Probability in the Engineering and Informational Sciences (2023), **37**:1 232–244
doi:10.1017/S0269964821000565

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Using a Chen-Stein identity to obtain low variance simulation estimators

Sheldon M. Ross

Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA.
E-mail: smross@usc.edu.

Keywords: Simulation, Chen-Stein identity, Low variance estimators, Patterns, Coupon collecting

Abstract
This paper is concerned with developing low variance simulation estimators of probabilities related to the sum of Bernoulli random variables. It shows how to utilize an identity used in the Chen-Stein approach to bounding Poisson approximations to obtain low variance estimators. Applications and numerical examples in such areas as pattern occurrences, generalized coupon collecting, system reliability, and multivariate normals are presented. We also consider the problem of estimating the probability that a positive linear combination of Bernoulli random variables is greater than some specified value, and present a simulation estimator that is always less than the Markov inequality bound on that probability.

1. Introduction and summary

For a given set of n events, let X_i be the indicator variable of event i, and let $W = \sum_{i=1}^{n} X_i$ denote the number of these events that occur. We are interested in using simulation to estimate $P(W \in A)$ for a specified set A. We show how to utilize an identity used by Chen and Stein in their work on bounding the error of a Poisson approximation to $P(W \in A)$ (see [1] or [5]) to yield a new approach for obtaining an unbiased simulation estimator of $P(W \in A)$.

In Section 2, we review the relevant Chen-Stein theory for Poisson approximations and illustrate our starting point in utilizing the identity to obtain a simulation estimator. To highlight the promise of the proposed approach, Section 3 considers the case where the Bernoulli random variables $X_1, \ldots, X_n$ are independent. In Section 4, we present a simulation estimator in the general case of dependent $X_1, \ldots, X_n$. We prove that the variance of the new simulation estimator of $P(W > 0)$ is at most $E^2[W]$. In Section 5, we consider the problem of estimating $P(N > m)$ where N is the time of the first occurrence in the sequence $Y_1, Y_2, \ldots$ of a certain pattern. In Section 5.1, we develop a simulation estimator when $Y_1, Y_2, \ldots$ are independent and identically distributed, and in Section 5.2, when they represent the sequence of states of a stationary Markov chain. Numerical examples compare the variance of the proposed estimators with those of the raw simulation estimator $I\{N > m\}$ and of the conditional Bernoulli sampling estimator. Numerical examples related to the generalized coupon collecting problem, partial sums of independent normal random variables, and system reliability are presented in Section 6.

To obtain the simulation estimator we are proposing, one has to be able to simulate the Bernoulli random variables $X_1, \ldots, X_n$ conditional on $X_i = 1$, which may be difficult in certain models. When this is so, we show in Section 7 how we can unconditionally simulate $X_1, \ldots, X_n$ and still make use of our proposed estimator, as long as none of the values $E[X_i]$ are very small. (This is analogous to using a post-stratification simulation estimator, see [6].) In Section 7, we consider the problem of estimating $P(\sum_{i=1}^{n} a_i X_i \geq k)$ where $a_1, \ldots, a_n$ are positive constants, and present a nonnegative unbiased simulation estimator that is always less than or equal to the Markov inequality bound on this probability.

2. Some relevant Chen-Stein theory

Suppose that $X_i,\ i = 1, \ldots, n$, are Bernoulli random variables with means $\lambda_i = E[X_i],\ i = 1, \ldots, n$; set $W = \sum_{i=1}^{n} X_i$ and let $\lambda = E[W]$. Also, let $P_\lambda(A) = \sum_{i \in A} e^{-\lambda}\lambda^i/i!$ be the probability that a Poisson random variable with mean λ lies in A.

For any set of nonnegative integers A, let f_A be recursively defined as follows:

$$f_A(0) = 0 \tag{1}$$

$$\lambda f_A(j+1) = j f_A(j) + I\{j \in A\} - P_\lambda(A), \quad j \geq 0. \tag{2}$$

The following lemma is key to the Chen-Stein approach.

Lemma 1. *For all* A, $|f_A(j) - f_A(i)| \leq ((1 - e^{-\lambda})/\lambda)|j - i|$.

