SSC Annual Meeting, June 1996
Proceedings of the Survey Methods Section

ON SAMPLING WITH UNEQUAL PROBABILITIES CLOSE TO REJECTIVE SAMPLING

Y. G. Berger[1]

ABSTRACT

Sampling n distinct units from a population with unequal probability without replacement is a problem often considered in the literature (*e.g.*, Hanif & Brewer, 1980). In this work, we investigate this problem from the maximum entropy point of view. First, we note that rejective sampling is maximum entropy sampling. We define the directed divergence of a design from maximum entropy sampling. We will see that this measure implies interesting properties concerning first and second order inclusion probabilities. After that, we compute a maximum value for the directed divergence for Rao & Sampford sampling (Rao, 1965 and Sampford, 1967) and successive sampling (Hájek, 1964). We show that these sampling methods are asymptotically equivalent to maximum entropy sampling.

KEY WORDS: Unequal probability sampling with and without replacement; Inclusion probabilities; Maximum entropy; Poisson sampling; Rejective sampling; Successive sampling.

RÉSUMÉ

L'échantillonnage avec probabilités inégales de n unités distinctes d'une population est un problème décrit fréquemment dans la littérature statistique (*p. ex.* Hanif et Brewer, 1980). La présente étude vise à examiner le problème sous l'angle de l'entropie maximale. Nous notons d'abord que l'échantillonnage avec rejet est celui qui présente l'entropie maximale. La divergence dirigée d'un plan d'échantillonnage est définie par rapport à l'échantillonnage à entropie maximale. Nous montrons que cette mesure donne à penser que les probabilités d'inclusion de premier et de deuxième ordres présentent des propriétés intéressantes. Nous calculons ensuite la valeur maximale de la divergence dirigée pour l'échantillonnage de Rao et Sampford (Rao, 1965 et Sampford, 1967) et pour l'échantillonnage répété (Hájek, 1964). Nous montrons que ces méthodes d'échantillonnage sont asymptotiquement équivalentes à l'échantillonnage à entropie maximale.

MOTS CLÉS: Échantillonnage avec probabilités inégales exhaustif et non exhaustif; probabilités d'inclusion; entropie maximale; échantillonnage de Poisson; échantillonnage avec rejet; échantillonnage répété.

1. INTRODUCTION

Let U be a finite population of N units labelled $i=1,\dots,N$. A sample s is a sub-set of distinct units from U. We note a sample as the following

$$s=\{i_1,i_2,\dots,i_{n(s)}\};$$

where i_j is the label of the j^{th} unit in s. The number of distinct components of s, $n(s)$, is called the sample size. Let S be the set of all the samples, s, of distinct units from U. Any function $P(s)$ on S satisfying

(A.1) $\qquad P(s)\geq 0$ for all $s\in S$,

(A.2) $\qquad \sum_{s\in S} P(s)=1;$

is called a sampling design. Note that we only consider sampling without replacement. Let n be the expected sample size; *i.e.*,

$$n=\sum_{s\in S} P(s)n(s).$$

Moreover, $P(s)$ is a fixed size design if $P(s)$ verifies (A.1), (A.2) and

(A.3) $\qquad P(s)\geq 0$ for all s such that $n(s)=n$,

(A.4) $\qquad P(s)=0$ for all s such that $n(s)\neq n$.

[1] Yves Berger, Université Libre de Bruxelles, C.P. 124, L.M.T.D. Avenue Jeanne, 44, B-1050 Bruxelles, BELGIUM, e-mail: yvberger@ulb.ac.be.

n is a constant, fixed a priori, called the sampling size.

The first and second-order inclusion probabilities are defined by

$$\pi_i(P)=\sum_{s\in S} P(s) I_i,$$

$$\pi_{ij}(P)=\sum_{s\in S} P(s) I_i I_j,$$

where $I_i=1$ if $i\in s$ and 0 otherwise.

