probability density on A_{k-1}^c. Now we show by induction that a random sample selected by Procedure 2 is a sample from the maximum entropy model. Let γ_k be the random index of the unit selected at the kth draw. Assuming

$$\text{pr}\,(A_{k-1} = A) = R(k-1, A)/R(k-1, S)$$

for any $A \subset S$ with $|A| = k-1$, which is true for $k = 2$ by the definition of P_2, we show that $\text{pr}\,(A_k = B) = R(k, B)/R(k, S)$ for any $B \subset S$ with $|B| = k$. By (*b*) and (*c*) in Proposition 1,

$$\text{pr}\,(A_k = B) = \sum_{j\in B} \text{pr}\,(A_{k-1} = B\backslash\{j\})\,\text{pr}\,(\gamma_k = j \mid A_{k-1} = B\backslash\{j\})$$

$$= \sum_{j\in B} \frac{R(k-1, B\backslash\{j\})}{R(k-1, S)} \left\{\sum_{i=0}^{k-1} \frac{w_j R(k-i-1, B^c)R(i, B\backslash\{j\})}{(k-i)R(k, S)}\right\}$$

$$= \frac{R(k, B)}{R(k, S)} \sum_{i=0}^{k-1} \frac{R(k-i-1, B^c)}{(k-i)R(k-1, S)} \left\{\sum_{j\in B} R(i, B\backslash\{j\})\right\}$$

$$= \frac{R(k, B)}{R(k, S)} \sum_{i=0}^{k-1} \frac{R(k-i-1, B^c)}{(k-i)R(k-1, S)} (k-i)R(i, B) = \frac{R(k, B)}{R(k, S)}.$$

By induction, the probability of obtaining a set A_n is

$$R(n, A_n)/R(n, S) = \text{pr}\,(X_t = 1, t \in A_n).$$

It is evident from the proof above that using Procedure 2

$$\text{pr}\,(A_k = \{i_1, \ldots, i_k\}) \propto \prod_{l=1}^{k} w_{i_l}.$$

Thus Procedure 2 does not depend on n.

The two procedures have different uses. By using (10), Procedure 1 requires less operations than Procedure 2, but cannot be used when n is nonfixed. Procedure 2 is useful for doing rotations in survey sampling. The preference between forward and backward procedures depends on the scale of n. Evidently, forward procedures are preferred when $n \leqslant N/2$ while backward procedures are preferred when $n > N/2$. When the w_i are all equal, both procedures reduce to simple random sampling without replacement.

4. Application to survey sampling

In the context of survey sampling, weighted sampling is often called sampling with unequal probabilities and without replacement. Its use in survey sampling was first suggested by Hansen & Hurwitz (1943). Hanif & Brewer (1980) reviewed and classified over 50 different such schemes. The goal is typically to estimate the population total $Y = \sum y_i$ for a finite population of N units, where y_i is a characteristic associated with the ith unit. A commonly-used unbiased estimator of Y as proposed by Horvitz & Thompson

(1952) is

$$\hat{Y} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} X_i,$$

where $X = (X_1, \ldots, X_N)$ represents a random sample of n distinct units drawn from the population. The variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2 \sum_{1 \leqslant i < j \leqslant N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j,$$

where π_{ij} is the second-order inclusion probability for both the units i, j to be in the sample. An alternative expression that is valid only when n is fixed is derived by Yates & Grundy (1953):

$$V(\hat{Y}) = \sum_{1 \leqslant i < j \leqslant N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \tag{11}$$

Yates & Grundy (1953) also suggest the following estimator of $V(\hat{Y})$, which is unbiased if $\pi_{ij} \neq 0$ for all pairs $i \neq j$ and for use when n is fixed:

$$v(\hat{Y}) = \sum_{1 \leqslant i < j \leqslant N} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 X_i X_j. \tag{12}$$

The maximum entropy model has the following properties.

