# A COMPARISON OF METHODS FOR SELECTING VALUES OF SIMULATION INPUT VARIABLES

Megdouda Ourbih-Tari<sup>1</sup> and Sofia Guebli<sup>1</sup>

**Abstract.** Refined descriptive sampling (RDS) is a method of sampling that can be used to produce input values for estimation of expectation of functions of output variables. This paper gives a generalization of RDS method for K input variables. An estimator of RDS is defined and shown to be unbiased and efficient compared to simple random sampling with respect to variance criterion for a class of estimators. The efficiency of RDS algorithm is discussed at the end of the paper.

Mathematics Subject Classification. 62D05, 65C05, 62H12, 37M05.

Received October 17, 2012. Revised April 11, 2014.

#### 1. INTRODUCTION

Suppose we have some device, the behavior of which depends on a random vector  $\mathbf{X} = (X_1, X_2, \dots, X_K)$  of fixed length K with known probability density function f(x) and known cumulative function F(x) for  $x \in \mathbb{R}^{K}$ . A mathematical model for the device is developed from which we can simulate the behavior of the device on a computer. So experiments are carried out on the model built and unknown parameter  $\theta$  of the output random variable Y of interest denoted as the unknown but observable univariate transformation of  $\mathbf{X}$  given by the function  $Y = h(\mathbf{X})$  is estimated. Thus, we have the problem of approximating  $\theta$ . Since  $h(\mathbf{X})$  may be difficult to compute for each new value of X, it is important to pick a sampling scheme that allows us to estimate h(X)well while keeping N, the number of replication, to a minimum. There exist several procedures for choosing  $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ . The simplest is simple random sampling (SRS) also known as Monte Carlo (MC) method which is due to Metropolis and Uhlam, first published in 1947, but was developed in Los Alamos during the World War II. MC is used to generate N independent identically distributed random vectors with the distribution of X but this method presents sampling errors [14], the set and sequence effect. In the literature, we can find several efficient Monte Carlo algorithms for generating random numbers, for instance, super-convergent MC and Adaptive MC algorithms [1] for practical computations which are developed using variance reduction techniques  $(VRT)_s$  but the obtained simulation results are still of modest accuracy [3]. Because of the limits of MC methods, a new paradigm emerged: it is not always necessary to resort to randomness. Then, new sampling methods for generating  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$  without using VRT<sub>s</sub> are developed, for instance, Quasi Monte Carlo (QMC) [4,8], latin hypercube sampling (LHS) [6,7], descriptive sampling (DS) [13] as well as refined descriptive sampling (RDS) [18]. DS method avoids the set effect and keeps the sequence effect but it is not without drawbacks. It is

Keywords and phrases. Sampling, variance, estimation, simulation.

<sup>&</sup>lt;sup>1</sup> Laboratoire de Mathématiques Appliquées, Faculté des Sciences Exactes, Université de Bejaia, 06000 Bejaia, Algérie. ourbihmeg@gmail.com; sofiaguebli33@gmail.com

#### M. OURBIH-TARI AND S. GUEBLI

known to have two problems. First, it can be biased, and second, more practically, its strict operation requires a prior knowledge of the required sample size [11]. Based on DS, RDS was proposed to make DS safe, efficient and convenient. It is safe by reducing substantially the risk of sampling bias, efficient by producing estimates with lower variances and convenient by removing the need to determine in advance the sample size. The efficiency of RDS over DS and SRS is proved by several comparisons, for instance, on a flow shop system and a production system [16, 17]. In [9], there is a discussion deducing that RDS outperforms DS, SRS, LHS and QMC methods and it is capable of substantially reducing the cost of running simulation experiments, if properly applied.

Let  $T_{\rm SRS}$  and  $T_{\rm RDS}$  be the estimators of  $\theta$  obtained using respectively a simple random sample and a refined descriptive sample of size N. In Section 2, we consider a class of estimators used for the comparison between SRS and RDS. In Section 3, we describe DS while in Section 4 RDS is described for one dimension and in Section 5, we give an example of how to use DS and RDS. Section 6 proposes the generalization of RDS for Kinput random variables. In Section 7, the estimator  $T_{\rm RDS}$  is shown to be unbiased and efficient by studying its variance which is proved to be less than that of  $T_{\rm SRS}$  with respect to a class of estimators. A discussion about the efficiency of RDS algorithm is finally given in Section 8.

### 2. Estimators

Monte Carlo methods [2, 15] are usually used for high-dimensional problems. That is, N values of the input random vector,  $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$  are generated in some manner such that the parameter  $\theta = E(g(Y))$  can be estimated by

$$T_{\text{SRS}} = \mathbf{T}(Y_1, Y_2, \dots, Y_N) = \frac{1}{N} \sum_{i=1}^N g(Y_i)$$

where E design the mathematical expectation, the arguments  $Y_1, Y_2, \ldots, Y_N$  constitute a random sample of Y and g(.) is an arbitrary known function.

The mean and variance of  $T_{\rm SRS}$  are denoted by  $\theta$  and  $\frac{\sigma^2}{N}$  where  $\sigma^2$  is the variance of g(Y) obtained using simple random sampling. In this paper, RDS is examined and compared to SRS with respect to this class of estimators.

### 3. Descriptive sampling

DS proposed by [13] is based on a fully deterministic selection of the input sample values and their random permutation. Either a discrete or a continuous or even a mixed distribution can be represented, provided that the respective inverse of the distribution function is available. This inverse function is always defined, although, in most cases, a numerical approximation may be necessary, as in the case of a normal distribution [12]. Formally, in descriptive sampling when the sample size N is known, set values are first computed for the input random variable X using the inverse transform method, as follows

$$x_i = F_1^{-1}(r_i)$$
 for  $i = 1, 2, \dots, N$  (3.1)

and then a random sequence  $\{x'_i, i = 1, ..., N\}$  is drawn without replacement from  $\{x_i, i = 1, ..., N\}$ .

