SSC Annual Meeting, June 2018
Proceedings of the Survey Methods Section

Unequal probability sampling through random partitions

Ayi Ajavon [1]

ABSTRACT

The article introduces a class of sampling methods with unequal probabilities consisting in retaining distinct elements of a sampling with replacement. A direct application is non-rejective schemes for some sampling designs. A contribution of the article resides in presenting sampling designs using sequential occupancy distributions. Examples of the Sampford design and the splitting methods are proposed.

KEY WORDS: sampling, unequal probability, partition.

RÉSUMÉ

L'article introduit une classe de méthodes d'échantillonnage avec probabilités inégales consistant à conserver les éléments distincts d'un échantillonnage avec remise. Comme application directe, on peut citer les méthodes d'échantillonnage non-rejectives. Une contribution de l'article réside dans l'utilisation de modèles d'urnes pour décrire les plans d'échantillonnage. Des examples du plan de Sampford et des méthodes de partitionnement des probabilités d'inclusion sont proposées.

MOTS CLÉS : échantillonnage; probabilité inégale; partition.

1 INTRODUCTION

1.1 Description of the Problem

A simple random sampling is a type of random sampling where all the units have the same probability of inclusion. If we consider π the probability of inclusion of the units of a population $U = (1, \ldots, N)$ in a simple random sample $s = (s_1, \ldots, s_N)$ of size n, s_k being the number of occurrences of the unit k in the sample, $k = 1, \ldots, N$, some possible designs are:

(a) the Bernoulli design,

$$P(s, \pi) = \prod_{k \in U} \pi^{s_k} (1 - \pi)^{1 - s_k}, s_k \in \{0, 1\}, k = 1, \ldots, N;$$

(b) the hypergeometric design or simple random sampling without replacement (SRSWOR)

$$P(s, \pi = 1/N) = \frac{1}{\binom{N}{n}}, s_k \in \{0, 1\}, k = 1, \ldots, N;$$

(c) the multinomial design or simple random sampling with replacement (SRSWR)

$$P(s, \pi = 1/N) = \frac{n!}{s_1! \cdots s_N!} \prod_{k \in U} \frac{1}{N^{s_k}}, s_k \in \{0, 1, 2, \ldots\}, k = 1, \ldots, N.$$

The subsets of the distinct units of the sample, with each subset containing the repeated elements, induce a random partition of the sample. Also, sampling could be considered as partitioning a population into a set of sampled units and non sampled units, the sampled units being divided later into classes. Among sampling models that adopted this point of view, we could cite the combinatorial sampling models in which a sample of fixed size is

[1] christianolivier.nambeu@canada.ca; Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario K1A 0T6

selected from a finite population with a finite number of classes. Combinatorial sampling models rely primarily on occupancy models. In occupancy theory, balls of different colors (different types of draws) are distributed into urns or cells (units) of different kinds. Even though occupancy models are not popular among the survey statisticians, they could help in unifying the presentation of the various sampling designs and devising efficient algorithms. The paper will put the focus on the size-biased sampling in relation with occupancy models and how it could be used in implementing some sampling designs.

1.2 Organization of the paper

The plan of the paper is as follows. Section 2 conducts a literature review on selecting distinct units from a population. The size-biased sampling appears naturally as the sampling design equivalent to sampling distinct units from a population. Section 3 presents the size-biased sampling in the context of occupancy models. Section 4 depicts a sampling design that can be generated by symmetrizing the size-biased sampling. In section 5, an extension of the splitting methods is presented. Section 6 concludes the paper.

2 LITERATURE REVIEW

Pathak (1961, 1964) investigated a sampling with replacement scheme where the information about the multiplicity of the units is suppressed, and only the n first distinct units are conserved. The papers show that the procedure is equivalent to a random sampling without replacement. In this scheme, the sampling design induced by the serial-statistic $(s_{[1]}, s_{[2]}, \ldots, s_{[n]})$, $s_{[i]}$ being the i^{th} element to be selected in the sample, $i = 1, \ldots, n$, is called inverse sampling. The sampling design induced by the order-statistic $(p_{(1)}, p_{(2)}, \ldots, p_{(N)})$, p_i being the probability of inclusion of element i into the sample, is called successive sampling (Patterson (1950); Rosen (1972); Holst (1973)), the Plackett-Luce model (Plackett (1968, 1975); Stern (1990)) or size-biased sampling (McCloskey (1965)). Size-biased sampling is widely used in sampling models with the population containing a countable number of classes of random weights and is related to random partitions generated by the jumps of a Poisson process (Kingman (1975); McCloskey (1965); Perman (1993)).The size-biased sampling can also be introduced more simply using occupancy models. Occupancy models are useful in the description of randomized phenomena and have a wide range of applications in Sciences, Engineering and in Statistics (Charalambides (2005)).

