Available online at www.sciencedirect.com
SCIENCE DIRECT®
Simulation Modelling Practice and Theory 14 (2006) 143–160
SIMULATION MODELLING PRACTICE AND THEORY
www.elsevier.com/locate/simpat

Refined descriptive sampling: A better approach to Monte Carlo simulation

Megdouda Tari *, Abdelnasser Dahmani

Laboratory of Applied Mathematics, Department of Mathematics, University of Bejaia, Algeria

Received 2 June 2004; received in revised form 1 March 2005; accepted 1 April 2005
Available online 8 June 2005

Abstract

Descriptive sampling could lead to biased results and require prior knowledge of the sample size. This paper analyses the conditions under which bias can occur and proposes an approach which is mainly concerned with a block of regular samples of prime size. This approach reduces the sampling bias and eliminates the problem of descriptive sampling related to the sample size. We evaluate performance measures of a problem whose response through simulation is a wave with regular period.

Keywords: Sampling; Variance reduction; Monte Carlo

1. Introduction

Descriptive sampling (DS) [13,14] is based on a deterministic selection of the input sample values and their random permutation. This method is known to have two problems: it can be biased, and its strict operation requires a prior knowledge of the sample size [10]. The concept of sample size using either random sampling (RS) and DS is stressed in [16].

* Corresponding author.
E-mail address: megatari@yahoo.fr (M. Tari).

1569-190X/$ - see front matter
doi:10.1016/j.simpat.2005.04.001

Several empirical comparisons on a PERT network, an M/M/1 queue and on an inventory system [14] show that the estimates of the output random variables parameters produced through simulation using the DS method are with lower variance than those obtained by the Monte Carlo (MC) method [1,12]. There is a discussion about the bias in [14] mentioning its small magnitude but no study is yet available on the bias that DS produces. As far as RS is used in simulation, the problem of biased estimates is not a matter of concern since the use of RS produces unbiased estimators.

Descriptive sampling and Latin Hypercube Sampling (LHS) [7] are both based on a random permutation of the input numbers but selection their values differently. In [17], there is a discussion suggesting that DS has a smaller variance than LHS, but no mention is made about the bias it produces. However, it is well known that LHS is unbiased. Hence, from the point of view of the mean square error, it is not clear that DS is better than both LHS and RS. In general, unbiased estimators are always preferred. The LHS estimates are also compared to Monte Carlo estimates in [19] (see [9,5] for more work concerning LHS).

Quasi Monte Carlo (QMC) methods are concerned with deterministic streams [8]. Streams of low discrepancy are used to reduce the variance in MC methods. A comparison between MC and QMC simulations was carried out [22] and led to a combined method which can be considered as a random permutation of QMC methods. Some theoretical results on QMC can be found, for example, in [2,6,23].

The DS method still lacks an adequate theoretical development and this limits its applicability. It is therefore important to carry out further studies on this method. In order to compare it with other methods, we need to show that it produces unbiased estimates or at least reduces significantly the risk of bias.

This paper improves descriptive sampling by reducing the bias introduced by non-random sampling. Our approach may be considered as a variance reduction technique similar to Latin Hypercube Sampling and QMC methods. The analysis is based on some stylized input–output transformations (more specifically, on sinusoidal curves) which allow to determine the types of problems where DS encounters the most bias. We also propose an improvement called the refined descriptive sampling (RDS) to safeguard against the possible occurrence of such problems. Mathematical arguments for the new approach are presented in this paper and a proof of its efficiency is given. A study of the proposed approach is also carried out in [20,21].

2. Problem formulation

A logical model is built in a simulation study and is used as a vehicle for experimentation. The model is illustrated in Fig. 1.

The distributions of input variables are assumed to be known, while the response variables' distributions are unknown. When the problem is simulated, input random variables are replaced by samples. As a result, response variables are also replaced by

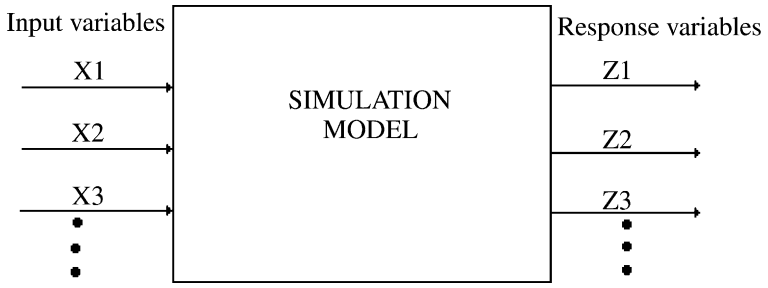

Fig. 1. Simulation model representation: a set of input random variables is transformed into a set of output random variables.

samples. So experiments are carried out on the model and unknown parameters of the response random variables of interest are estimated.