Let X_i be the value of a known auxiliary characteristic for the i-th unit. In this paper, we treat the case of $\pi_i(P)$ proportional to a positive value X_i; *i.e.*,

$$\pi_i(P)=n\frac{X_i}{\sum_{j\in U} X_j}. \tag{1}$$

We also suppose that $nX_i \le \sum_{j\in U} X_j$. A sampling which verifies (1) is called an unequal probability sampling design. Very often, we prefer implementing such a sampling because a common problem in survey sampling is to estimate the total $Y=\sum_{i\in U} Y_i$ of an unknown characteristic measured without error; Y_i being the unknown value of this characteristic for the i-th unit of U. One of the best strategies is to use the Horvitz & Thompson (1951) estimator, since the variance of this estimator becomes zero, if Y_i is exactly proportional to X_i, and if $\pi_i(P)$'s are equal to (1).

In section 2, we start with general considerations concerning entropy. In section 3, Poisson sampling and rejective sampling are introduced and analysed from the entropy point of view. Section 4 deals with directed divergence and inclusion probabilities for sampling which is not maximum entropy sampling. In the last two sections, we compute a maximum value of the directed divergence for Rao & Sampford sampling and successive sampling; and we conclude that under special conditions these sampling methods are close to maximum entropy sampling.

2. INFORMATION AND ENTROPY OF A DESIGN

Sampling from a finite population may be considered as a random experiment whose outcomes are samples s. We gain some information from U because we know that s occurred. Let $I(s)$ give a quantitative measure of the information gained if the event s results from the outcome of the experiment. We impose three natural requirements on $I(s)$:

(i) $I(s)$ only depends on the sampling design $P(s)$;

(ii) $I(s)$ is a continuous function of $P(s)$;

(iii) $I(s)\ge 0$;

(iv) $I(s^1 \text{ and } s^2) = I(s^1)+I(s^2)$ if and only if s^1 and s^2 are independent;

(v) if $P(s^1)<P(s^2)$ then $I(s^1)<I(s^2)$.

The only functions that satisfy (i), (ii), (iii), (iv) and (v) are

$$I(s)=-\mathrm{Log}_a[P(s)] \text{ for } a\ge 1.$$

We choose $a=\mathrm{Exp}(1)=2.718...$ to be the base throughout this paper.

The Shannon entropy (Shannon, 1948), defined by $H(P)$ is the expected value of the information.

$$H(P)=-\sum_{s\in S} P(s)\,\mathrm{Log}[P(s)].$$

We always assume that

$$0\,\mathrm{Log}(0)=0. \tag{2}$$

This means $H(P)$ exists even if there is s such that $P(s)=0$. We simply call $H(P)$ the entropy of $P(s)$. $H(P)$ represents the average amount of information contained in the sampling design, $P(s)$. Since the function $-x\,\mathrm{Log}(x)$ is strictly concave, there is a single maximum entropy sampling for a set S^*. Our aim is to find the sampling design that maximises $H(P)$.

Let S^* be a sub-set of S. Let $P^*(s)$ be the maximum entropy sampling, with $\pi_i(P^*)$'s for first-order inclusion probabilities. $P^*(s)$ must be such that $\sum_{s\in S^*} P^*(s)=1$. The usual method of the Lagrange multiplicator gives (see Hájek, 1959)

$$P^*(s)=\prod_{i\in s}\lambda_i^* \text{ if } s\in S^*; \tag{3}$$

where λ_i^*'s are such that

$$\sum_{s\in S^*} P^*(s) I_i=\pi_i(P^*) \tag{4}$$

since the function $-x\,\mathrm{Log}(x)$ is strictly concave, there is a single maximum entropy sampling for a set S^*.

A maximum entropy for a sub-set S^{**} of S^* can be derived by means of conditionning. The following lemma gives a formal expression of this fact.

Lemma 1 Let $P^*(s)$ *be the maximum entropy sampling for a set* S^* *and with* $\pi_i(P^*)$*'s for first-order inclusion probabilities. Thus the conditional sampling*

$$P^{**}(s) = c^{**} P^*(s) \text{ , if } s \in S^{**}$$

with respect to a sub-set S^{**} *of* S^*, *is a maximum entropy sampling for* S^{**} *with* $\pi_i(P^{**})$*'s for first-order inclusion probabilities, where*

$$\pi_i(P^{**}) = \sum_{s \in S^*} P^{**}(s) I_i.$$

c^{**} *is chosen as to achieve* $\sum_{s \in S^{**}} P^{**}(s) = 1$.