- F<sub>1</sub><sup>-1</sup> is the inverse cumulative input distribution.
  The stream of regular numbers {r<sub>i</sub> = <sup>i-0.5</sup>/<sub>N</sub>, i = 1, 2, ..., N} belongs to [0, 1[.

A descriptive sample  $\{x'_i, i = 1, ..., N\}$  is then defined as a set of input values  $x_i$  taken in a random sequence. Using DS, sample values are generated in advance and stored in memory to be used by the simulation.

The use of DS procedure leads to the following estimates of the parameter  $\theta$ 

$$T_{DS} = \frac{1}{N} \sum_{i=1}^{N} g(h(x'_i)).$$

### 4. Refined descriptive sampling

In a simulation run, RDS is based on a block that distributes descriptive subsets of sizes, randomly generated prime numbers as required by the simulation. We stop the process when the simulation terminates, say when m prime numbers  $p_j$ , have been used which derives m sub-runs, where j stands for the sub-run. In [18], the procedure was described for one input real-valued random variable X having a cumulative function  $F_1$  where its values are given by

$$x_i^j = F_1^{-1}(R_i^j)$$
 for  $i = 1, 2, \dots, p_j$  and  $j = 1, 2, \dots, m$ 

where each subset  $\{R_i^j, i = 1, ..., p_j\}$  is randomly selected without replacement from the following subset of regular numbers

$$\left\{ r_i^j = p_j^{-1} \left( i - 0.5 \right) \quad i = 1, 2, \dots, p_j \right\}$$

such as  $r_i^j$  is the midpoint of the  $p_j$  subintervals in which the interval [0, 1] is subdivided.

The primes  $p_j$  are randomly generated over [7, RANDMAX] where RANDMAX is a given value fixed by the user that the generated prime number does not exceed while 7 is the minimum value of generated primes chosen to avoid a possible sampling bias. Note that any prime below 7 has too many multiples and therefore if the underlying frequency is periodic, it could be a multiple of these particular primes [18].

The use of RDS generating  $\sum_{j=1}^{m} p_j$  values of the input random variable X leads to the following m estimates of the parameter  $\theta$  in each sub-run

$$\widehat{\theta}_j = \frac{1}{p_j} \sum_{i=1}^{p_j} g\left(h\left(F_1^{-1}\left(R_i^j\right)\right)\right) \qquad j = 1, \dots, m.$$

Therefore, in a given run, the use of refined descriptive sample of size  $N = \sum_{j=1}^{m} p_j$  leads to the following sampling estimate of  $\theta$  defined by the average of those estimates observed on different sub-runs

$$\widehat{\theta}_{\text{RDS}} = T_{\text{RDS}} = \frac{1}{N} \sum_{j=1}^{m} \sum_{i=1}^{p_j} g\left(h\left(F_1^{-1}\left(R_i^j\right)\right)\right)$$

## 5. An example of how to use DS and RDS

Let us suppose a variable X following an exponential distribution with mean  $\theta = 1$  and variance  $\theta' = 1$ .

To estimate  $\theta$ , we take  $g(x_i) = x_i$  and to estimate  $\theta'$  we take  $g(x_i) = (x_i - \overline{X})^2$ . The sample values are obtained by  $x_i = -\ln(1 - ri)$ , respectively for i = 1, ..., 31 using DS and for  $i = 1, ..., p_j$  and j = 1, 2, 3 for RDS.

Using DS, the estimates  $\hat{\theta}$  and  $\hat{\theta}'$  of E(X) and Var(X) are computed as sample mean and sample variance of descriptive sample of size N. The estimates are given in Table 1 and the sample values can be found in the appendix (see Tab. A.3).

TABLE 1. The observed mean and variance of a negative exponential distribution using a descriptive sample of size 31.

$\widehat{\theta} = T_{DS}(mean)$	0.98886
$\widehat{\theta}' = T_{DS}(var)$	0.89652

Using RDS and taking  $p_1 = 7$ ,  $p_2 = 11$  and  $p_3 = 13$ , we first computes the estimates  $\hat{\theta}_j$  and  $\hat{\theta}'_j$  of E(X) and Var(X) in each sub-run as sample mean and sample variance of different descriptive samples of size  $p_j$ .

The estimates  $\hat{\theta}$  and  $\hat{\theta}'$  of E(X) and var(X) are computed as the mean of the samples means and the mean of the samples variances as follows.

$$\widehat{\theta} = T_{\text{RDS}}(mean) = \frac{1}{31} \sum_{j=1}^{3} p_j \times \widehat{\theta}_j$$

and

$$\widehat{\theta}' = T_{\text{RDS}}(var) = \frac{1}{31} \sum_{j=1}^{3} p_j \times \widehat{\theta}'_j.$$

The estimates are given in Table 2 and the sample values can be found in the appendix (see Tabs. A4-A6).

TABLE 2. The observed mean and variance of a negative exponential distribution using a refined descriptive sample of size 31.

j	1	2	3	$T_{\rm RDS}$
$p_j$	7	11	13	
$\widehat{ heta}_j$	0.95134	0.96884	0.97359	$\widehat{\theta} = 0.966$
$\widehat{\theta'_i}$	0.60025	0.77260	0.79889	$\widehat{\theta}' = 0.744$

### 6. The proposed generalization of RDS

#### 6.1. Sample space

Let the sample space  $\Omega$  of **X** be partitioned into m disjoint finite sets  $A_i$  associated to the prime number  $p_i$ of vectors of size K, such as,

$$A_{j} = \left\{ r_{\underline{i}}^{j} = \left( r_{i_{1}}^{j}, \dots, r_{i_{K}}^{j} \right), \quad \underline{i} = (i_{1}, \dots, i_{K}) \in \{1, \dots, p_{j}\}^{K} \right\}$$

where card  $(A_j) = p_j^K$ .