We assume, for simplicity, that only one input random variable drives the simulation and one response variable with k parameters to be estimated is observed through the simulation.

In a given run, the use of RS procedure leads to the following estimates of the parameters θ_j, $j = 1, 2, \ldots, k$:

$$Y_j = F_j(u_1, u_2, \ldots, u_n) \quad j = 1, 2, \ldots, k$$

where F_j, $j = 1, 2, \ldots, k$ is a set of the simulation functions and $u_1, u_2, \ldots, u_n$ are independent random numbers uniformly distributed over the range 0 and 1.

The surface represented by $F_j(u_1, u_2, \ldots, u_n)$, is called the jth response surface.

Using RS, two types of variations are present in a randomly generated input sample: one is related to the set of values and the other to their sequence [3]. The simulation estimates (which is a function of the input values) is affected by sampling errors, and the random behaviour of an input stochastic variable is not be well represented. Therefore, the simulation estimates vary between different runs even if the model remains unchanged. Thus, as it was stressed by [13,15], the variability of the simulation estimates are subject to RS procedure.

DS entails a full control over the input set of sample values. As such, it avoids the set variability in simulation studies [13,15] but keeps the sequence variability. So the resulting estimates are more precise than those of random sampling. The input random variable is then well represented.

In a given run, the use of DS procedure leads to the following estimates of the parameters θ_j, $j = 1, 2, \ldots, k$:

$$Y_{rj} = F_j(r_1, r_2, \ldots, r_n) \quad j = 1, 2, \ldots, k$$

where $r_1, r_2, \ldots, r_n$ are dependent regular numbers uniformly distributed over the range 0 and 1.

The surface represented by $F_j(r_1, r_2, \ldots, r_n)$, is called the jth regular response surface.

3. The use of descriptive sampling

3.1. Descriptive sampling

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 [11]. Formally, in descriptive sampling when the sample size n is known, the set values are defined for the input random variable X using the inverse transform method, such as

$$xd_i = H^{-1}(r_i) \quad \text{for } i = 1, 2, \ldots, n$$

where $H^{-1}(r)$, $r \in [0, 1]$ is the inverse transform for the input distribution and the dependent regular numbers r_i, $i = 1, 2, \ldots, n$ are defined by

$$r_i = \frac{i - 0.5}{n} \quad i = 1, 2, \ldots, n$$

Using DS, set values are generated in advance and stored in memory for later use. A descriptive sample is defined as one in which observations are not fully independent random variables, but each one following the population distribution. For such sampling, a supply of independent samples uniformly distributed over the range 0 and 1 is needed. For this reason, a congruential generator is used to randomize the set of input values. The descriptive sample is then, the set of input values taken in a random sequence.

To complete the DS generation process, the set of input values is used in a random sequence in each simulation run. Unlike the RS, the set of input values are the same for all replicated runs in the simulation.

3.2. Problem of bias

The bias problem in descriptive sampling is due to the fact that a regular sampling grid is used to select the input values. If the input function is periodic, there is a risk that the set of output values will be biased as they may hit the same point on the cycle. However, if the input function is not periodic, when sampling from unimodal probability distributions there is no risk of sampling bias.

The aim of descriptive sampling is to represent truly the population distribution from which it is generated so that any simulation estimates drawn from it can be safely implemented in the real system. There are two ways in which we can preserve regular sampling. We can sample at a high frequency, i.e. increase the sample size, and thereby increase the cost of the experiment, or refine the regular sampling procedure.

3.3. Description of the proposed refinement

To reduce the risk of bias, we propose an approach which is concerned with a block that must be situated inside a generator aiming to distribute regular subsets

of prime number sizes $p_q, p_{q+1}, \ldots$ for any q in a random order as required by the simulation. We stop when the simulation terminates.

In this approach, each run is determined by a block of different prime numbers. If M replicated runs are needed, it is necessary to consider M blocks of $m_1, m_2, \ldots, m_M$ regular subsets. The prime numbers and the input subsets values are not the same for all replicated runs. This approach removes the need to determine in advance the sample size.

3.4. Bias reduction

By definition, in each simulation run, the simulation delivers one response surface for each parameter under study. If the input function is periodic and high bias is generated for any subset, then, it will not be present for other subsets considered in a block. Furthermore, the use of regular subsets of prime number size in the proposed approach, ensure that the combination of prime numbers does not have a frequency that is a function of the frequency of the input distribution. This should ensure that the proposed refinement reduces the sampling bias even if the input function is periodic.

4. Model building of the response surface

We consider, without loss of generality, the simulation problem described in Section 2 in the case when the response random variable has only one parameter θ to be estimated. In general, in a given simulation problem, the bias of an estimate is defined by

$$\text{Bias}(Y_r) = E(Y_r) - \theta$$

The problem of bias in DS is caused by the use of a regular grid to select input values. As a consequence, if the response surface observed through simulation has regularities which match those of regular numbers, bias can achieve its maximum level.

