Computational Statistics and Data Analysis 105 (2017) 11–23

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Quasi-systematic sampling from a continuous population

Matthieu Wilhelm *, Yves Tillé, Lionel Qualité

Institute of Statistics, University of Neuchâtel, Switzerland

ARTICLE INFO

Article history:
Received 22 December 2015
Received in revised form 18 July 2016
Accepted 18 July 2016
Available online 25 July 2016

Keywords:
Binomial process
Point process
Poisson process
Renewal process
Systematic sampling

ABSTRACT

A specific family of point processes is introduced that allow to select samples for the purpose of estimating the mean or the integral of a function of a real variable. These processes, called quasi-systematic processes, depend on a tuning parameter $r > 0$ that permits to control the likeliness of jointly selecting neighbor units in a same sample. When r is large, units that are close tend to not be selected together and samples are well spread. When r tends to infinity, the sampling design is close to systematic sampling. For all $r > 0$, the first and second-order unit inclusion densities are positive, allowing for unbiased estimators of variance. Algorithms to generate these sampling processes for any positive real value of r are presented. When r is large, the estimator of variance is unstable. It follows that r must be chosen by the practitioner as a trade-off between an accurate estimation of the target parameter and an accurate estimation of the variance of the parameter estimator. The method's advantages are illustrated with a set of simulations.

1. Introduction

We propose to use a specific family of point processes to select samples for the purpose of estimating the mean or the integral of a function of a real variable. We draw a parallel with sampling designs which are themselves point processes on finite spaces. Systematic sampling is widely used in finite population. It has been introduced by Madow and Madow (1944) and Madow (1949). It is easily implemented and, by spreading the sample over the population, it results in precise mean and total estimators when the variable of interest is similar for neighboring units. The main drawback of systematic sampling is that most of the unit joint inclusion probabilities are null, making it impossible to estimate the variance of the Horvitz–Thompson estimator without bias (see Horvitz and Thompson, 1952).

The aim of this paper is to develop a method that is a compromise between a base point process such as the Poisson process or the binomial process and the systematic process for sample selections in a continuous population. A similar objective is pursued in Breidt (1995) in a finite population setting supported by a superpopulation model. Breidt (1995) considers one-per-stratum sampling designs from a population that is split into strata of a successive units where a divides the population size. He introduces a class of sampling procedures that encompasses systematic sampling with constant rate $1/a$ and simple random sampling of one unit per stratum.

Point processes, that we refer to as *sampling processes* in the context of sampling, are the subject of a vast literature (see for example Daley and Vere-Jones, 2002, 2008, and references therein). Cordy (1993) and Deville (1989) introduced independently the continuous analogue to the Horvitz–Thompson estimator for infinite population sampling. Different communities have studied point processes: mathematical physicists, probabilists and statisticians. A detailed state of the art in the study and simulation of some complex point processes can be found in Møller and Waagepetersen (2003, 2007). Many simulation methods for point processes are implemented in the **R** package `spatstat` (Baddeley and Turner, 2005).

* Correspondence to: Institute of Statistics, University of Neuchâtel, Avenue de Bellevaux 51, 2000 Neuchâtel, Switzerland.
E-mail addresses: matthieu.wilhelm@unine.ch (M. Wilhelm), yves.tille@unine.ch (Y. Tillé), lionel.qualite@unine.ch (L. Qualité).

http://dx.doi.org/10.1016/j.csda.2016.07.011
0167-9473/

We introduce a new family of sampling methods that enable to continuously tune the distance between units in the sample. These processes allow to obtain small probabilities of jointly selecting neighboring units. These sampling methods are particularly efficient when the function of interest is smooth. Moreover, joint inclusion densities are positive and it is possible to estimate the sampling variance without bias.

The paper is organized as follows: in Section 2, we give a definition of sampling processes in continuous populations and we define the Poisson process, the binomial process and the systematic process. Important results of renewal process theory are recalled in Section 3. In Section 4, we define the systematic-Poisson and the systematic–binomial processes with tuning parameter r, and compute the joint densities. Section 5 contains proofs for the asymptotic processes when r tends to infinity. Simulations are presented in Section 6 and our ideas on the choice of the tuning parameter in Section 7. Finally, we give a brief discussion of the method and its advantages in Section 8.