It follows from (2) that

$$\lambda f_A(W+1) - W f_A(W) = I\{W \in A\} - P_\lambda(A).$$

Taking expectations yields that

$$\lambda E[f_A(W+1)] - E[W f_A(W)] = P(W \in A) - P_\lambda(A). \tag{3}$$

Now,

$$\begin{aligned} E[W f_A(W)] &= \sum_{i=1}^{n} E[X_i f_A(W)] \\ &= \sum_{i=1}^{n} E[f_A(W) \mid X_i = 1]\lambda_i \\ &= \sum_{i=1}^{n} E[f_A(1 + V_i)]\lambda_i \end{aligned} \tag{4}$$

where $V_1, \ldots, V_n$ are any random variables such that $V_i =_{st} \sum_{j \neq i} X_j \mid X_i = 1$. It follows from (3) and (4) that

$$\sum_i \lambda_i \left(E[f_A(W+1)] - E[f_A(1+V_i)]\right) = P(W \in A) - P_\lambda(A).$$

Hence,

$$P(W \in A) - P_\lambda(A) = \sum_i \lambda_i E\left[f_A(W+1) - f_A(1+V_i)\right] \tag{5}$$

which, from Lemma 1, yields that

$$|P(W \in A) - P_\lambda(A)| \leq \frac{1 - e^{-\lambda}}{\lambda} \sum_i \lambda_i E[|W - V_i|] \tag{6}$$

which is the Chen-Stein bound. (It is important to note that the preceding bound holds for any random vector $V_1, \ldots, V_n$ for which the distribution of V_i is the conditional distribution of $\sum_{j \neq i} X_j$ given that $X_i = 1$.) For more about Poisson approximation bounds, see [1] or [5].

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

We propose to use the identity (5), which we rewrite as

$$P(W \in A) = P_\lambda(A) + \sum_i \lambda_i E\left[f_A(W+1) - f_A(1+V_i)\right], \tag{7}$$

as the starting point of our simulation approach for estimating $P(W \in A)$. In all cases, we will make modifications so as to increase the efficiency of the simulation estimators. We first consider the case where the X_i are independent.

3. The independent case

Suppose $X_1, \ldots, X_n$ are independent. Because $\sum_{j\neq i} X_j$ is independent of X_i, it follows that $\sum_{j\neq i} X_j =_{st} \sum_{j\neq i} X_j \mid X_i = 1$, which allows us to let $V_i = \sum_{j\neq i} X_j = W - X_i$. Hence,

$$\begin{aligned} E[f_A(W+1) - f_A(V_i+1)] &= E[f_A(W+1) - f_A(V_i+1) \mid X_i = 1]\lambda_i \\ &= E[f_A(W - X_i + 2) - f_A(W - X_i + 1)]\lambda_i. \end{aligned}$$

Thus, from (7), we see that

$$P(W \in A) = P_\lambda(A) + \sum_i \lambda_i^2 E[f_A(W - X_i + 2) - f_A(W - X_i + 1)]. \tag{8}$$

We propose to simulate $X_1, \ldots, X_n$ and to estimate $P(W \in A)$ by the unbiased estimator

$$\mathcal{E} = P_\lambda(A) + \sum_i \lambda_i^2 (f_A(W - X_i + 2) - f_A(W - X_i + 1)). \tag{9}$$

Note that it follows from Lemma 1 that

$$\left| \sum_i \lambda_i^2 (f_A(W - X_i + 2) - f_A(W - X_i + 1)) \right| \leq \frac{1 - e^{-\lambda}}{\lambda} \sum_i \lambda_i^2$$

and so

$$P_\lambda(A) - \frac{1 - e^{-\lambda}}{\lambda} \sum_i \lambda_i^2 \leq \mathcal{E} \leq P_\lambda(A) + \frac{1 - e^{-\lambda}}{\lambda} \sum_i \lambda_i^2.$$

Because $P(a \leq X \leq b) = 1$ implies that $\mathrm{Var}(X) \leq (b-a)^2/4$, it follows from the preceding that

$$\mathrm{Var}(\mathcal{E}) \leq \left(\frac{1 - e^{-\lambda}}{\lambda} \sum_i \lambda_i^2 \right)^2. \tag{10}$$

For instance, if $n = 100$, $\lambda_i = i/c$, then when $c = 1,000$ we have $\mathrm{Var}(\mathcal{E}) \leq 0.00443$. Even in cases where the Poisson approximation is not particularly good, the approach works well. For instance, when $c = 200$ (and so λ_i becomes as large as 0.5), we have $\mathrm{Var}(\mathcal{E}) \leq 0.1122$. That is, even in the latter case, the variance of the estimator of any probability concerning W (even one having probability close to $1/2$) cannot exceed 0.1122. However, in any particular case, the actual variances may be quite a bit smaller than the preceding bounds. We illustrate by an example.