We know that the response surface can have several components, so taking X and Z to represent the horizontal and vertical line respectively, the most suitable regular response surface chosen to provide a high bias is described as follows. The first component is a straight line defined by the equation $Z = 0$ and the second component is a wave with a regular period and a big amplitude which is located on the first component.

In this case, $\theta = 0$ and

$$\text{Bias}(Y_r) = E(Y_r) \tag{4.1}$$

We now define the suggested models related to the behaviour of these components. Let a be the wavelength and d the distance between two consecutive regular numbers. Consider the underlying frequency f_w to be the wave frequency, and the sampling frequency f_s to be the n regular numbers r_i, $i = 1, 2, \ldots, n$.

Definition 1. We call worst model, the described regular response surface with

$$f_w = kf_s \quad \text{for any } k = 1, 2, 3, \ldots$$

or

$$d = ka \quad \text{for any } k = 1, 2, 3, \ldots$$

and the first regular number generated must either be on the wave peak or on the wave trough. Illustration of the worst model for $k = 1$ is given in Fig. 2.

Definition 2. We call best case 1, the case where the described regular response surface has regularities which do not match those of the regular numbers, that is

$$d = \frac{(2k+1)a}{2} \quad \text{where } k = 0, 1, 2, \ldots$$

This case is shown in Fig. 3 where $k = 0$.

Definition 3. We call best case 2, the case where the described regular response surface has regularities which match those of the regular numbers, as in the worst model, but the first regular number generated is on the horizontal axis. This case is illustrated in Fig. 4 for $k = 1$.

5. Bias evaluation for descriptive sampling

In this section, we evaluate the bias of the estimate in the worst model using DS, and prove that it is of the same magnitude as the amplitude of the wave. We also show mathematically that bias does not exist in the best cases.

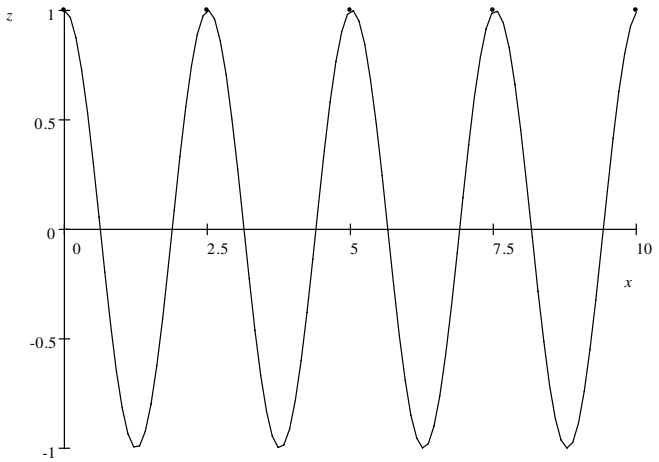

Fig. 2. Worst model for $k = 1$ when the 1st regular number is on the wave peak.

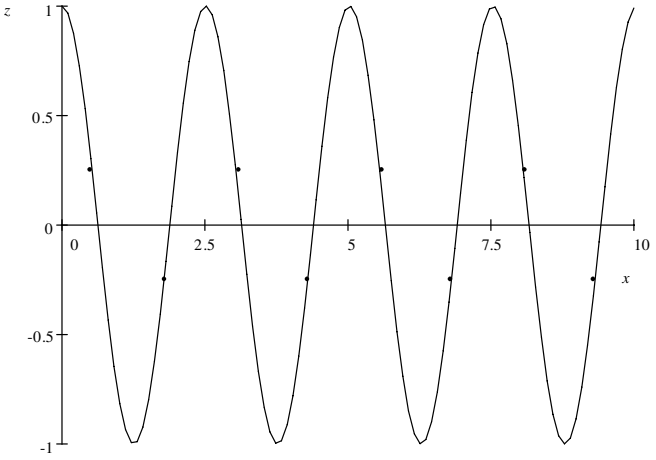

Fig. 3. Best case 1 for $k = 0$.

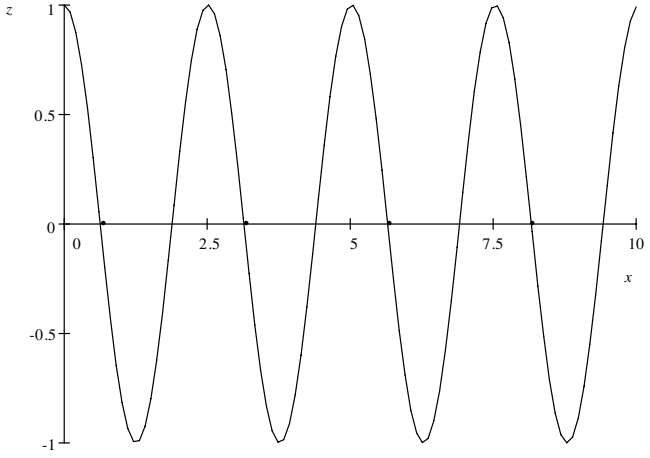

Fig. 4. Best case 2 for $k = 1$.