Then,  $A_j$  is partitioned into  $p_j^K$  disjoint K dimensional hypercubes labeled by  $S_{\underline{i}}^j$ . The center of each hypercube  $S_i^j$  is defined by a K dimensional vector  $r_i^j$  of regular numbers. The sample space  $\Omega$  of **X** can be written by

$$\Omega = \bigcup_{j=1}^{m} A_j = \bigcup_{j=1}^{m} \bigcup_{i} S_i^j.$$

### 6.2. Sample values generation

Let  $\mathbf{X}_{\underline{i}}^{j}$  be the <u>i</u>th simulated vector of the jth sub-run situated in  $S_{\underline{i}}^{j}$  with a probability density function given in Section 6.5. Let  $\mathbf{R}_{\underline{i}}^{j}$  be the <u>i</u>th vector of the *j*th sub-run, where each component is randomly selected without replacement from the following K identically subsets of regular numbers:

$$\left\{r_1^j, \dots, r_i^j, \dots, r_{p_j}^j\right\}^K.$$
(6.1)

A refined descriptive sample of size  $N = \sum_{j=1}^{m} p_j$  is defined for the input random vector **X** using the inverse transform method as successive *m* descriptive samples of size  $p_j$ , j = 1, 2, ..., m where the <u>i</u>th vector of the jth descriptive sample is defined by

$$\mathbf{X}_{\underline{i}}^{j} = F^{-1} \left( \mathbf{R}_{\underline{i}}^{j} \right)$$

### 6.3. Bernoulli random variables

To have the appropriate form of the estimator  $T_{\text{RDS}}$  and try to calculate its variance, we introduce  $p_j^K$ Bernoulli independent random variables of parameter  $\alpha = \frac{p_j}{p_j^K} = \frac{1}{p_j^{K-1}}$  on the *j*th sub-run given by:

$$w_{\underline{i}}^j: S_{\underline{i}}^j \to \{0, 1\}$$

such as

 $w_{\underline{i}}^{j} = \begin{cases} 1 & \text{if } S_{\underline{i}}^{j} \text{ is in the descriptive sample of size } p_{j} \\ 0 & \text{otherwise} \end{cases}$ 

where each random variable having the following properties which are immediate:

$$\begin{array}{ll} (1) \ P(w_{\underline{i}}^{j} = 1) = E(w_{\underline{i}}^{j}) = E((w_{\underline{i}}^{j})^{2}) = \frac{1}{p_{j}^{K-1}} \\ (2) \ \operatorname{Var}(w_{\underline{i}}^{j}) = \left(\frac{1}{p_{j}^{K-1}}\right) \left(1 - \frac{1}{p_{j}^{K-1}}\right) \\ & 3. \ E\left(w_{\underline{i}}^{j} \times w_{\underline{t}}^{j}\right) = E\left(w_{\underline{i}}^{j} \times w_{\underline{t}}^{j}/w_{\underline{t}}^{j} = 0\right) P\left(w_{\underline{t}}^{j} = 0\right) + E\left(w_{\underline{i}}^{j} \times w_{\underline{t}}^{j}/w_{\underline{t}}^{j} = 1\right) P\left(w_{\underline{t}}^{j} = 1\right) \\ & = \frac{1}{p_{j}^{K-1}(p_{j} - 1)^{K-1}} \quad \text{if} \quad \underline{i} \neq \underline{t} \\ & \text{since} \quad E\left(w_{\underline{i}}^{j}/w_{\underline{t}}^{j} = 1\right) = \frac{1}{(p_{j} - 1)^{K-1}} \quad \text{when} \quad \underline{i} \neq \underline{t} \end{array}$$

#### 6.4. The estimator of RDS

In the *j*th sub-run, using RDS the parameter  $\theta = E(g(Y))$  of the output variable Y is estimated as follows

$$\widehat{\theta}_{j} = \mathbf{T} \left( Y_{1}, Y_{2}, \dots, Y_{p_{j}} \right)$$
$$= \frac{1}{p_{j}} \sum_{\underline{i}} w_{\underline{i}}^{j} \times g \left( y_{\underline{i}}^{j} \right)$$
(6.2)

and in the simulation experiment defined by m sub-runs, by the following  $T_{\text{RDS}}$  estimator defined by the average of the sub-runs estimates

$$T_{\rm RDS} = \mathbf{T}\left(\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_m\right) = \frac{1}{N} \sum_{j=1}^m p_j \times \widehat{\theta}_j$$
(6.3)

$$= \frac{1}{N} \sum_{j=1}^{m} \sum_{\underline{i}} w_{\underline{i}}^{j} \times g\left(y_{\underline{i}}^{j}\right), \tag{6.4}$$

where the arguments  $Y_j = \{Y_i^j, i = 1, ..., p_j\}$  constitute the *j*th descriptive sample of size  $p_j$  of Y and the arguments  $\{Y_j, j = 1, ..., m\}$  constitute a refined descriptive sample of size N of Y observed through simulation.