3 Sized-biased sampling as an occupancy distribution

Let us consider a supply of indistinguishable balls randomly and independently distributed into N distinguishable urns $z_1, z_2, \ldots, z_N$ and assume that the probability for any ball to fall into the jth urn is p_{z_j}, $j = 1, 2, ..., N$. An urn cannot contain more than one ball. We consider the random variable $X_1 = 1$, the first time we select a distinct unit, z_1 the chosen unit, X_2 the first time we select the second distinct unit, z_2 the chosen unit,..., X_n the first time we select the nth distinct unit, z_n the chosen unit. Given $X_1, \ldots, X_k$ and $z_1, \ldots, z_{X_k}$, the variable $X_{k+1} - X_k$ is a geometric $Geo(1 - \sum_{j=1}^{k} p_{z_j})$ random variable and the joint distribution $(z_{X_{k+1}}, X_{k+1})$ is given by

$$P(z_{k+1} = z, X_{k+1} = x_{k+1} | X_1, \ldots, X_k) = (\sum_{j=1}^{k} p_{z_j})^{x_{k+1} - X_k - 1} p_z;$$

as a consequence, we have

$$\begin{aligned} P(z_{k+1} = z | X_{k+1} = x_{k+1}, X_k, \ldots, X_1) &= \frac{(\sum_{j=1}^{k} p_{z_j})^{x_{k+1} - X_k - 1} p_z}{(\sum_{j=1}^{k} p_{z_j})^{x_{k+1} - X_k - 1} (1 - \sum_{j=1}^{k} p_{z_j})} \\ &= \frac{p_z}{1 - \sum_{j=1}^{k} p_{z_j}}. \end{aligned} \tag{1}$$

We will write $P(z_k) = p_{z_k}$ as a simplification for $P(z_k = z) = p_z$. Therefore, the probability of selecting $(z_1, \ldots, z_n)$ given $X_1, \ldots, X_n$ is:

$$\begin{aligned} P(z_1, \ldots, z_n | X_1, \ldots, X_n) =& P(z_1|X_1) P(z_2|X_1, X_2, z_{X_1}) \cdots \\ & P(z_n|X_1, \ldots, X_n, z_{X_1}, \ldots, z_{n-1}) \\ =& p_{z_1} \frac{p_{z_2}}{(1 - p_{z_1})} \cdots \frac{p_{z_n}}{1 - \sum_{j=1}^{n-1} p_{z_j}} \\ =& p_{z_n} \prod_{j=1}^{n-1} \frac{p_{z_j}}{(1 - \sum_{l=1}^{j} p_{z_l})}. \end{aligned}$$

The design is equivalent to a size-biased sampling: the first unit z_1 is selected with probability p_{z_1}, and subsequent units being taken sequentially among units not yet sampled.

4 Final comments

Sampling designs consist of multivariate distributions with given margins characterized by the inclusion probabilities. The paper proposed sampling designs obtained by symmetrizing the size-biased sampling. Applications were proposed for the splitting methods and the Sampford design. The symmetric mixtures of size-biased sampling discussed in this article are part of the much larger families of occupancy distributions and discrete distributions invariant by size-biased permutation. Other sampling designs not explored in this paper are also part of these families. Using the sequential occupancy distributions, one could design performent drawing procedures for the sampling designs involved.

Acknowledgment

Thanks to Jean-Francois Fillon, Martin Beaulieu, Abel Dasylva, Jean-Francois Dubois and Christian Nambeu for correcting the first version of the paper.

References

Charalambides, C. A. (2005). *Combinatorial methods in discrete distributions*, Volume 600. John Wiley & Sons.

Deville, J.-C. and Y. Tillé (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika 85*(1), 89–101.

Holst, L. (1973, 07). Some limit theorems with applications in sampling theory. *Ann. Statist. 1*(4), 644–658.

Kingman, J. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–22.

McCloskey, J. W. (1965). *A Model for the Distribution of Individuals by Species in an Environment, unpublished Ph. D.* Ph. D. thesis, Michigan State University.

Pathak, P. K. (1961). Use of'order-statistic'in sampling without replacement. *Sankhyā: The Indian Journal of Statistics, Series A*, 409–414.

Pathak, P. K. (1964). On inverse sampling with unequal probabilities. *Biometrika 51*(1-2), 185–193.

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society. Series B (Methodological) 12*(2), 241–255.

Perman, M. (1993). Order statistics for jumps of normalised subordinators. *Stochastic processes and their applications 46*(2), 267–281.

Plackett, R. L. (1968). Random permutations. *Journal of the Royal Statistical Society. Series B (Methodological) 30*(3), 517–534.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 24*(2), 193–202.

Rosen, B. (1972, 04). Asymptotic theory for successive sampling with varying probabilities without replacement, i. *Ann. Math. Statist. 43*(2), 373–397.

Stern, H. (1990). Models for distributions on permutations. *Journal of the American Statistical Association 85*(410), 558–564.