Theorem 1. *Let A be the amplitude of the wave. In the worst model we have*

$$\mathrm{Bias}(Y_r) = A$$

Proof. Suppose that

(1) the first regular number is situated on the wave peak, and
(2) the underlying frequency is close to sampling frequency by a real number c, $c \geqslant 0$. Let a increase or decrease by c (i.e going from a to $a + c$ or from a to $a - c$). The movement of the underlying frequency f_w to the sampling frequency f_s is then $d = a + c$ or $d = a - c$. We evaluate the bias as f_w moves to f_s.

The moving regular response from its equilibrium value is given by

$$G(x) = A\cos(\omega x + L)$$

where ω is the circular frequency and L is the phase.

To simplify the calculation, we take $A = 1$ and $L = 0$. Then we have

$$G(x) = \cos(\omega x)$$

In order to evaluate bias we consider the wave function given by Dirichlet's Kernel expression $D_N(x)$ from Fourier Analysis in [4]

$$D_n(x) = 1 + 2\sum_{k=1}^{n} \cos(kx) \tag{5.1}$$

When $x = 0$, we have

$$D_n(0) = 2n + 1$$

Using (4.1), the bias can be expressed as follows:

$$\text{Bias}(Y_r) = E(G(x)) = \frac{1}{n}\sum_{k=1}^{n} G(k) = \frac{1}{n}\sum_{k=1}^{n} \cos(\omega k)$$

where n is the input sample size.

The quantity c is defined as the circular frequency ω so

$$\text{Bias}(Y_r) = \frac{1}{n}\sum_{k=1}^{n} \cos(kc)$$

It follows from (5.1) that

$$\text{Bias}(Y_r) = \frac{1}{n}\left(\frac{D_n(c) - 1}{2}\right) \tag{5.2}$$

Consequently, for $c = 0$, i.e. when the underlying frequency equals the sampling frequency, we have

$$\text{Bias}(Y_r) = 1$$

Note that if we assume that the first input number is the minimum of the wave, the bias will be equal to -1. □

Theorem 2. *In the best case* 1 *we have*

$$\text{Bias}(Y_r) = \circ\left(\frac{1}{n}\right).$$

Proof. A result in [18] shows that

$$D_n(x) = \frac{\sin\left(n + \frac{1}{2}\right)x}{\sin\frac{x}{2}} \quad \text{for } x \neq 0$$

Using this result, (5.2) becomes

$$\mathrm{Bias}(Y_r) = \frac{\sin\left(n+\frac{1}{2}\right)c}{2n\sin\frac{c}{2}} - \frac{1}{2n}$$

Suppose that the 2nd assumption in the proof of Theorem 1 holds. Then, for $c > 0$,

$$\mathrm{Bias}(Y_r) = \circ\left(\frac{1}{n}\right) \tag{5.3}$$

In particular, when $c = d = \frac{a}{2} > 0$, i.e. when the underlying frequency equals half the sampling frequency, we obtain (5.3). Consequently, the bias is irrelevant as simulation samples are typically large. □

Theorem 3. *In the best case* 2 *we have*

$$\mathrm{Bias}(Y_r) = 0.$$

Proof. We assume that the 2nd assumption in the proof of Theorem 1 holds and suppose that the first regular number is on the horizontal axis. To satisfy the latter condition we consider the following moving response:

$$G(x) = \sin(\omega x)$$

Using (4.1), the bias is given by

$$\mathrm{Bias}(Y_r) = \frac{1}{n}\sum_{k=1}^{n}\sin(kc)$$

Thus, for $c = 0$, we have

$$\mathrm{Bias}(Y_r) = 0 \qquad \square$$

Remark 1. In most simulation problems, the response surface might have the following two components:

- A wave form with big amplitude.
- A wave with a regular period but small amplitude situated on the first component.

This sort of response surface may be regarded, if we pull both edges of the first component, as a response surface composed of a wave with a regular period and *small amplitude* situated on a straight line. It is therefore clear that the produced simulation estimates have small bias since these depend on the amplitude of the wave in the worst case model.

6. The use of the proposed approach

6.1. Simulation run

Let p_q, $q = 1, 2, 3, \ldots$ be distinct prime numbers. Suppose that:

- the response variable has k parameters to be estimated,
- the simulation terminates when m prime numbers have been used, which derives m sub-runs.