# 6.5. The density function of the arguments $Y_j$

Firstly, we notice that

$$P\left(\mathbf{X}_{\underline{i}}^{j} \in S_{\underline{i}}^{j}\right) = \frac{1}{p_{j}^{K}}$$

$$(6.5)$$

and the conditional density function of

$$\mathbf{X}_{\underline{i}}^{j} \text{ given } \left\{ \mathbf{X}_{\underline{i}}^{j} \in S_{\underline{i}}^{j} \right\} is \ p_{j}^{K} f(x)$$

$$(6.6)$$

it follows that,

$$\begin{split} P(y_{\underline{i}}^{j} \leq y) &= \sum_{A_{j}} P\left(y_{\underline{i}}^{j} \leq y/\mathbf{X}_{\underline{i}}^{j} \in S_{\underline{i}}^{j}\right) \times P\left(\mathbf{X}_{\underline{i}}^{j} \in S_{\underline{i}}^{j}\right) \\ &= \sum_{\underline{i}} \int_{S_{\underline{i}}^{j}/h(x) \leq y} p_{j}^{K} f(x) \mathrm{d}x \times \frac{1}{p_{j}^{K}} \\ &= \int_{A_{j}/h(x) \leq y} f(x) \mathrm{d}x. \end{split}$$

Then, distributions of  $Y_j$ , j = 1, ..., m are all the same as the distribution of **X**.

# 7. Properties of the estimator $T_{\rm RDS}$

### **Propriety 1.** $T_{\text{RDS}}$ is asymptotically unbiased estimator.

*Proof.* It has been shown in [18] that  $\operatorname{Bias}(T_{\mathrm{RDS}})$  is insignificant if the underlying frequency  $f_w$  of the output random variable Y is different from  $M \times \prod_{j=1}^m p_j$  where  $M \in N^*$ . This condition is usually verified because the simulation will certainly terminates before the product of all prime numbers used in a run or a multiple of it can be equal to the underlying frequency as this product  $M \times \prod_{j=1}^m p_j$  has a very high frequency.

We can deduce that if the sample size  $N \to +\infty$  then  $\prod_{j=1}^{m} p_j \to +\infty$  faster and  $M \times \prod_{j=1}^{m} p_j$  tends to infinity even faster and since  $f_w$  is a finite frequency, then  $f_w \neq M \times \prod_{j=1}^{m} p_j$ . As a consequence,  $\lim_{N\to+\infty} \text{Bias}(T_{\text{RDS}}) = 0$ then,  $T_{\text{RDS}}$  is asymptotically unbiased estimate.

**Propriety 2.**  $T_{\text{RDS}}$  is an unbiased estimator of  $\theta$ .

*Proof.* Let  $\mu_{\underline{i}}^{j}$  and  $(\sigma_{\underline{i}}^{j})^{2}$  be respectively the mean and variance of  $g(y_{\underline{i}}^{j})$  in the K dimensional hypercubes. Let  $\mu_{j}$  and  $\sigma_{j}^{2}$  be respectively the mean and variance of  $g(Y_{j})$  in the sub-runs defined by

$$\mu_j = \frac{1}{p_j^K} \sum_{\underline{i}} \mu_{\underline{i}}^j. \tag{7.1}$$

1) Let us calculate first the mathematical expected value of  $\hat{\theta}_j$ . Considering the formulae (6.2), we have,

$$E\left(\widehat{\theta}_{j}\right) = E\left(\frac{1}{p_{j}}\sum_{\underline{i}}w_{\underline{i}}^{j} \times g\left(y_{\underline{i}}^{j}\right)\right) = \frac{1}{p_{j}}\sum_{\underline{i}}E\left(w_{\underline{i}}^{j} \times g\left(y_{\underline{i}}^{j}\right)\right)$$

given the independence of  $w_{\underline{i}}^{j}$  and  $g(y_{\underline{i}}^{j})$  then,

$$E\left(\widehat{\theta}_{j}\right) = \frac{1}{p_{j}} \sum_{\underline{i}} E\left(w_{\underline{i}}^{j}\right) \times E\left(g\left(y_{\underline{i}}^{j}\right)\right)$$

according to the 1st property of  $w_i^j$ , it follows,

$$E\left(\widehat{\theta}_{j}\right) = \frac{1}{p_{j}} \sum_{\underline{i}} \frac{1}{p_{j}^{K-1}} E\left(g\left(y_{\underline{i}}^{j}\right)\right) = \frac{1}{p_{j}^{K}} \sum_{\underline{i}} \mu_{\underline{i}}^{j}$$

using the formulae (7.1), we write

$$E(\widehat{\theta}_j) = \mu_j. \tag{7.2}$$

2) Let us calculate now the mathematical expected value of  $T_{\text{RDS}}$ . Considering the 4th formulae, we obtain,

$$E(T_{\rm RDS}) = \frac{1}{N} \sum_{j=1}^{m} p_j E\left(\widehat{\theta}_j\right)$$

and using (7.2), we have

$$E(T_{\rm RDS}) = \frac{1}{N} \sum_{j=1}^{m} p_j \mu_j = E(g(Y)) = \theta.$$
(7.3)

We deduce then, that  $T_{\text{RDS}}$  is an unbiased estimator of  $\theta$ .

**Propriety 3.**  $\forall \underline{i} \neq \underline{t}$  we have

$$\operatorname{Var}(T_{\mathrm{RDS}}) = \operatorname{Var}(T_{\mathrm{SRS}}) + \frac{1}{N^2} \sum_{j=1}^m p_j (p_j - 1) \operatorname{Cov}\left(\mu_{\underline{i}}^j, \mu_{\underline{t}}^j\right).$$

*Proof.* Given that the distribution of  $Y_j$  is f(x) as shown in Section 5.5, then

$$\sigma_j^2 = \operatorname{Var}(g(Y_j))$$

$$= E\left(g\left(y_{\underline{i}}^j\right) - \mu_j\right)^2 \qquad (7.4)$$

$$= \int_{A_j} \left(g\left(y_{\underline{i}}^j\right) - \mu_j\right)^2 f(x) dx$$

$$= \sum_{\underline{i}} \int_{S_{\underline{i}}^j} \left(g\left(y_{\underline{i}}^j\right) - \mu_j\right)^2 f(x) dx \quad where \quad j = 1, \dots, m.$$

$$(7.5)$$

In refined descriptive sampling method, the variance of g(Y) is defined by

$$\sigma^{2} = \operatorname{Var}(g(Y)) = \frac{1}{N} \sum_{j=1}^{m} p_{j} \sigma_{j}^{2}.$$
(7.6)