The proof is obvious using (3) and (4).

$P^{**}(s)$ can be implemented by selecting a sample according to $P^*(s)$. If $s \notin S^{**}$, we select a new sample until $s \in S^{**}$. Note that usually the $\pi_i(P^*)$'s are different from $\pi_i(P^{**})$'s.

3. MAXIMUM ENTROPY SAMPLING

If no auxiliary characteristic is available, we take n/N for first-order inclusion probabilities. In this case, it is clear from (3) that the maximum entropy sampling is the simple Poisson sampling,

$$SPO(s) = \left(1 - \frac{n}{N}\right)^N \left(\frac{n}{N-n}\right)^{n(s)}, \text{ with } s \in S.$$

This sampling is implemented by selecting each unit with a probability n/N.

Moreover, if we want a fixed-size design with maximum entropy, $SPO(s)$ must be conditioning. By lemma 1, the maximum entropy sampling is given by

$$SRS(s) \propto \left(1 - \frac{n}{N}\right)^N \left(\frac{n}{N-n}\right)^n, \text{with } s \in S_n;$$

where S_n is the set of all samples of n distincts units from U. As $\sum_{s \in S_n} SRS(s) = 1$, it is clear that $SRS(s)$ is the well known simple random sampling.

If an auxiliary characteristic is known for each unit of the population, we prefer to use an unequal probability sampling design with $\pi_i(PO)$'s equal to (1). We may use the same procedure as equal probability sampling. In this case, the maximum entropy sampling is given by Poisson sampling

$$PO(s) = c(PO) \prod_{i \in s} \alpha_i(PO), \text{ with } s \in S;$$

where

$$\alpha_i(PO) = \lambda(PO) \frac{\pi_i(PO)}{1 - \pi_i(PO)},$$

$\lambda(PO)$ is such that $\sum_{i=1}^N \alpha_i(PO) = 1$ and $c(PO)$ is chosen as to achieve $\sum_{s \in S} PO(s) = 1$. Practically, Poisson sampling may be easily performed by selecting each unit with a probability equal to $\pi_i(PO)$.

If we focus our interest on fixed size design, the conditional Poisson sampling gives the maximum entropy sampling in the class of all the fixed size sampling designs without replacement with first-order inclusion probabilities fixed to (1).

$$R(s) = c(R) \prod_{i \in s} \alpha_i(R), \text{ for all } s \in S_n; \qquad (5)$$

where

$$\alpha_i(R) = \lambda(R) \frac{p_i(R)}{1 - p_i(R)},$$

where $c(R)$ is such that $\sum_{s \in S} R(s) = 1$ and $\lambda(R)$ is such that $\sum_{i=1}^N \alpha_i(R) = 1$. In fact, $R(s)$ is the rejective sampling introduced by Hájek (1959).

We suppose that the drawing probabilities $p_i(R)$ are such that $R(s)$ has first-order inclusion probabilities equal to $\pi_i(R)$ and such that $\sum_{i=1}^N p_i(R) = n$.

The $p_i(R)$'s have to be distinguished from the inclusion probabilities $\pi_i(R)$. It has been proved by Dupačová (1979) that there always exists a unique set of $p_i(R)$ such that the first-order inclusion probabilities of $R(s)$ are exactly equal to $\pi_i(R)$.

Hájek (1981) gives an approximation for the $\alpha_i(R)$ which implies $\pi_i(R)$ for first-order inclusion probabilities.

$$\alpha_i(R) = c(R) \frac{\pi_i(R)}{1 - \pi_i(R)} A_i;$$

where

$$A_i = 1 + \frac{\bar{\bar{\pi}} - \pi_i(R)}{d} + O(d^{-1}),$$

$$d = \sum_{i=1}^N \pi_i(R)(1 - \pi_i(R)),$$

$$\bar{\bar{\pi}} = d^{-1} \sum_{i=1}^N \pi_i^2(R)(1 - \pi_i(R)),$$

where $O(d^{-1})$ is a number such that $d\, O(d^{-1}) \to 0$ if $d \to \infty$.