In a given run, the use of RDS procedure leads to the following m estimates of the parameters θ_j, $j = 1, 2, \ldots, k$

$$Yr_j^1 = F_j(r_1^1, r_2^2, \ldots, r_{p_1}^{p_1}) \quad \text{for } j = 1, 2, \ldots, k$$
$$Yr_j^2 = F_j(r_{1+p_1}^1, r_{2+p_1}^2, \ldots, r_{p_2+p_1}^{p_2}) \quad \text{for } j = 1, 2, \ldots, k$$
$$Yr_j^m = F_j\left(r_{1+\sum_{i=1}^{m-1} p_i}^1, r_{2+\sum_{i=1}^{m-1} p_i}^2, \ldots, r_{\sum_{i=1}^{m} p_i}^{p_m}\right) \quad \text{for } j = 1, 2, \ldots, k$$

conventionally

$$\sum_{i=1}^{0} p_i = 0$$

and

$$\left(r_{1+\sum_{i=1}^{q-1} p_i}^1, r_{2+\sum_{i=1}^{q-1} p_i}^2, \ldots, r_{\sum_{i=1}^{q} p_i}^{p_q}\right) \quad \text{for } q = 1, 2, \ldots, m$$

are considered as subsets of dependent regular numbers of prime size $p_1, p_2, \ldots, p_m$ that are uniformly distributed between $[0, 1]$, and are obtained by the following formula (as in DS method)

$$r_{i+\sum_{i=1}^{q-1} p_i}^i = \frac{i - 0.5}{p_q} \quad \text{for } i = 1, 2, \ldots, p_q \text{ and } q = 1, 2, \ldots, m.$$

Therefore, in a given run, the use of RDS procedure leads to the following sampling estimates of θ_j, $j = 1, 2, \ldots, k$, defined by the average of these estimates

$$Yr_j = \frac{1}{m} \sum_{i=1}^{m} Yr_j^i \quad \text{for } j = 1, 2, \ldots, k$$

6.2. Subset values generation

Using RDS, subset values for the input random variable X are generated as required by the simulation. The general method of the inverse transform can produce regular subset values given by

$$xd_{i+\sum_{i=1}^{q-1} p_i}^i = H^{-1}\left(r_{i+\sum_{i=1}^{q-1} p_i}^i\right) \quad \text{for } i = 1, 2, \ldots, p_q \text{ and } q = 1, 2, \ldots, m$$

Table 1
Regular subsets of prime numbers size for a negative exponential distribution with mean $E(X) = 1$, $p_1 = 7$, $p_2 = 11$ and $p_3 = 13$

i	r_i	xd_i	i	r_i	xd_i	i	r_i	xd_i
1	0.071	0.074	1	0.045	0.047	1	0.038	0.039
2	0.214	0.241	2	0.136	0.147	2	0.115	0.123
3	0.357	0.442	3	0.227	0.248	3	0.192	0.214
4	0.500	0.693	4	0.318	0.383	4	0.269	0.314
5	0.643	1.030	5	0.409	0.526	5	0.346	0.425
6	0.786	1.540	6	0.500	0.693	6	0.423	0.550
7	0.929	2.639	7	0.591	0.894	7	0.500	0.693
	Mean	0.951	8	0.682	1.145	8	0.577	0.860
			9	0.773	1.482	9	0.654	1.061
			10	0.864	1.992	10	0.731	1.312
			11	0.955	3.091	11	0.808	1.649
				Mean	0.969	12	0.885	2.159
						13	0.962	3.258
							Mean	0.974

Table 2
The observed mean of a negative exponential distribution using three regular subsets of prime number size

	p_1	p_2	p_3	Overall mean
Mean	0.951	0.969	0.974	0.965

We illustrate the method in Tables 1 and 2 by taking $p_1 = 7$, $p_2 = 11$ and $p_3 = 13$ for a negative exponential distribution with mean of 1 obtained by $xd_i = -\ln(1 - r_i)$.

Therefore, in a given run and using RDS, the estimate of the mean and its bias are obtained as follows:

$$y_r = \frac{1}{3}\sum_{i=1}^{3} y_r^i = 0.965$$

$$\mathrm{Bias}(y_r) = \frac{1}{3}\sum_{i=1}^{3} \mathrm{Bias}(y_r^i) = -0.035$$

For this simulation run, if DS is used, we take a sample of size $n = \sum_{i=1}^{3} p_i = 31$.