Applying the independence of the sub-runs and using the 6.4th formulae, the variance of the general form of  $T_{\text{RDS}}$  is written by

$$\operatorname{Var}\left(T_{\mathrm{RDS}}\right) = \frac{1}{N^2} \sum_{j=1}^{m} \sum_{\underline{i}} \operatorname{Var}\left(w_{\underline{i}}^j \times g\left(y_{\underline{i}}^j\right)\right) + \frac{1}{N^2} \sum_{j=1}^{m} \sum_{\underline{i}} \sum_{\underline{t}, \underline{i} \neq \underline{t}} \operatorname{Cov}\left(w_{\underline{i}}^j \times g\left(y_{\underline{i}}^j\right), w_{\underline{t}}^j \times g\left(y_{\underline{t}}^j\right)\right)$$
(7.7)

141

1) Let us calculate the first part of the right hand side of the last equality. We first notice that,

$$\operatorname{Var}\left(w_{\underline{i}}^{j} \times g\left(y_{\underline{i}}^{j}\right)\right) = E\left(\left(w_{\underline{i}}^{j}\right)^{2}\right) \operatorname{Var}\left(g\left(y_{\underline{i}}^{j}\right) + E^{2}\left(g\left(y_{\underline{i}}^{j}\right)\right) \times \operatorname{Var}\left(w_{\underline{i}}^{j}\right).$$

By virtue of the 1st and 2nd probability properties of  $w^j_{\underline{i}},$  we have,

$$\sum_{\underline{i}} \operatorname{Var}\left(w_{\underline{i}}^{j} \times g\left(y_{\underline{i}}^{j}\right)\right) = \sum_{\underline{i}} \frac{1}{p_{j}^{K-1}} \operatorname{Var}\left(g\left(y_{\underline{i}}^{j}\right)\right) + \sum_{\underline{i}} \left(\mu_{\underline{i}}^{j}\right)^{2} \left(\frac{1}{p_{j}^{K-1}}\right) \left(1 - \frac{1}{p_{j}^{K-1}}\right).$$
(7.8)

By adding and reducing  $\mu_j$  in the expectation of  $(\sigma_{\underline{i}}^j)^2$  given by

$$\operatorname{Var}\left(g\left(y_{\underline{i}}^{j}\right)\right) = E\left(g\left(y_{\underline{i}}^{j}\right)\right) - \left(\mu_{\underline{i}}^{j}\right)^{2}$$

$$(7.9)$$

we obtain

$$\operatorname{Var}\left(g\left(y_{\underline{i}}^{j}\right) = E\left(g\left(y_{\underline{i}}^{j}\right) - \mu_{j}\right)^{2} - \left(\mu_{\underline{i}}^{j} - \mu_{j}\right)^{2}$$
$$= \int_{S_{\underline{i}}^{j}} \left(g\left(y_{\underline{i}}^{j}\right) - \mu_{j}\right)^{2} p_{j}^{K} f(x) \mathrm{d}x - \left(\mu_{\underline{i}}^{j} - \mu_{j}\right)^{2}.$$
(7.10)

Substituting the expression (7.10) in the relation (7.8), we write

$$\sum_{\underline{i}} \operatorname{var} \left( w_{\underline{i}}^{j} \times g\left( y_{\underline{i}}^{j} \right) \right) = \frac{1}{p_{j}^{K-1}} \left( \sum_{\underline{i}} \int_{S_{\underline{i}}^{j}} \left( g\left( y_{\underline{i}}^{j} \right) - \mu_{j} \right)^{2} p_{j}^{K} f(x) \mathrm{d}x \right)$$
$$-p_{j}^{-K+1} \sum_{\underline{i}} \left( \mu_{\underline{i}}^{j} - \mu_{j} \right)^{2}$$
$$+ \left( p_{j}^{-K+1} - p_{j}^{-2K+2} \right) \sum_{\underline{i}} \left( \mu_{\underline{i}}^{j} \right)^{2}.$$

Considering (7.5) we have then,

$$\frac{1}{N^2} \sum_{j=1}^m \sum_{\underline{i}} \operatorname{Var}\left(w_{\underline{i}}^j \times g\left(y_{\underline{i}}^j\right)\right) = \frac{1}{N^2} \sum_{j=1}^m p_j \sigma_j^2 - \frac{1}{N^2} \sum_{j=1}^m \left[p_j^{-K+1} \sum_{\underline{i}} \left(\mu_{\underline{i}}^j - \mu_j\right)^2 + \left(p_j^{-K+1} - p_j^{-2K+2}\right) \sum_{\underline{i}} \left(\mu_{\underline{i}}^j\right)^2\right]$$
(7.11)

and finally taking into account the formula (7.6), we get

$$\frac{1}{N^2} \sum_{j=1}^m \sum_{\underline{i}} \operatorname{Var}(w_{\underline{i}}^j \times g(y_{\underline{i}}^j)) = \frac{\sigma^2}{N} - \frac{1}{N^2} \sum_{j=1}^m \left[ p_j^{-K+1} \sum_{\underline{i}} (\mu_{\underline{i}}^j - \mu_j)^2 + \left( p_j^{-K+1} - p_j^{-2K+2} \right) \sum_{\underline{i}} \left( \mu_{\underline{i}}^j \right)^2 \right]$$
(7.12)

2) We calculate now the second part of the right hand side of the equality (7.7).