The rejective sampling is performed by drawing each unit i, with replacement, with a probability proportional to $\alpha_i(R)$. If a unit was selected at a previous draw, we reject the whole sample and start again until we have selected a sequence of n distinct units. Thus $R(s)$ is a conditional sampling with replacement. This implies that

$$c(R) = \frac{n!}{P(A)}; \tag{6}$$

where $P(A)$ is the probability that the units drawn at n independent draws are distinct if the drawing probabilities $\{\alpha_i(R): 1 \le i \le N\}$ are used. The term $n!$ is due to the fact that there are $n!$ permutations of the element of a sample s having the same probability $R(s)$.

Of course, rejective sampling is impracticable since the probability of rejecting a sample may be too large. Finally, rejective sampling is to sampling with varying first inclusion probabilities what simple random sampling is to sampling with constant first inclusion probabilities. Moreover, when sampling with varying probabilities without replacement is used, the difficulty is due to the fixed sample size.

4. DIRECTED DIVERGENCE

But most sampling methods used in practice are different from rejective sampling. One now wishes to find fixed size designs which are as 'close' as possible to the rejective sampling $R(s)$. We define the measure of 'closeness' of a sampling method $P(s)$ to $R(s)$ by,

$$D(P \| R) = \sum_{s \in S_n} P(s) \text{ Log} \left[\frac{P(s)}{R(s)}\right]; \tag{7}$$

$D(P \| R)$ is called the directed divergence of $P(s)$ from $R(s)$. By (5) we see that $R(s) > 0$ for all $s \in S_n$. Thus if there exists $s \in S_n$ such that $P(s) = 0$, $D(P \| R)$ still exists since we have the assumption (2). $D(P \| R)$ represents a positive quantity since $Log(x) \ge 1 - x^{-1}$. Moreover,

$$D(P \| R) = 0 \text{ if and only if } P(s) = R(s) \tag{8}$$

for all $s \in S_n$. We can prove (8) by considering the function $Log(x)$ and its tangent $x - 1$ at point $x = 1$. That implies

$$Log\left[\frac{R(s)}{P(s)}\right] = \frac{R(s)}{P(s)} - 1 - \delta(s),$$

where $\delta(s) \ge 0$. Thus

$$D(P \| R) = \sum_{s \in S_n} P(s)\delta(s).$$

Now, observe that $\delta(s) = 0$ if and only if $P(s) = R(s)$. Thus (8) follows.

What happens if the directed divergence of a sampling design tends to zero? We will see that in this case, such a design has interesting properties.

Consider arrays $\{n_1, n_2, \dots, n_k, \dots\}$ and $\{N_1, N_2, \dots, N_k, \dots\}$ to be sequences of sample size and population size, where n_k and N_k both increase as $k \to \infty$. For the sake of simplicity, we shall suppress the dependence on k.

Lemma 2

$$\sum_{s \in S_n} | P(s) - R(s) | \le \sqrt{2D(P \| R)}.$$

One may find the proof in Kemperman (1969).

With the following theorem, we can conclude that if $D(P \| R) \to 0$ as $k \to \infty$, then the first-order inclusion probabilities of $P(s)$ are asymptotically equivalent to the first-order inclusion probabilities of $R(s)$. We may conclude the same with second-order inclusion probabilities.

Theorem 1

$$Sup_i \mid \pi_i(P) - \pi_i(R) \mid \le \sqrt{2D(P \| R)},$$

$$Sup_i \mid \pi_{ij}(P) - \pi_{ij}(R) \mid \le \sqrt{2D(P \| R)},$$

where $\pi_i(P)$'s are the first-order inclusion proba- bilities of the design $P(s)$; $\pi_{ij}(P)$'s are the respective second-order inclusion probabilities of the design $P(s)$; $\pi_i(R)$'s and $\pi_{ij}(R)$'s are the first and the second-order inclusion probabilities of the rejective sampling.