6.3. Descriptive sub-sets

The descriptive sub-sets of prime number size p_q are obtained by:

(1) generating the subsets of regular numbers

$$\left(r^1_{1+\sum_{i=1}^{q-1} p_i},\ r^2_{2+\sum_{i=1}^{q-1} p_i}, \ldots,\ r^{p_q}_{\sum_{i=1}^{q} p_i}\right)$$

of size a prime number p_q defined in Section 6.1,

(2) randomizing their sequence for any p_q, $q = 1, 2, \ldots, m$ randomly chosen,
(3) computing the regular subsets values of the input random variable $xd^i_{i+\sum_{i=1}^{q-1} p_i}$ for $i = 1, 2, \ldots, p_q$ as required by the simulation.

This type of generation is motivated by the fact that the simulation can terminate before the complete use of the last descriptive sub-set. This reduces the cost of the experiment since we compute only a portion of the last subset values generated from the last prime number p_m.

6.4. Data structure

For each input random variable, we define a record with the following structure:

p	a prime number defining the size of the regular numbers subset;
R	array $[1 \ldots p]$ of real numbers containing the subset of regular numbers;
ip	integer pointing to the first available r element to be drawn. If $ip = 1$, no element has been drawn yet. If $ip > p$, a full subset of regular numbers has already been drawn.

6.5. Algorithm

(a) Initialization for the experiment.
  (a_1) Randomly generate a prime number.
  (a_2) Generate the subset of regular numbers r_i, $i = 1, 2, \ldots, p$ and store them in an array R.
(b) Initialization for the sub-run. At the beginning of every sub-run, let $ip := 1$.
(c) Sampling without replacement during the sub-run:
  (c_1) if $ip > p$ then go to (d)
  (c_2) randomly generate an integer $iaux \in [ip, p]$
  (c_3) interchange $r(ip)$ with $r(iaux)$
  (c_4) generate one observation xd_i.
    If a descriptive sub-set value is not required stop and collect the final results from the last prime number used and go to (e)
  (c_5) otherwise let $ip := ip + 1$ and go to (c_1).
(d) Collect the results after each sub-run and go to (a_1).
(e) Collect the results after each run.

6.6. Bias evaluation

According to Section 4, if the response surface is a wave with a regular period and big amplitude and if the first regular number from each subset is on the wave peak or on the wave trough, then to generate high bias, the underlying frequency must be

$$f_w = n_q p_q \quad \forall q = 1, 2, \ldots, m$$

where n_q is a suitable integer and p_q, $q = 1, 2, \ldots, m$ are the prime numbers used in a run.

That is, if

$$f_w = Mp_1p_2 \ldots p_m \quad \text{for any integer } M$$

then, bias can arise in RDS. Since the product of all prime numbers used in a run has a very high frequency, the simulation will certainly terminate before the latter can be equal to the underlying frequency.

Theorem 4. *If*

$$f_w \neq Mp_1p_2 \ldots p_m \quad \text{where } M = 1, 2, \ldots$$

then Bias(*Yr*) *is insignificant.*

Proof. We consider, without loss of generality, the following example where $f_w = 30$.

- Suppose that the prime numbers chosen are $p_1 = 5$ and $p_2 = 2$. Then, f_w is a multiple of $p_1$and p_2 where the multiple integers are $n_1 = 6$ and $n_2 = 15$. Thus, all regular numbers from p_1 to p_2 are situated on the wave peaks if each first regular number generated from each prime number is situated on a wave peak.

  Consequently, bias achieves its maximum level since

  $$f_w = Mp_1p_2 \quad \text{where } M = 3$$
  $$f_w > p_1p_2$$

- Suppose now that an extra prime number $p_3 = 3$ is required by the simulation. Since

  $$f_w = n_3p_3 \quad \text{where } n_3 = 10$$

  the regular numbers from p_3 are also on wave peaks if the first regular number generated from p_3 is on a wave peak.

  As a result, bias also achieves its maximum level as

  $$f_w = p_1p_2p_3$$

- In the same way, suppose that another prime number $p_4 = 7$ is required by the simulation. As f_w is not a multiple of p_4, the regular numbers from p_4 are not situated on wave peaks. Hence, bias becomes smaller than its maximum since

  $$f_w < p_1p_2p_3p_4$$

It is clear that bias becomes insignificant as a run is composed of more and more prime numbers different from p_1, p_2 and p_3.

In general, bias is insignificant if the product of prime numbers used in a run composed by $p_1, p_2, \ldots, p_m$ is greater than the underlying frequency, i.e. if

$$f_w < p_1p_2 \cdots p_m \tag{6.1}$$

Suppose now that the early prime numbers required by the simulation are not all divisors of f_w, for instance, when $p_1 = 7$ and $p_2 = 2$. In this case, bias is insignificant since

$$f_w \neq Mp_1p_2 \quad \text{where } M = 2, 3, \ldots$$
$$f_w > p_1p_2$$

In general, bias is irrelevant if

$$f_w > p_1p_2 \cdots p_m \tag{6.2}$$

where $p_1p_2 \cdots p_m$ is not a divisor of f_w. We conclude the proof by using (6.1) and (6.2). □

7. Experiment and empirical results

7.1. Experiment

To check the proposed approach, we consider a game whose response surface behaves as a wave with a regular period. At the beginning, the performance capability of the player is evaluated using a function X. After much apparent effort, a performance directly related to X is delivered as the answer for a given run. In this problem, there is one input variable and one response variable with two parameters μ and σ to be estimated.