Property 1. The inclusion probabilities of any order are uniquely determined by the π_i and can be explicitly expressed in closed form. For example,

$$\pi_{ij} = w_i w_j R(n-2, \{i, j\}^c)/R(n, S).$$

Thus $V(\hat{Y})$ and $v(\hat{Y})$ can be evaluated directly from the w_i. In general, the kth-order $(1 \leqslant k \leqslant n)$ inclusion probability for the units $i_1, \ldots, i_k$ to be in the sample is $\pi_{i_1 \ldots i_k}$ as given in (9). As we have shown in § 2, these inclusion probabilities can be easily calculated from the w_i.

Property 2. For the maximum entropy model, $0 < \pi_{ij} < \pi_i \pi_j$, for any pair $i \neq j$.

It is easy to see that the condition

$$\pi_{ij} \leqslant \pi_i \pi_j \quad (i \neq j) \tag{13}$$

is a sufficient condition for the Yates & Grundy's variance estimator $v(\hat{Y})$ to be always nonnegative. By some algebra, it can be shown that (13) is a sufficient condition for a sampling procedure without replacement to be more efficient, i.e. to have smaller variance, than sampling with replacement. Few plans have this nice property, especially when $n > 2$. For example, the procedure by Hartley & Rao (1962) satisfies (13) only when $n = 2$. Furthermore, for $v(\hat{Y})$ to be unbiased and definable for all samples, the condition that $\pi_{ij} > 0$ for any $i \neq j$ is also desirable.

Property 3. A direct consequence by applying Property 2 to (11) and (12) is as follows:

$$\pi_i \propto y_i \Leftrightarrow V(\hat{Y}) = 0 \Leftrightarrow v(\hat{Y}) = 0 \tag{14}$$

for all possible samples.

For an arbitrarily given sampling with unequal probabilities and without replacement scheme, $\pi_i \propto y_i$ is only sufficient for $V(\hat{Y}) = 0$ and $v(\hat{Y}) = 0$. Therefore, the resulting estimator of the variance of $\hat{Y}$, $v(\hat{Y})$, may be zero even when $\hat{Y}$ is clearly not a constant. The result in (14) implies, however, that the maximum entropy model displays the desirable property that $V(\hat{Y})$ or $v(\hat{Y})$ is zero if and only if the π_i are proportional to the y_i.

Property 4 (*rotatability*). In continuing surveys conducted at regular intervals or multistage sampling, it is often desired that the sampling units should not stay in the survey indefinitely, but rather that old units that have been in the survey for a specified period should be rotated, i.e. replaced, by new units. It is also desirable that the new units are selected in such a way that the new sample follows the same probability distribution as the old one. We find that doing rotations under the maximum entropy model is especially convenient. To see this, we present two schemes. Let A denote the set of the indices in the current sample.

SCHEME 1.

Step 1. Draw a unit $i \in A$ with uniform probability, and draw a unit $j \in A^c$ with uniform probability.

Step 2. Replace the unit i by the unit j with the acceptance probability

$$H(j|i) = \begin{cases} 1 & (w_j \geqslant w_i), \\ w_j/w_i & (w_j < w_i). \end{cases}$$

Step 3. If the replacement is accepted, take $A \cup \{j\} \backslash \{i\}$ as the new sample and stop; otherwise go back to Step 1.

Scheme 1 is in fact a Metropolis–Hasting-type algorithm (Smith & Roberts, 1993). When all the weights are equal, this algorithm reduces to the celebrated Bernoulli–Laplace model. It can be shown that the acceptance rate of Scheme 1 is

$$\text{pr (accept)} = \frac{1}{N-n}\left(2N - n + 1 - \frac{2}{n}\sum_{i=1}^{N} i\pi_i\right),$$

where the π_i are assumed to be in ascending order. Two extreme cases: pr (accept) = 1 when the π_i are all equal, and pr (accept) = 0 when the largest n π_i are 1 and the rest are 0.