Example 1. *Let $X_1, \ldots, X_{20}$ be independent Bernoullis with means $E[X_i] = \lambda_i = i/50$, $i = 1, \ldots, 20$. In this case, $E[W] = \sum_{i=1}^{20} (i/50) = 4.2$. Suppose we want to estimate (a) $P(W \geq 5)$ and (b) $P(W \geq 11)$.*

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

Using that $P(Z \geq 5) = 0.410173$, $P(Z \geq 11) = 0.004069$, *where* Z *is Poisson with mean* 4.2, *a simulation consisting of* 10,000 *runs yielded that for* $A = \{W \geq 5\}$

$$E[\mathcal{E}_A] \approx 0.414851, \quad Var(\mathcal{E}_A) \approx 0.003741,$$

whereas, for $B = \{W \geq 11\}$

$$E[\mathcal{E}_B] \approx 0.000517, \quad Var(\mathcal{E}_B) \approx 0.000044,$$

where $\mathcal{E}_C$ *refers to (9) when* $A = C$. *Consequently, the variance of the estimators is much below* 0.072487, *the upper bound given by (10). The variances are also much below those of the indicator estimators which have* $Var(I\{W \in A\}) \approx 0.242750$ *and* $Var(I\{W \in B\}) \approx 0.000517$.

Remark. *We have considered the case of independent* X_i *not because a simulation is needed to determine the distribution of their sum but to indicate the promise of our approach. (To compute the mass function of* W *when* $X_1, \ldots, X_n$ *are independent, let* $P(i, j) = P(X_1 + \cdots + X_i = j)$. *Starting with* $P(1, 0) = 1 - \lambda_1$, $P(1, 1) = \lambda_1$, *and using that* $P(i, j) = 0$ *if* $i < j$, *we can recursively compute these values by using that* $P(i, j) = P(i - 1, j - 1)\lambda_i + P(i - 1, j)(1 - \lambda_i)$.*) The exact values in Example 1 are* $P(W \geq 5) = 0.4143221438$ *and* $P(W \geq 11) = 0.0004586525$.

4. The general case

In this section, we no longer suppose that $X_1, \ldots, X_n$ are independent, but allow them to have an arbitrary joint mass function. The proposed simulation approach for estimating $P(W \in A)$ starts by noting that if we let I be independent of all the other random variables and be such that

$$P(I = i) = \lambda_i/\lambda, \quad i = 1, \ldots, n$$

then we obtain from (7) that

$$P(W \in A) = P_\lambda(A) + \lambda E[f_A(W + 1) - f_A(1 + V_I)].$$

This yields the unbiased estimator

$$\mathcal{E} = P_\lambda(A) + \lambda(f_A(W + 1) - f_A(1 + V_I)). \tag{11}$$

That is, the simulation procedure is to generate I, and if $I = i$ to then generate W and V_i to obtain the value of the preceding estimator. To utilize this approach, we must be able to generate V_i for each i. In addition, we want to couple the generated values of W and V_I to be close to each other, so as to result in a small variance of the estimator.

One important case where we can analytically show that $\mathrm{Var}(\mathcal{E})$ is very small in comparison to $P(W \in A)$, at least when the latter probability is itself small, is when $A = \{0\}$. The following examples illustrate the ubiquity of this important case, which occurs when we are interested in the probability of a union of events.

1. Consider independent trials that each result in any of the outcomes $1, \ldots, n$ with probabilities $p_1, \ldots, p_n$, $\sum_{i=1}^n p_i = 1$, and suppose that there are specified numbers $r_1, \ldots, r_n$.
   The generalized birthday problem is interested in M, the number of trials until there have been r_i type i outcomes for some $i = 1, \ldots, n$. (The classical problem has $n = 365$, $p_i = 1/n$, $r_i = 2, i = 1, \ldots, n$.) With N_i being the number of type i outcomes in the first k trials, $i = 1, \ldots, n$, and $X_i = I\{N_i \geq r_i\}$, we have that $P(M > k) = P(\sum_{i=1}^n X_i = 0)$.
   The generalized coupon collecting problem concerns N, the number of trials until there have been at least r_i type i outcomes for every $i = 1, \ldots, n$. (The classical coupon collecting problem has all

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

$r_i = 1$.) With N_i defined as the number of type i outcomes in the first k trials, $i = 1, \ldots, n$, and $X_i = I\{N_i < r_i\}$, we have that $P(N \le k) = P(\sum_{i=1}^{n} X_i = 0)$.