By lemma 2, using triangle inequalities, it is clear that

$$\sum_{s \in S_n} [P(s) - R(s)] I_i < \sqrt{2D(P \| R)},$$

$$\sum_{s \in S_n} [R(s) - P(s)] I_i < \sqrt{2D(P \| R)}.$$

These two last inequalities imply that

$$-\sqrt{2D(P \| R)} < \pi_i(P) - \pi_i < \sqrt{2D(P \| R)}$$

which implies the first part of the theorem. By a similar method, we obtain the second part of the theorem.

5. DIRECTED DIVERGENCE OF RAO & SAMPFORD SAMPLING

In Rao & Sampford sampling, $Q(s)$ is a fixed size design defined by

$$Q(s)=c(Q)\left[\sum_{i\notin s}\pi_i(Q)\right]\prod_{j\in s}\frac{\pi_i(Q)}{1-\pi_i(Q)}, \tag{9}$$

for all $s\in S_n$ (see Hájek, 1981, p87), where $c(Q)$ is such that $\sum_{s\in S} Q(s)=1$. The first-order inclusion probabilities of $Q(s)$ are equal to $\pi_i(Q)$ (see Hájek, 1981).

We can perform this sampling by selecting the first unit with drawing probabilities $\pi_i(Q)/n$ $(i=1,\dots,N)$. The remaining $n-1$ units are drawn with drawing probabilities proportional to $\pi_i(Q)/(1-\pi_i(Q))$, $(i=1,\dots,N)$. We accept the sample if the units drawn are all distinct; otherwise, we reject s and start over again. This sampling is a modified rejective sampling. This modification is used to achieve precisely a set of first-order inclusion probabilities. The Rao & Sampford sampling is simpler than rejective sampling as it doesn't require computing the $\alpha_i(R)$.

The following theorem gives a maximum value for the directed divergence $D(Q\|R)$.

Theorem 2

$$D(Q\|R)\le\frac{\bar{\bar{p}}_2}{d}-\bar{\bar{p}}_1\left[\frac{\bar{\bar{\pi}}}{d}+O(d^{-1})\right],$$

where

$$\bar{\bar{p}}_1 = d_{(p;\pi)}^{-1}\sum_{i=1}^{N} p_i(R)^2(1-\pi_i(R)), \tag{10}$$

$$\bar{\bar{p}}_2 = d_{(p;\pi)}^{-1}\sum_{i=1}^{N} p_i(R)^2\pi_i(R)(1-\pi_i(R)), \tag{11}$$

$$d(p;\pi)=\sum_{i=1}^{N}p_i(R)(1-\pi_i(R)), \tag{12}$$

$Q(s)$ is the Rao & Sampford sampling with first-order inclusion probabilities $\pi_i(Q)$'s equal to $p_i(R)$'s.

With this last theorem, we see that $D(Q\|R)\to 0$ if $d\to\infty$. Thus by theorem 1, the first-order inclusion probabilities of $Q(s)$ can be approximated by π_i. We conclude that the Rao & Sampford sampling is close to rejective sampling, for $\pi_i(Q)=p_i(R)$.

6. DIRECTED DIVERGENCE OF SUCCESSIVE SAMPLING

The successive sampling denoted by $S(s)$ is a fixed size sampling design defined by (see Hájek, 1981)

$$S(s)=J(\alpha;s)\prod_{i\in s}\alpha_i(R),\ \text{for all } s\in S_n, \tag{13}$$

where,

$$J(\alpha;s)=\sum_{\sigma(i_1,\dots,i_n)} m[\alpha_{i_1}(R),\dots,\alpha_{i_n}(R)]^{-1}, \tag{14}$$

$$\begin{aligned} m[\alpha_{i_1}(R),\dots,\alpha_{i_n}(R)] &= [1-\alpha_{i_1}(R)]\\ &\cdot[1-\alpha_{i_1}(R)-\alpha_{i_2}(R)]\dots\\ &\cdot[1-\alpha_{i_1}(R)-\alpha_{i_2}(R)-\dots-\alpha_{i_{n-1}}(R)] \end{aligned} \tag{15}$$

and

$$\sum_{i=1}^{N}\alpha_i(R)=1 .$$

$\sum_{\sigma(i_1,\dots,i_n)}$ is the sum over all permutations of $\{i_1,\dots,i_n\}$ of all elements contained in the sample s.