(1) The input variable has the following observations

$$x_i = \sum_{j=1}^{5} A_j \times \cos(2\pi\omega_j u_i) \quad \text{for } i = 1, 2, \ldots, n$$

where A_j, $j = 1, 2, \ldots, 5$ are the amplitudes, ω_j, $j = 1, 2, \ldots, 5$ are the circular frequencies and $u_1, u_2, \ldots, u_n$ are uniformly distributed between $[0, 1]$ and

- are dependent regular numbers when the usual descriptive sampling is used,
- are considered as subsets of dependent regular numbers of prime size when the proposed approach is used, or
- are independent random numbers when random sampling is used.

(2) The response variable has the following observations:

$$f_i = \sum_{j=1}^{5} A_j \times \cos(2\pi\omega_j u_i) + 25 \quad \text{for } i = 1, 2, \ldots, n.$$

The theoretical values of the parameters under study are

$$\mu = 25 \quad \text{and} \quad \sigma = \left(\frac{1}{2}\sum_{j=1}^{5} A_j^2\right)^{1/2}$$

The estimates of μ and σ are respectively given by

$$\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i \quad \text{and} \quad S = \left(\frac{1}{n-1}\sum_{i=1}^{n}(f_i - \overline{f})^2\right)^{1/2}$$

7.2. Comparison between DS and RDS

To compare these methods, we chose a simple case where estimates are not affected by the input sample sequence. Therefore, there is no need to permute the input regular numbers nor to replicate simulation runs. Then, one run is carried out. We summarize each experiment by computing the mean, variance and bias of the estimates (we observe that $\sigma = 26.823$). In Section 6, the prime numbers must be chosen randomly. Here, the prime numbers are chosen as selected in ascending order as this does not affect the results. The first prime number used to generate the regular numbers is 7. Any prime number below 7 has too many multiples and therefore the wave frequencies could be chosen to be a multiple of these particular primes. The amplitudes A_j, $j = 1, 2, \ldots, 5$ of the waves are chosen arbitrary. For example, $A = (10, 5, 20, 17, 25)$.

In the first and second experiments, the sampling frequency is taken to be equal to the first wave frequency. For the remaining ones, we take $\omega_2 = 50$, $\omega_3 = 40$, $\omega_4 = 60$, $\omega_5 = 30$.

For the first experiment, we set $n = \omega_1 = 90$. The last prime number generated is 23 and the sum of all prime numbers is 90. The observed results are listed in Table 3.

For the second experiment, we take $n = \omega_1 = 100$. The last prime number generated is 29. Table 4 summarizes the observed results.

For the third experiment, ω_1 was set equal to the product of all prime numbers used. We consider a set of $n = 119$ observations, so, $\omega_1 = 215\,656\,441$ and the remaining sampling frequencies are chosen to be respectively $\omega_2 = 10$, $\omega_3 = 20$, $\omega_4 = 30$ and $\omega_5 = 40$. Note that all regular numbers from the last prime number generated were used by the simulation. The summarized results for the new approach are given in Table 5.

Tables 3 and 4 suggest that the bias of the estimates is insignificant when using the proposed approach and highly biased when using descriptive sampling. A small difference is observed between the results drawn in both tables. We conclude that the

Table 3
Empirical results when all regular numbers generated from the last prime number were used by the simulation

Estimator	DS			RDS		
	Mean	Variance	Bias	Mean	Variance	Bias
$\overline{f}$	15	0	−10	25	0	0
S	12.09	0	−14.74	28.47	0	1.65

Table 4
Empirical results when only a portion of the regular numbers generated from the last prime number were used

Estimator	DS			RDS		
	Mean	Variance	Bias	Mean	Variance	Bias
$\overline{f}$	15	0	−10	24.80	0	−0.20
S	17.89	0	−8.93	29.22	0	2.40

Table 5
Empirical results showing that the only wave frequency capable of creating high bias is the frequency generated by the product of all prime numbers

Estimator	Mean	Variance	Bias
$\overline{f}$	15	0	−10
S	26.46	0	0.36

approach developed in this paper is better if all regular numbers generated from the last prime number are used by the simulation.

It can be seen from Table 5 that the use of our approach produces highly biased mean estimate but an unbiased standard deviation estimate. The bias of the mean estimate has a magnitude of the first amplitude A_1 as expected. Therefore, the results from this run are in line with the argument made in Section 6.6.