2. Sampling from a continuous population

Following Macchi (1975) (see also Moyal, 1962), a finite sample of size n from a bounded and open subset Ω of $\mathbb{R}$ is a collection of units $X = \{x_1, \ldots, x_n\}$ without consideration for the order of the x_i's. This definition matches those commonly used in finite population sampling (see for example Cochran, 1977, for an introduction to finite population sampling theory). A sampling process is a probability distribution on the space $\mathcal{S}$ of all such collections, for all $n \in \mathbb{N}$. Note that it is not directly a distribution on $\Omega^{\mathbb{N}}$ equipped with the tensor product of Borel sigma algebras $\mathcal{B}(\Omega)$ as the sample units are not ordered. An extensive discussion on the definition of a sampling point process on Ω and the corresponding symmetric measure on $(\Omega^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}(\Omega))$ is given in Macchi (1975). It is sufficient for our purpose to know that a sampling point process is a probability distribution on $(\mathcal{S}, \mathcal{B})$ where $\mathcal{S} = \bigcup_{n\in\mathbb{N}} \Omega^n/\mathcal{R}^n$, with x and y in Ω^n being in the same class for the equivalence relation $\mathcal{R}^n$ if x is a permutation of elements of y, and $\mathcal{B}$ is the sigma algebra generated by the family of counting events:

$$\{s \in \mathcal{S} \text{ such that } N(s, A) = p,\ A \in \mathcal{B}(\Omega),\ p \in \mathbb{N}\},$$

and $N(s, A)$ is the number of elements of s that are in A.

The first and second factorial moment measures of a sampling point process X (Moyal, 1962) are defined respectively as

$$M_1 = \begin{pmatrix} \mathcal{B}(\Omega) & \to & \mathbb{R}_+ \\ A & \mapsto & \mathrm{E}\,[N(X, A)] \end{pmatrix},$$

where $N(X, A)$ is the random number of elements of X that are in A, and the second factorial moment measure is the extension to $\mathcal{B}(\Omega)^{\otimes 2}$ of

$$M_2 = \begin{pmatrix} \mathcal{B}(\Omega) \times \mathcal{B}(\Omega) & \to & \mathbb{R}_+ \\ A \times B & \mapsto & \mathrm{E}\,[N_2(X, A \times B)] \end{pmatrix},$$

where $N_2(X, A \times B)$ is the random number of pairs (x_i, x_j), $i \neq j$ of elements of X such that $x_i \in A$ and $x_j \in B$.

We call first and joint (second) order inclusion densities the respective densities of M_1 and M_2 with respect to the Lebesgue measure on Ω and Ω^2 when they exist. In that case, the first order inclusion density π is such that $M_1(A) = \int_A \pi(x)dx$, for all $A \in \mathcal{B}(\Omega)$, and the second-order inclusion density $\pi^{(2)}$ satisfies $M_2(A \times B) = \int_A \int_B \pi^{(2)}(x, y)dxdy$ for all $A \times B \in \mathcal{B}(\Omega) \times \mathcal{B}(\Omega)$. Heuristically, the term $\pi(x)dx$ can be viewed as the probability that one unit of the sample lies between x and $x + dx$, and $\pi^{(2)}(x, y)dxdy$ as the probability that one unit of the sample lies between x and $x + dx$ and another between y and $y + dy$, disregarding what happens outside of these sets. Likewise, one can define kth order factorial moments and, when they exist, inclusion densities for $k \geq 3$.

We now turn to the problem of estimating the mean of a Lebesgue integrable function z defined on Ω:

$$\bar{z} = \frac{1}{|\Omega|} \int_\Omega z(x)dx,$$

where $|\Omega|$ denotes the Lebesgue measure of Ω, using a finite random sample $X = \{x_1, \ldots, x_n\}$ of points in Ω. Assuming that Ω is bounded, $|\Omega|$ is known and X is a sampling process with inclusion density π, Cordy (1993) defines the continuous analogue of the Horvitz–Thompson estimator as:

$$\widehat{\bar{z}} = \frac{1}{|\Omega|} \sum_{i=1}^{n} \frac{z(x_i)}{\pi(x_i)},$$

and gives its properties. Under the assumption that $\pi(x) > 0$ on Ω and that z is bounded or non-negative, this estimator is unbiased (Cordy, 1993, Theorem 1). If, moreover, $\int_\Omega 1/\pi(x)dx < +\infty$, the variance of this estimator is given by:

$$\mathrm{var}\left(\widehat{\bar{z}}\right) = \frac{1}{|\Omega|^2} \int_\Omega \frac{[z(x)]^2}{\pi(x)} dx + \int_\Omega \int_\Omega z(x)z(y) \left[\frac{\pi^{(2)}(x, y) - \pi(x)\pi(y)}{\pi(x)\pi(y)}\right] dxdy,$$