2. In reliability systems with components $1, \ldots, m$, we often suppose that there are specified subsets of components $C_1, \ldots, C_n$, none of which is a subset of another, such that the system fails if and only if all components in at least one of these subsets are failed. Thus, with X_i being the indicator of the event that all components in C_i are failed, $P(\text{system works}) = P(\sum_{i=1}^{n} X_i = 0)$. The subsets $C_1, \ldots, C_n$ are called the minimal cut sets of the system.
3. With $Y_1, Y_2, \ldots$ being the successive states of a stationary Markov chain, a quantity of interest is N, the first time that the pattern $y_1, y_2, \ldots, y_r$ appears. With X_i defined as the indicator of the event that $Y_i = y_1, Y_{i+1} = y_2, \ldots, Y_{i+r-1} = y_r$, then $P(N > n + r - 1) = P(\sum_{i=1}^{n} X_i = 0)$.
4. In DNA matching problems (see [2]), we are often interested in the largest common subsequence in the sequences $Z_1, \ldots, Z_{r+k-1}$ and $Y_1, \ldots, Y_{s+k-1}$. In particular, we often want to determine the probability that there would be a common subsequence of length k if the $r + s + 2k - 2$ data values were independent and identically distributed with a specified mass function $\alpha_t,\ t \ge 1$. If we let $X_{i,j},\ i, j \ge 1$, equal the indicator of the event that $Z_{i+m} = Y_{j+m},\ m = 0, \ldots, k - 1$ the probability there are such subsequences is $P(\sum_{i \le r,\ j \le s} X_{i,j} > 0)$.
5. If $Y_1, \ldots, Y_n$ is multivariate normal, then $P(\max_i Y_i \le x) = P(\sum_{i=1}^{n} I\{Y_i > x\} = 0)$.

The unbiased estimator of $P(W = 0)$ given by (11) is

$$\mathcal{E} = e^{-\lambda} + \lambda(f_0(W + 1) - f_0(1 + V_I)) \tag{12}$$

where $f_0 = f_{\{0\}}$. To bound the variance of $\mathcal{E}$, we use, as shown in [7], that

$$f_0(j) = \int_0^1 e^{-\lambda t} t^{j-1}\, dt, \quad j \ge 1 \tag{13}$$

from which it follows that $f_0(j)$ is, for $j > 0$, a decreasing, convex, positive function. Because this implies that for $i > 0,\ j > 0$

$$|f_0(i) - f_0(j)| \le f_0(\min(i, j)) \le f_0(1) = \frac{1 - e^{-\lambda}}{\lambda},$$

we obtain from (12) that

$$|\mathcal{E} - e^{-\lambda}| \le 1 - e^{-\lambda}. \tag{14}$$

Consequently,

$$\text{Var}(\mathcal{E}) \le (1 - e^{-\lambda})^2 \le \lambda^2.$$

Because typically $P(W > 0) \approx \lambda$ when $P(W > 0)$ is small, it appears in this case that $\text{Var}(\mathcal{E})$ is an order of magnitude lower than $P(W > 0)(1 - P(W > 0)) \approx \lambda(1 - \lambda) \approx \lambda$, which is the variance of the raw simulation estimator $I\{W > 0\}$.

The bound given by (14) can be strengthened when W and V_I can be generated so that $W \ge V_I$, which is possible when W is stochastically larger than V_i for all i, as is the case in the generalized birthday and coupon collecting problems. With such a coupling, it follows from (12) and the fact that f_0 is decreasing that $\mathcal{E} \le e^{-\lambda}$, and so

$$2\, e^{-\lambda} - 1 \le \mathcal{E} \le e^{-\lambda}$$

showing, in this case, that

$$\text{Var}(\mathcal{E}) \le \frac{(1 - e^{-\lambda})^2}{4} \le \frac{\lambda^2}{4}.$$

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

Remarks.