According to the covariance properties and the 1st properties of  $w_i^j$ , it follows that,

$$\sum_{\underline{i}} \sum_{\underline{t}, \underline{i} \neq \underline{t}} \operatorname{Cov} \left( w_{\underline{i}}^{j} \times g\left( y_{\underline{i}}^{j} \right), w_{\underline{t}}^{j} \times g\left( y_{\underline{t}}^{j} \right) \right) = \sum_{\underline{i}} \sum_{\underline{t}, \underline{i} \neq \underline{t}} \mu_{\underline{i}}^{j} \times \mu_{\underline{t}}^{j} \times E \left( w_{\underline{i}}^{j} \times w_{\underline{t}}^{j} \right) - \frac{1}{p_{j}^{2K-2}} \sum_{\underline{i}} \sum_{\underline{t}, \underline{i} \neq \underline{t}} \mu_{\underline{i}}^{j} \times \mu_{\underline{t}}^{j}$$

By virtue of the 3rd probability properties of  $w_{\underline{i}}^{j},$  we have

$$\frac{1}{N^2} \sum_{j=1}^m \sum_{\underline{i}} \sum_{\underline{t}, \underline{i} \neq \underline{t}} \operatorname{Cov}(w_{\underline{i}}^j \times g(y_{\underline{i}}^j), w_{\underline{t}}^j \times g(y_{\underline{t}}^j)) = \frac{1}{N^2} \sum_{j=1}^m \left[ (p_j - 1)^{-K+1} p_j^{-K+1} \sum_{\underline{i}} \sum_{\underline{t}, \underline{i} \neq \underline{t}} \mu_{\underline{i}}^j \mu_{\underline{t}}^j - p_j^{-2K+2} \sum_{\underline{i}} \sum_{\underline{t}, \underline{i} \neq \underline{t}} \mu_{\underline{i}}^j \mu_{\underline{t}}^j \right]$$
(7.13)

Substituting (7.12) and (7.13) in (7.7), we write,

$$\operatorname{Var}(T_{\mathrm{RDS}}) = \operatorname{Var}(T_{\mathrm{SRS}}) + \frac{1}{N^2} \sum_{j=1}^{m} \left[ (p_j - 1)^{-K+1} p_j^{-K+1} \sum_Z \sum_{j=1}^{k} \mu_{\underline{i}}^j \mu_{\underline{t}}^j - p_j^{-K+1} \sum_{\underline{i}} \left( \mu_{\underline{i}}^j - \mu_j \right)^2 + \left( p_j^{-K+1} - p_j^{-2K+2} \right) \sum_{\underline{i}} \left( \mu_{\underline{i}}^j \right)^2 - p_j^{-2K+2} \sum_Z \mu_{\underline{i}}^j \mu_{\underline{t}}^j \right].$$
(7.14)

Where Z means the restricted space of  $p_j^K (p_j - 1)^K$  pairs  $(\mu_{\underline{i}}^j, \mu_{\underline{t}}^j)$  corresponding to the hypercubes  $S_{\underline{i}}^j$  having no hypercube coordinates in common. After some algebra and using (7.1), we demonstrate that

$$\sum_{\underline{i}} \left(\mu_{\underline{i}}^{j} - \mu_{j}\right)^{2} = \sum_{\underline{i}} \left(\mu_{\underline{i}}^{j}\right)^{2} - \mu_{j}^{2} p_{j}^{K}$$
(7.15)

and

$$-p_{j}^{-K+1} \sum_{\underline{i}} \left(\mu_{\underline{i}}^{j} - \mu_{j}\right)^{2} + (p_{j}^{-K+1} - p_{j}^{-2K+2}) \sum_{\underline{i}} \left(\mu_{\underline{i}}^{j}\right)^{2} - p_{j}^{-2K+2} \sum_{Z} \sum_{Z} \mu_{\underline{i}}^{j} \mu_{\underline{t}}^{j}$$

$$= \mu_{j}^{2} p_{j} - p_{j}^{-2K+2} \left(\sum_{Z} \sum_{Z} \mu_{\underline{i}}^{j} \mu_{\underline{t}}^{j} + \sum_{\underline{i}} \left(\mu_{\underline{i}}^{j}\right)^{2}\right)$$

$$= \mu_{j}^{2} p_{j} - p_{j}^{2} p_{j}^{-2K} \left(\sum_{\underline{i}} \mu_{\underline{i}}^{j}\right) \left(\sum_{\underline{t}} \mu_{\underline{t}}^{j}\right)$$

$$= \mu_{j}^{2} p_{j} - p_{j}^{2} \mu_{j}^{2}.$$
(7.16)

Therefore, substituting (7.16) in (7.14), the Var $(T_{RDS})$  becomes

$$\operatorname{Var}(T_{\mathrm{RDS}}) = \operatorname{Var}(T_{\mathrm{SRS}}) + \frac{1}{N^2} \sum_{j=1}^{m} \left[ (p_j - 1)^{-K+1} p_j^{-K+1} \sum_Z \sum \mu_{\underline{i}}^j \mu_{\underline{t}}^j - \mu_j^2 p_j (p_j - 1) \right]$$
$$= \operatorname{Var}(T_{\mathrm{SRS}}) + \frac{1}{N^2} \sum_{j=1}^{m} p_j (p_j - 1) \left[ (p_j - 1)^{-K} p_j^{-K} \sum_Z \sum \mu_{\underline{i}}^j \mu_{\underline{t}}^j - \mu_j^2 \right]$$
(7.17)