7.3. Comparison between RS and RDS

RDS is a sampling procedure which reduces bias, so to compare it with RS we carried out M replicated runs for both methods. We summarize the experiment by computing the mean and variance of the estimates. In RDS, the prime numbers are randomly selected as stressed in Section 6. The results are given in Table 6 for $n = 100$, $\omega_1 = 10$, $\omega_2 = 20$, $\omega_3 = 30$, $\omega_4 = 40$, $\omega_5 = 50$.

As shown in Table 6, the refined descriptive sampling gives lower variance than random sampling.

Table 6
Empirical results showing the efficiency of our approach versus RS

M	Estimator	RS		RDS	
		Mean	Variance	Mean	Variance
10	$\overline{f}$	25.464	6.91	25	0
	S	27.693	5.074	28.910	0.040
100	$\overline{f}$	25.285	7.927	25	0
	S	26.982	4.943	28.783	0.037

8. Conclusion

We showed that descriptive sampling can produce sampling bias and proposed a refinement which guarantees the insignificance of bias and removes the need to determine in advance the sample size. We gave a mathematical argument together with a proof of the efficiency of the new approach by studying a problem whose input variable is sinusoidal function. Therefore, the designed approach offers a better alternative to descriptive sampling and RS. It can be fitted to any simulation in an economical and undemanding manner.

References

[1] G.S. Fishman, Monte-Carlo: Concepts, Algorithms and Applications, Springer-Verlag, 1997.
[2] F.J. Hickernell, H.S. Hong, P. L'Ecuyer, C. Lemieux, Extensible lattice sequences for quasi Monte-Carlo quadrature, SIAM J. Sci. Comput. 22 (3) (2000) 1117–1138.
[3] M.G. Kendall, S. Babington, Randomness and random sampling numbers, J. Roy. Stat. Soc. 101 (1938) 147–166.
[4] T.W. Körner, Fourier Analysis, Cambridge University Press, Trinity Hall, 1988.
[5] W.L. Loh, On Latin hypercube sampling, Ann. Stat. 24 (1996) 2058–2080.
[6] W.J. Morokoff, R.E. Caflisch, Quasi random sequences and their discrepancies, SIAM J. Sci. Comput. (1994) 1571–1599.
[7] M.D. McKay, W.J. Conover, R.J. Beckman, 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, CBMS-SIAM, 63, Philadelphia, 1992.
[9] A.B. Owen, A central limit theorem for Latin hypercube sampling, J. Roy. Stat. Soc. Ser. B. 54 (1992) 541–551.
[10] M. Pidd, Computer Simulation in Management Science, fourth ed., John Wiley and Sons, Chichester, 1998.
[11] J.S. Ramberg, B.W. Schmeiser, An approximate method for generating symmetric random variables, Commun. ACM 15 (1972) 987–990.
[12] K.W. Ross, D. Tsang, J. Wang, Monte-Carlo summation and integration applied to multichain queueing networks, J. Assoc. Comput. Mach. 41 (6) (1994) 1110–1135.
[13] E. Saliby, A reappraisal of some simulation fundamentals, Ph.D. Thesis, Department of Operational Research, University of Lancaster, 1980.
[14] E. Saliby, Descriptive sampling: a better approach to Monte Carlo simulation, J. Operat. Res. Soc. 41 (12) (1990) 1133–1142.
[15] E. Saliby, Understanding the variability of simulation results: an empirical study, J. Operat. Res. Soc. 41 (4) (1990) 319–327.
[16] E. Saliby, R.J. Paul, Implementing descriptive sampling in three phase discrete event simulation models, J. Operat. Res. Soc. 44 (1993) 147–160.
[17] E. Saliby, Descriptive sampling: an improvement over Latin hypercube sampling, in: Winter Simulation Conference, 1997, pp. 230–233.
[18] R. Spiegel, Murray, Fourier Analysis with Applications to Boundary Value Problems, McGraw-Hill Inc., New-York, 1974.
[19] M. Stein, Large sample properties of simulations using Latin hypercube sampling, Technometrics 29 (1987) 143–151.
[20] M. Tari, Making descriptive sampling safe and efficient, M.Ph. Thesis, Department of Operational Research, University of Lancaster, 1987.

[21] M. Tari, A. Dahmani, Descriptive sampling improved. Available from: <http://interstat.stat.vt.edu>, 2002.
[22] B. Tuffin, Variance reductions applied to product-form multi-class queueing networks, ACM Trans. Modell. Comput. Simul. 7 (4) (1997) 478–500.
[23] B. Tuffin, L.M. Le Ny, Parallélisation d'une combinaison des méthodes de Monte-Carlo et quasi Monte-Carlo et application aux réseaux de files d'attente, RAIRO Operat. Res. 34 (2000) 85–98.