SCHEME 2.

Step 1. Select a unit $i \in A^c$ using Procedure 2 and then select a unit $j \in A \cup \{i\}$ using Procedure 2′, where Procedure 2 is described in § 3 and Procedure 2′ is the 'backward' version of Procedure 2.

Step 2. If $i \neq j$, take $A \cup \{i\} \backslash \{j\}$ as the new sample and stop; otherwise go back to Step 1.

In Scheme 2, we use Procedure 2 to add a unit to the current sample A and then use Procedure 2′ to remove a unit from A plus this new unit. If the unit added and the unit removed are different, we accept the final sample as the new sample; otherwise, we repeat the procedure until a sample different from A is obtained. Procedure 1 cannot be used in place of Procedure 2 for adding units to the current sample because $P_1(., A^c)$ is not defined when $|A| = n$. There is no simple form for the acceptance rate of Scheme 2, though we know that, in the same extreme cases as for Scheme 1 above, it is $(N-n)/(N-n+1)$, and 0, respectively.

5. NUMERICAL EXAMPLES

Example 1. We use the following simulation to illustrate the properties of the mapping between w and π described in Lemma 1. Let $N = 100$. A vector π^* is generated uniformly from the simplex

$$\{\pi = (\pi_1, \ldots, \pi_{100}) : 0 < \pi_i < 1, i = 1, \ldots, 100; \sum \pi_i = 50\}$$

and its corresponding w^* is found from (6) via the iterative procedure described in (7). For convenience of comparison, this particular w^* is used as the weights for five different maximum entropy models with sample size $n = 2, 5, 15, 30, 50$, respectively. The coverage probabilities π for each of these five models are obtained by using (5). Then by using (7), each of these five π's is converted to a w, which should be the same as w^* up to a scalar. We use $\max |w_i^{(t)}/\tilde{w}_i - 1| < 0{\cdot}01$ as the stopping rule, where $w^{(t)}$ is the value of w at step t and $\tilde{w} = w^* \pi_N / w_N^*$ is the fixed point. The results are summarized in Table 1.

Table 1. *Number of iterations for computing w from π and the range of w and π for sample size $n = 2, 5, 15, 30, 50$*

n	Number of iterations	π_{100} (w_{100})	π_1	w_1	w_1/π_1
2	3	0·2634	$5{\cdot}350 \times 10^{-5}$	$4{\cdot}651 \times 10^{-5}$	0·8736
5	4	0·4576	$1{\cdot}489 \times 10^{-4}$	$9{\cdot}026 \times 10^{-5}$	0·5958
15	10	0·7730	$6{\cdot}201 \times 10^{-4}$	$1{\cdot}547 \times 10^{-4}$	0·2301
30	20	0·9165	$2{\cdot}020 \times 10^{-3}$	$1{\cdot}813 \times 10^{-4}$	0·08851
50	16	0·9903	$7{\cdot}627 \times 10^{-3}$	$1{\cdot}947 \times 10^{-4}$	0·02553

The procedure does not take many iterations even when π_N is close to 1 and π_1 is close to 0. Although the value of w_N differs from experiment to experiment, actually it is always set to be π_N, the ratios w_N/w_i $(i = 1, \ldots, N-1)$ remain nearly the same for all five experiments, which is explained by the fact that w is determined up to rescaling (Theorem 1). The phenomenon that the ratio w_1/π_1 approaches 1 as N/n gets large is supported by (d) of Lemma 1.

Example 2. Let $N = 4$ and $n = 2$. The π_i are 0·1, 0·4, 0·7 and 0·8, respectively. The corresponding w_i found from (6) are 0·05193, 0·2355, 0·52 and 0·8, respectively. The second-order inclusion probabilities π_{ij} are given in Table 2.