Given that,

$$\operatorname{Cov}\left(\mu_{\underline{i}}^{j}, \mu_{\underline{t}}^{j}\right) = p_{j}^{-K}(p_{j}-1)^{-K} \sum_{Z} \sum \mu_{\underline{i}}^{j} \mu_{\underline{t}}^{j} - \mu_{j}^{2}$$
(7.18)

and substituting (7.18) in (7.17), we have

$$\operatorname{Var}(T_{\mathrm{RDS}}) = \operatorname{Var}(T_{\mathrm{SRS}}) + \frac{1}{N^2} \sum_{j=1}^m p_j (p_j - 1) \operatorname{Cov}\left(\mu_{\underline{i}}^j, \mu_{\underline{t}}^j\right).$$

**Theorem 7.1.** If  $Y = h(X_1, \ldots, X_K)$  is monotonic in each of its arguments and if g(Y) is monotonic function of Y, we have then  $\operatorname{Var}(T_{\mathrm{RDS}}) \leq \operatorname{Var}(T_{\mathrm{SRS}})$ 

*Proof.* To proof the theorem, we need to verify the conditions of Lehmann's theorem given in appendix [7]. Let us select two hypercubes defined by their center as given in Section 6.1, say,  $r_{\underline{i}}^{j} = (r_{i_{1}}^{j}, \ldots, r_{i_{K}}^{j})$  and  $r_{\underline{t}}^{j} = (l_{i_{1}}^{j}, \dots, l_{i_{K}}^{j})$  with no coordinates in common.

- 1. The pairs  $(r_{i_k}^j, l_{i_k}^j), k = 1, \ldots, K$  are mutually independent by definition
- 2. Each pair  $(r_{i_k}^j, l_{i_k}^j)$  is negatively quadrant dependent since:

$$P(r_{i_k}^j \le x, l_{i_k}^j \le y) = \frac{[xy - \min(x, y)]}{p_j(p_j - 1)} \le P(r_{i_k}^j \le x) P\left(l_{i_k}^j \le y\right)$$

where [.] represents the integer part function 3. Given that  $\mu_{\underline{i}}^{j} = E(g(y_{\underline{i}}^{j}))$  then,  $\mu_{\underline{i}}^{j} = \mu(r_{i_{1}}^{j}, \ldots, r_{i_{K}}^{j})$  and  $\mu_{\underline{t}}^{j} = \mu(l_{i_{1}}^{j}, \ldots, l_{i_{K}}^{j})$  are monotonic in each argument under the assumptions of the theorem.

The conditions are verified, so, using Lehmann's theorem, we conclude that  $\mu_i^j$  and  $\mu_t^j$  are negatively quadrant dependent defined by the following inequality:

$$P(\mu_{\underline{i}}^{j} \leq x, \mu_{\underline{t}}^{j} \leq y) \leq P(\mu_{\underline{i}}^{j} \leq x) P(\mu_{\underline{t}}^{j} \leq y) \ \forall \ x \ and \ y$$

and finally, using Hoeffding's equation, we have

$$\operatorname{Cov}(\mu_i^j, \mu_t^j) \le 0 \quad \forall \ \underline{i} \neq \underline{t} \ and \ j = 1, \dots, m.$$

Then, using the 3rd propriety of the estimator  $T_{\rm RDS}$ , given in Section 7, the result is then obtained, that is  $\operatorname{Var}(T_{\mathrm{RDS}}) \leq \operatorname{Var}(T_{\mathrm{SRS}})$ 

We conclude that RDS is better than SRS for estimating E(q(Y)) when the output variable Y =  $h(X_1,\ldots,X_K)$  is monotonic in each of its arguments and q(Y) is monotonic function of Y.

#### 8. A discussion about the efficiency of RDS algorithm

 $T_{\rm RDS}$  is shown to be unbiased and from the point of view of the variance, the last theorem shows that the RDS algorithm given in [18] is more accurate than the simple Monte Carlo algorithm when the output variable  $Y = h(X_1, \ldots, X_K)$  is monotonic in each of its arguments and q(Y) is monotonic function of Y. Nevertheless. the simple Monte Carlo algorithm is preferred, from the algorithmic point of view, because its computational complexity  $C(P) = t \times var(T_{SRS})$  may be less than the complexity of the RDS algorithm  $C(RDS) = t' \times var(T_{SRS})$  $var(T_{RDS})$  where t and t' are the mean time (or number of operations) required to compute one value of the random variable using respectively SRS and RDS. This mean time t' depends on the runtime for generating a prime number which depends on the value of RANDMAX defined in Section 4 that is on the integer to be tested for the primality. In addition, the computation of the set of regular numbers of size  $p_i$  requires  $2p_i$  unit of time and the randomization of the set of regular numbers requires  $4p_i$  unit of time. Then, basically, we can say that RDS algorithm is efficient since its running time is 0(P(N)) where P(N) is a polynomial of size N.

Moreover, a software component getRDS is designed to generate numbers using RDS method and it is fully tested for both criteria: independence and uniformity between 0 and 1 [10] then a refined descriptive sample can be considered as one in which observations are independent random variables and each one following the population distribution and as a consequence, the random behavior of an input stochastic variable will be then well represented by a refined descriptive sample, it must be then preferred.

## APPENDIX A.

Lehmann's theorem:

If 1)  $(X_1, Y_1), (X_2, Y_2), \ldots, (X_K, Y_K)$  are independent, 2)  $(X_i, Y_i)$  is negatively quadrant dependent for all i, 3)  $X = r(x_1, \ldots, x_K)$  and  $Y = s(y_1, \ldots, y_K)$  are monotonic in each argument, then (X, Y) is negatively quadrant dependent.