Successive sampling is implemented by selecting each unit by independent draws with the drawing probabilities $\alpha_1(R),\dots,\alpha_N(R)$. This sequence of draws is continued until there are n distinct units in the sample s. Successive sampling removes the disadvantage of rejective sampling that lies in the frequent rejection of a part of sample in the case when a replication occurs. By comparing (5) and (13), we see that the successive sampling is not a maximum entropy sampling since $J(\alpha;s)$ depends on s.

The following theorem gives a maximum value for the directed divergence.

Theorem 3 *Let*

$$t=\frac{K}{N}\frac{n(n-1)}{2}<1,$$

we have,

$$D(S\|R)\le Log\left[\frac{P(A)}{1-t}\right]; \tag{16}$$

and

$$P(A) \leq 1 \leq \frac{P(A)}{1-t}, \qquad (17)$$

where K is such that

$$\frac{Max_i\, \alpha_i(R)}{Min_i\, \alpha_i(R)} < K \text{ ,for all } N. \qquad (18)$$

We suppose that K exists and is independent of N.

Assumption (18) is common; indeed, we may find the same assumption in Rosén (1972, p380).

If $t \to 0$, by (17) and (16), we have $P(A) \to 1$ and $D(S \| R) \to 0$. Thus the successive sampling is close to rejective sampling. The asymptotic approach adopted here is $t \to 0$; it can be achieved if $n^2/N \to 0$.

Furthermore, using theorem 1, if $t \to 0$, the inclusion probabilities of this sampling may be approximated by the inclusion probabilities of the rejective sampling.

7. CONCLUSION

The rejective sampling is the maximum entropy sampling in the class of all the fixed-size sampling designs without replacement with fixed first-order inclusion probabilities. We propose interesting results concerning Rao & Sampford sampling and successive sampling. We show that these sampling methods are close to maximum entropy sampling. We also give a sufficient condition for inclusion probabilities of any design to be asymptotically equal to inclusion probabilities of rejective sampling.

ACKNOWLEDGEMENT

The author is very grateful to Jean-Claude Deville (INSEE) and Alain Berlinet (Université de Montpellier II) for their interesting remarks concerning entropy.

REFERENCES

Dupačová, J. (1979). "A note on rejective sampling", *Contributions to Statistics (J. Hájek memorial volume), Academia Prague, 71-78.*

Hájek, J. (1959). "Optimum strategy and other problems in probability sampling", *Casopis Pro Pestovani Mathematiky (Journal for Cultivation of Mathematics)*, 84, 387-423.

Hájek, J. (1964). "Asymptotic theory of rejective sampling with varying probabilities from a finite population", *Annals of Mathematical Statistics*, 35, 1491-1523.

Hájek, J. (1981). *Sampling from a Finite Population*, New York and Bassel: Marcel Dekker.

Hanif, M., and Brewer, K.R.W. (1980). "Sampling with unequal probabilities without replacement: a review", *International Statistical Review*, 48, 317-355.

Horvitz, D.G., and Thompson, D.J. (1951). "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, 47, 663-685.

Kemperman, J.H.B. (1969). "On the optimum rate of transmitting information", *The Annals of Mathematical Statistics*, 40, 2156-2177.

Rao, J.N.K. (1965). "On two simple schemes of unequal probability sampling without replacement", *Journal of the Indian Statistical Association*, 3, 173-180.

Rosén, B. (1972). "Asymptotic theory for successive sampling with varying probabilities without replacement I, II", *Annals of Mathematical Statistics*, 43, 373-397 and 748-776.

Sampford, M.R. (1967). "On sampling without replacement with unequal probabilities of selection", *Biometrika*, 54, 494-513.

Shannon, C.E. (1948). "A mathematical theory of communication", *Bell System Technology Journal*, 27, 379-423, 623-656.

Yates, F., and Grundy, P.M. (1953). "Selection without replacement from within strata with probability proportional to size", *Journal of the Royal Statistical Society, Series B*, 1, 253-261.