Table 2. *Second-order inclusion probabilities π_{ij}*

i	$j=2$	$j=3$	$j=4$
1	0·01514	0·03343	0·05143
2	—	0·1516	0·2333
3	—	—	0·5151

ACKNOWLEDGEMENT

Partial support was provided by an Army Research Office Grant and a National Science Foundation Grant to Harvard University.

APPENDIX

Proofs

The proofs presented in the Appendix are only a sketch. Detailed proofs can be found in a technical report by X.-H. Chen at Department of Statistics, Harvard University.

PROPOSITION 2. *We have* $R(k-2, C)R(k, C) < \{R(k-1, C)\}^2$ *for any* $C \subset S$ *and* $2 \leqslant k \leqslant |C|$.

Proof. Each term in $R(k-2, C)R(k, C)$ has the form $w_{i_1}^2 \ldots w_{i_j}^2 w_{i_{j+1}} \ldots w_{i_{2k-j-2}}$, where $0 \leqslant j \leqslant k-2$, corresponding to the case when both $R(k-2, C)$ and $R(k, C)$ choose $w_{i_1}, \ldots, w_{i_j}$ and in addition, $k-j-2$ of $w_{i_{j+1}}, \ldots, w_{i_{2k-j-2}}$ are from $R(k-2, C)$ and the other $k-j$ are from $R(k, C)$. This particular term appears

$$\binom{2k-2j-2}{k-j}$$

times in $R(k-2, C)R(k, C)$. By analogy, the same term appears

$$\binom{2k-2j-2}{k-j-1}$$

times in $\{R(k-1, C)\}^2$. Since a binomial function $(2i)!/\{(2i-m)!m!\}$ reaches its maximum when $m = i$, we find

$$\binom{2k-2j-2}{k-j} < \binom{2k-2j-2}{k-j-1}$$

and thus $R(k-2, C)R(k, C) < \{R(k-1, C)\}^2$. □

Proof of Lemma 1. Take the ratio of the ith equation and the jth equation of (5):

$$\frac{\pi_i}{\pi_j} = \frac{w_i R(n-1, \{i\}^c)}{w_j R(n-1, \{j\}^c)} = \frac{w_i w_j R(n-2, \{i, j\}^c) + w_i R(n-1, \{i, j\}^c)}{w_j w_i R(n-2, \{i, j\}^c) + w_j R(n-1, \{i, j\}^c)}. \tag{A1}$$

Then (*a*) and (*b*) can be directly deduced from (A1). Since the right-hand side of (A1) $\leqslant w_i/w_j$ if and only if $w_i \geqslant w_j$, (*c*) is also straightforward by the condition $\pi_i \geqslant \pi_j$ and (*a*) and (*b*). By manipulating the equation in (A1) for π_1 instead of π_j and using (*c*), we can show

$$\frac{\pi_i}{\pi_1} \leqslant \frac{w_i}{w_1} \leqslant \left\{\frac{\pi_i}{\pi_1}\left(\frac{N}{n} - 1\right)\right\} \Big/ \left(\frac{N}{n} - \frac{\pi_i}{\pi_1}\right). \tag{A2}$$

As $N/n \to \infty$, the right-hand side of (A2) approaches π_i/π_1. Thus for any pair $i \neq j$, we have $w_i/w_j \to \pi_i/\pi_j$ as $N/n \to \infty$. □

PROPOSITION 3. *Let* $\mathcal{X}$ *be a convex subset of* $\Re^m$, *and* $g = (g_1, \ldots, g_m)$ *a function* $g: \Re^m \to \Re^m$. *Given any two vectors* x *and* y *in* $\mathcal{X}$, *let*

$$l(x, y) = \{z \mid z = \lambda x + (1-\lambda)y, 0 \leqslant \lambda \leqslant 1\}.$$

If each g_i *is*

(*a*) *continuous at every point of* $l(x, y)$,

(*b*) *differentiable at every interior point of* $l(x, y)$,

then

$$\|g(x) - g(y)\| \leqslant \left\{\sup_{z \in l(x,y)} \|g'(z)\|\right\} \|x - y\|,$$

where $\|.\|$ *denotes the* l_∞ *norm; i.e. for a vector* x, $\|x\| = \max |x_i|$, *and for a matrix* $A = (a_{ij})_{m \times m}$, $\|A\| = \max (|a_{i1}| + \ldots + |a_{im}|)$.