Hoeffding's equation:

$$\operatorname{Cov}(X,Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [P(X \le x, Y \le y) - P(X \le x)(P(Y \le y))] dxdy$$

TABLE A.3. Regular set and descriptive sample of size n = 31 for a negative exponential distribution with mean E(X) = 1.

i	$r_i = (i - 0, 5)/31$	$x_i = -\ln(1 - ri)$
1	0.016129032	0.016260521
2	0.048387097	0.049596941
3	0.080645161	0.084083117
4	0.112903226	0.1198012
5	0.14516129	0.156842471
6	0.177419355	0.195308752
7	0.209677419	0.235314087
8	0.241935484	0.276986783
9	0.274193548	0.320471895
10	0.306451613	0.365934269
11	0.338709677	0.413562318
12	0.370967742	0.463572739
13	0.403225806	0.516216472
14	0.435483871	0.571786324
15	0.467741935	0.630626824
16	0.5	0.693147181
17	0.532258065	0.759838555
18	0.564516129	0.831297519
19	0.596774194	0.90825856
20	0.629032258	0.991640169
21	0.661290323	1.082611947
22	0.693548387	1.182695406
23	0.725806452	1.293921041
24	0.758064516	1.419084184
25	0.790322581	1.562185028
26	0.822580645	1.729239112
27	0.85483871	1.929909808
28	0.887096774	2.181224236
29	0.919354839	2.517696473
30	0.951612903	3.028522096
31	0.983870968	4.127134385

#### M. OURBIH-TARI AND S. GUEBLI

i	$r_i = (i - 0, 5)/7$	$x_i = -\ln(1 - ri)$
1	0.071428571	0.074107972
2	0.214285714	0.241162057
3	0.357142857	0.441832752
4	0.5	0.693147181
5	0.642857143	1.029619417
6	0.785714286	1.540445041
$\overline{7}$	0.928571429	2.63905733

TABLE A.4. Regular subsets and descriptive sample of size prime number  $p_1 = 7$  for a negative exponential distribution with mean E(X) = 1.

TABLE A.5. Regular subsets and descriptive sample of size prime number  $p_2 = 11$  for a negative exponential distribution with mean E(X) = 1.

i	$r_i = (i - 0, 5)/11$	$x_i = -\ln(1 - ri)$
1	0.045454545	0.046520016
2	0.136363636	0.146603474
3	0.22727272727	0.257829109
4	0.318181818	0.382992252
5	0.409090909	0.526093096
6	0.5	0.693147181
7	0.590909091	0.893817876
8	0.681818182	1.145132304
9	0.772727273	1.481604541
10	0.863636364	1.992430165
11	0.954545455	3.091042453

TABLE A.6. Regular subsets and descriptive sample of size a prime number  $p_3 = 13$  for a negative exponential distribution with mean E(X) = 1.

i	$r_i = (i - 0, 5)/13$	$x_i = -\ln(1 - ri)$
1	0.038461538	0.039220713
2	0.115384615	0.122602322
3	0.192307692	0.2135741
4	0.269230769	0.313657559
5	0.346153846	0.424883194
6	0.423076923	0.550046337
7	0.5	0.693147181
8	0.576923077	0.860201265
9	0.653846154	1.060871961
10	0.730769231	1.312186389
11	0.807692308	1.648658626
12	0.884615385	2.159484249
13	0.961538462	3.258096538

#### References

- [1] I.T. Dimov, Monte Carlo methods for applied scientists. World Scientific Publishing Co. Pte, Ltd Singapore (2008).
- [2] G.S. Fishman, Monte-Carlo Concepts Algorithms and Applications. Springer-Verlag, Berlin (1997).
- [3] M. Grigoriu, A spectral-based Monte Carlo algorithm for generating samples of nonstationary Gaussian processes. Monte Carlo Methods Appl. 16 (2010) 143–165.

- [4] A. Keller, S. Heinrich and H. Niederreiter, Monte Carlo and Quasi Monte Carlo Methods 2008. Springer (2008).
- [5] E.L. Lehmann, Some concepts of dependence. Ann. Math. Statist. 35 (1966) 1137-1153.
- [6] W.L. Loh, On Latin hypercube sampling. Ann. Statist. 24 (1996) 2058–2080.
- [7] M.D. MacKay, R.J. Beckman and W.J. Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21 (1979) 239-245.
- [8] H. Niederreiter, Random number generation and quasi monte Carlo methods. Society for Industrial and Applied Mathematics, Philadelphia (1992).
- M. Ourbih-Tari and A. Aloui, Sampling methods and parallelism into Monte Carlo simulation. J. Statist. Adv. Theory Appl. 1 (2009) 169–192.
- [10] M. Ourbih-Tari, A. Aloui and A. Alioui, A software component which generates regular numbers from refined descriptive sampling, in *Proc. of the European Simulation Modelling conference*, edited by Marwan Al-Akaidi. Leicester United Kingdom (2009) 23–25.
- [11] M. Pidd, Computer simulation in management science, 5 edn. John Wiley and Sons (2004).
- [12] J.S. Ramberg and B.W. Schmeiser, An Approximate Method for generating Symmetric Random Variables. Communications of the ACM 15 (1972) 987–990.
- [13] E. Saliby, Descriptive Sampling A better approach to Monte Carlo simulation. J. Oper. Res. Soc. 41 (1990) 1133–1142.
- [14] T.L. Schmitz and H.S. Kim, Monte Carlo evaluation of periodic error uncertainty. Precision Engrg. 31 (2007) 251-259.
- [15] I.M. Sobol, A primer for the Monte Carlo method. CRS Press (1994).
- [16] M. Tari and A. Dahmani, Flowshop simulator using different sampling methods. Oper. Res. Inter. J. 5 (2005) 261–272.
- [17] M. Tari and A. Dahmani, The three phase discrete event simulation using some sampling methods. Int. J. Appl. Math. Statist. 3 (2005) 37–48.
- [18] M. Tari and A. Dahmani, Refined Descriptive Sampling A better approach to Monte Carlo simulation. Simul. Model. Pract. Theory 14 (2006) 143–160.