Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 5471 – 5476

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Short Communication

Simulating from a multinomial distribution with large number of categories

Sonia Malefaki, George Iliopoulos*

Department of Statistics and Insurance Science, University of Piraeus, 80 Karaoli & Dimitriou Str., 18534 Piraeus, Greece

Received 4 August 2006; received in revised form 30 March 2007; accepted 31 March 2007
Available online 5 April 2007

Abstract

The multinomial distribution is a key-distribution for several applications. For this reason, many methods have been proposed so far in the literature in order to deal with the problem of simulation from it. A slight modification is suggested which can be used in conjunction with any of the standard schemes. The proposed variation is a two-stage procedure based on the property of the multinomial distribution that for any partition of the set of outcomes the vector of total frequencies of each part follows also a multinomial distribution with parameters adjusted accordingly. It is empirically exhibited that this variation is faster than the original procedures in case the numbers of independent trials and possible outcomes are both large. The time reduction is illustrated via a simulation study for several programming languages such as R, Matlab, and others.

Keywords: Multinomial distribution; Random number generation

1. Introduction

Nowadays, as computational power is rapidly growing, computer simulation is widely used in many disciplines. Simulation of random variables is often used when either no explicit theoretical results can be obtained or in order to confirm existing ones. There is a huge number of general as well as special methods for generating random variates from discrete or continuous distributions. An introduction on this topic can be found in Devroye (1986); for more recent results see Robert and Casella (2004). In this paper we focus on generating random samples from the multinomial distribution.

The multinomial distribution is a multidimensional generalization of the binomial distribution and is one of the most frequently used distributions in practical applications. Its most frequent use is in the analysis of contingency tables but it may be applied as well whenever data are grouped in a finite number of categories. Some recent applications can be found in McCulloch and Rossi (1994) and Imai and Van Dyk (2005). For more on multinomial distribution see Johnson et al. (1996) and the references therein.

* Corresponding author. Tel.: +30 210 414 2406, fax: +30 210 414 2340.
E-mail address: geh@unipi.gr (G. Iliopoulos).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.03.022

Let $\mathbf{P} = (p_1, \ldots, p_n)$ be a probability vector, that is, $0 < p_i < 1$, $i = 1, \ldots, n$, and $\sum_{i=1}^{n} p_i = 1$. A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is said to follow the multinomial distribution $\mathscr{M}_n(T; \mathbf{p})$ if its probability mass function (pmf) is

$$\mathrm{P}(\mathbf{X} = \mathbf{x}) = \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{T!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} p_i^{x_i}, \tag{1}$$

for $0 \leqslant x_i \leqslant T$, $i = 1, \ldots, n$, and $\sum_{i=1}^{n} x_i = T$. If X_i represents the frequency of the ith outcome in T independent replications of an experiment having n possible outcomes with respective probabilities $p_1, \ldots, p_n$, then the pmf of $\mathbf{X}$ is given by (1).

The most straightforward method to generate a sample from $\mathscr{M}_n(T; \mathbf{p})$ is the sequential search (cf. Devroye, 1986), also called "the direct method". Denoting the cumulative probabilities of $\mathbf{p}$ by $(P_1, \ldots, P_n)$ with $P_i = \sum_{j=1}^{i} p_j$, the method initializes a vector $\mathbf{X}$ to zeros and generates independent random variates $U_1, \ldots, U_T$ from the uniform distribution $\mathscr{U}(0, 1)$. Then, the ith component of $\mathbf{X}$ is increased by one if $P_{i-1} < U_j \leqslant P_i$, for $j = 1, 2, \ldots, T$.

Another popular method is based on the following well-known property of the multinomial distribution: X_1 follows the binomial distribution $\mathscr{B}(T; p_1)$ and the conditional distribution of X_i given $\sum_{j=1}^{i-1} X_j = t_{i-1}$ is also binomial $\mathscr{B}(T - t_{i-1}; p_i/(1 - P_{i-1}))$, $i = 2, \ldots, n$. According to this, $X_1, \ldots, X_n$ can be generated sequentially from the corresponding binomial distributions. Obviously, the key point here is the generation of binomial random variates. For this purpose, algorithms have been proposed by Ahrens and Dieter (1980), Kachitvichyanukul (1982) and Kemp (1986). This method is called "the conditional method".

In addition to the above schemes, there exists the alias method of Walker (1977) which is a general method for sampling from discrete distributions with a finite number of outcomes as well as Brown and Bromberg's (1984) two-stage algorithm which is based on the property that the conditional distribution of independent Poisson random variables given their sum is multinomial. Davis (1993) compared all these methods and concluded that the conditional method in conjunction with Kemp's (1986) algorithm for the binomial distribution "...is a good all-purpose algorithm for multinomial random variate generation, since it provides a nice balance between the competing criteria of execution speed and simplicity".

Generation from the multinomial distribution is part of many contemporary computational methods such as sampling/importance resampling (SIR, Rubin, 1987, 1988) and population Monte Carlo (PMC, Cappé et al., 2004). These algorithms proceed by generating at each step a large number of observations and assigning to each of them a weight. The step concludes by resampling with replacement from the observations with probabilities proportional to their weights. This is an application of multinomial sampling. In these methods, T and n are traditionally quite large. For example, McAllister and Ianelli (1997) illustrated a SIR scheme on a 54-dimensional model where in the resampling step $T = 10^4$ draws were performed from a sample of size $n = 3 \times 10^6$.

While we were implementing SIR and PMC to estimate some particular models, we found the computational time needed until convergence quite disturbing. In order to circumvent this drawback we tried many tricks, including splitting the multinomial sampling into two parts. Surprisingly, the reduction of the computational time was substantial and so we decided to write this short note.

The remaining of the paper proceeds as follows. In Section 2 we describe our two-stage variation for simulation from the multinomial distribution. In Section 3 we present some simulation results in several programming languages illustrating the CPU time reduction achieved by our variation. Finally, Section 4 contains our final conclusions.

2. The two-stage procedure

Let $A = \{1, \ldots, n\}$ and $\mathscr{B} = \{B_1, \ldots, B_m\}$ be an arbitrary partition of A consisting of non empty sets. Denote the cardinal number of B_j by n_j so that $\sum_{j=1}^{m} n_j = n$. Let $\mathbf{X} = (X_1, \ldots, X_n) = (X_i; i \in A)$ be a random vector and $\mathbf{p} = (p_i; i \in A)$ a probability vector. Moreover, for $j = 1, \ldots, m$ define $\mathbf{X}^{(j)} = (X_i; i \in B_j)$, $\mathbf{p}^{(j)} = (p_i; i \in B_j)$ and $q_j = \sum_{i \in B_j} p_i$. Then we have the following:

Lemma 2.1. *If* $\mathbf{K} = (K_1, \ldots, K_m) \sim \mathscr{M}_m(T; \mathbf{q})$ *and* $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}$ *are independent conditional on* $\mathbf{K} = \mathbf{k}$ *with* $\mathbf{X}^{(j)} | \mathbf{K} = \mathbf{k} \sim \mathscr{M}_{n_j}(k_j; \mathbf{p}^{(j)}/q_j)$, $j = 1, \ldots, m$, *then* $\mathbf{X} \sim \mathscr{M}_n(T; \mathbf{p})$.