Proof. Use the mean value theorem for multivariate functions and basic calculus. □

Proof of Theorem 2. (*a*) By Theorem 1, there is a unique $\tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_N)$ for the maximum entropy model if we let $\tilde{w}_N = \pi_N$. By Lemma 1, $\tilde{w}_i/\tilde{w}_N \leqslant \pi_i/\pi_N$. Thus $\tilde{w}_i \leqslant \pi_i$ and (6) has exactly one solution in $\mathscr{W}$.

(*b*) Let $x_i = \log w_i$ for each $i \in S$,

$$g_i(x) = \log \pi_i + \log R(n-1, \{N\}^c) - \log R(n-1, \{i\}^c)$$

for $i = 1, \ldots, N-1$ and $g_N(x) = \log \pi_N$. For convenience, we consider the logarithmic transformation of (6), which can be written as

$$x = g(x),$$

where $x = (x_1, \ldots, x_N)$ and $g(x) = (g_1(x), \ldots, g_N(x))$. Define

$$\mathscr{X} = \{x : \tilde{x}_i \leqslant x_i \leqslant \log \pi_i, i \in S\},$$

where $\tilde{x}_i = \log \tilde{w}_i$ for each $i \in S$. It can be checked that $\mathscr{X}$ and g satisfy all conditions in Proposition 3 and $g(\mathscr{X}) \subseteq \mathscr{X}$. It is obvious that $\partial g_i(x)/\partial x_j = 0$ when $i = N$ or $j = N$. It can be easily checked that

$$\frac{\partial g_i(x)}{\partial x_j} = \begin{cases} \dfrac{w_i R(n-2, \{N, i\}^c)}{R(n-1, \{N\}^c)} & \text{if } i = j \neq N; \\ \dfrac{w_j R(n-2, \{N, j\}^c)}{R(n-1, \{N\}^c)} - \dfrac{w_j R(n-2, \{i, j\}^c)}{R(n-1, \{i\}^c)} & \text{if } i, j \text{ and } N \text{ are distinct.} \end{cases}$$

Using Proposition 1 and Proposition 2, we can show that

$$\sum_{j=1}^{N} \left| \frac{\partial g_i(x)}{\partial x_j} \right| = \frac{w_N R(n-2, \{i, N\}^c)}{R(n-1, \{i\}^c)} < \frac{w_N R(n-1, \{N\}^c)}{R(n, S)}.$$

Using Proposition 2, it can be shown that $w_N R(n-1, \{N\}^c)/R(n, S)$ is decreasing in w_i for each $i = 1, \ldots, N-1$. Thus

$$\sup_{x \in \mathscr{X}} \| g'(x) \| = \sup_{x \in \mathscr{X}} \left\{ \max_{1 \leqslant i \leqslant N} \sum_{j=1}^{N} \left| \frac{\partial g_i(x)}{\partial x_j} \right| \right\} < \sup_{w \in \mathscr{W}} \left\{ \frac{w_N R(n-1, \{N\}^c)}{R(n, S)} \right\}$$

$$= \left. \frac{w_N R(n-1, \{N\}^c)}{R(n, S)} \right|_{w = \tilde{w}} = \pi_N.$$

The last step is obtained by the Nth equation of (5) since $\tilde{w}$ is a solution of (5). By Proposition 3,

$$\| x^{(t+1)} - x^{(t)} \| = \| g(x^{(t)}) - g(x^{(t-1)}) \| < \pi_N \| x^{(t)} - x^{(t-1)} \|. \tag{A3}$$

It is well known that the inequality in (A3) is a sufficient condition for any fixed point procedure to converge. Thus the sequence $x^{(t)}$ $(t = 1, 2, \ldots)$ converges to $\tilde{x}$ since $\tilde{x}$ is the unique fixed point in $\mathscr{X}$. Accordingly, we have that the sequence $w^{(t)}$ $(t = 1, 2, \ldots)$ converges monotonically and geometrically to $\tilde{w}$ and the convergence rate is bounded by π_N. Although we do not know $\tilde{w}$ beforehand, we can always choose $\pi \in \mathscr{W}$ as the starting point. □

Proof of Theorem 3. It is easy to check that (8) holds for $k = 1$. For any $k \geqslant 2$, define

$$Q_d = \textstyle\sum_{i \in C} w_i^d R(k-d, C \backslash \{i\}) \quad (d = 1, \ldots, k).$$

It can be shown that, for $d = 1, \ldots, k-1$,

$$T(d, C) R(k-d, C) = Q_d + Q_{d+1}. \tag{A4}$$

Using (A4) repeatedly for $d = 1, \ldots, k-1$, we get

$$Q_1 = T(1, C)R(k-1, C) - Q_2 = T(1, C)R(k-1, C) - \{T(2, C)R(k-2, C) - Q_3\}$$
$$= \ldots = T(1, C)R(k-1, C) - T(2, C)R(k-2, C) + \ldots + (-1)^{k+1}Q_k.$$

By Proposition 1 and the definition of R,

$$Q_1 = \sum_{i \in C} w_i R(k-1, C\backslash\{i\}) = kR(k, C), \quad Q_k = \sum_{i \in C} w_i^k R(0, C\backslash\{i\}) = T(k, C).$$

Thus

$$R(k, C) = \frac{1}{k} Q_1 = \frac{1}{k} \sum_{i=1}^{k} (-1)^{i+1} T(i, C)R(k-i, C). \qquad \square$$

Proof of Property 2. Since the w_i are all positive, it is easy to see that $\pi_{ij} > 0$. Comparing π_{ij} and $\pi_i \pi_j$ term by term, it is seen that we only need to show

$$R(n, S)R(n-2, \{i, j\}^c) < R(n-1, \{i\}^c)R(n-1, \{j\}^c).$$

By combinatorial expansion, it can be shown that

$$R(n, S)R(n-2, \{i, j\}^c) - R(n-1, \{i\}^c)R(n-1, \{j\}^c) = R(n-2, \{i, j\}^c)R(n, \{i, j\}^c)$$
$$- \{R(n-1, \{i, j\}^c)\}^2$$
$$< 0.$$

The last step is a direct consequence of Proposition 2 and thus completes the proof. $\square$

References

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families (with Applications in Statistical Decision Theory)*. Hayward, CA: Institute of Mathematical Statistics.

Chaudhuri, A. & Vos, J. W. E. (1988). *Unified Theory and Strategies of Survey Sampling*. New York: Elsevier Science.

Darroch, J. N. & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43**, 1470–80.

Deming, W. E. & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11**, 427–44.

Hanif, M. & Brewer, K. R. W. (1980). Sampling with unequal probabilities without replacement: A review. *Int. Statist. Rev.* **48**, 317–35.

Hansen, M. H. & Hurwitz, W. N. (1943). On the theory of sampling from a finite population. *Ann. Math. Statist.* **14**, 333–62.

Hartley, H. O. & Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Statist.* **33**, 350–74.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–85.

Joe, H. (1990). A winning strategy for lotto games? *Can. J. Statist.* **18**, 233–44.

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

Stern, H. & Cover, T. M. (1989). Maximum entropy and the lottery. *J. Am. Statist. Assoc.* **84**, 980–85.

Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *J. R. Statist. Soc.* B **15**, 253–61.

[*Received June* 1993. *Revised April* 1994]