1. *Another unbiased estimator of* $P(W = 0)$ *is the conditional Bernoulli sampling estimator* $\mathcal{E}_{\text{CBSE}} = 1 - \lambda/(1 + V_I)$ *(see [6] Sect. 10.1). However, in our simulation experiments, it turns out that CBSE is typically not competitive with* $\mathcal{E}$*, and that the variance of the best linear combination of these two estimators is only marginally less than* $Var(\mathcal{E})$.
2. *When computing* $f_A(j)$ *by using the recursive equations given by Eqs. (1) and (2), one must be very careful that the computation of* $P_\lambda(A)$ *is very precise. For otherwise, round off errors build up quickly. This can be avoided for the function* $f_0(j)$ *by using Eq. (13) and standard numerical approximation techniques. (For instance, as it is easily shown that* $g_j(t) \equiv e^{-\lambda t} t^{j-1}$ *is, for* $j \geq 2$*, an increasing function of* t *for* $t \in [0, 1]$*, it follows in this case that for any* m*,* $\sum_{i=1}^{m} g_j((i-1)/m)/m \leq f_0(j) \leq \sum_{i=1}^{m} g_j(i/m)/m$.)
3. *Although the values* $f_A(j)$ *must be computed with great precision, their computation does not add much time to the total simulation. This is because not only does the estimator from each simulation run require only* 2 *values of* $f_A(j)$*, but once these values are computed they can be used for other runs needing them. Consequently, when compared with using the Monte-Carlo estimator* $I\{W \in A\}$*, the additional computation time needed for our estimator primarily depends on the additional simulation beyond generating* W *that is needed to generate* V_I.
4. *As will be seen in the Examples to follow, the coupling of* W *and* V_I *often results in only a minimal additional simulation effort, beyond that of generating* W*, needed to generate our estimator.*

5. Pattern problem examples

In this section, we consider some examples where we are interested in whether a certain pattern occurs at some point within the sequence of random variables $Y_1, \ldots, Y_s$ (see [3,4] for applications).

5.1. A pattern problem with independent data

Suppose Y_i, $i \geq 1$ are independent and identically distributed with mass function $P(Y_i = j) = p_j$, $\sum_{j=1}^{k} p_j = 1$. Let N be the first time that there is a run of r consecutive equal values. That is, $N = \min\{m : Y_m = Y_{m-1} = \cdots = Y_{m-r+1}\}$, and suppose we are interested in estimating $p \equiv P(N > n + r - 1)$. To utilize our approach, we generate $Y_1, \ldots, Y_{n+r-1}$ to determine $W = \sum_{i=1}^{n} X_i$, where X_i is the indicator of the event that $Y_i = \cdots = Y_{i+r-1}$. We now generate I which, because $\lambda_i = \sum_{j=1}^{k} p_j^r$, is equally likely to be any of the values $1, \ldots, n$. Suppose $I = i$, we then generate J such that $P(J = j) = p_j^r / \sum_{i=1}^{k} p_i^r$, $j = 1, \ldots, k$ and, if $J = j$, reset the values of $Y_i, \ldots, Y_{i+r-1}$ to now all equal j. Letting V be the number of times there is a run of r equal values when using the reset values, then $V =_{st} 1 + \sum_{j \neq i} X_j \mid X_i = 1$. Consequently, with $\lambda = n \sum_{j=1}^{k} p_j^r$, the estimator of $P(N > n + r - 1) = P(W = 0)$ is

$$\mathcal{E} = e^{-\lambda} + \lambda(f_0(W + 1) - f_0(V)). \tag{15}$$

This estimator can be improved by taking its conditional expectation given all variables except J. That is, consider

$$\mathcal{E}^* \equiv E[\mathcal{E} \mid I, Y_1, \ldots, Y_{n+r-1}].$$

Now, letting V_j^* be the number of times there is a run of r consecutive equal values when $Y_I, \ldots, Y_{I+r-1}$ are reset to all equal j then, with $\alpha_j = p_j^r / \sum_{i=1}^{k} p_i^r$, $j = 1, \ldots, k$, we have

$$\mathcal{E}^* = e^{-\lambda} + \lambda \sum_{j=1}^{k} \alpha_j (f_0(W + 1) - f_0(V_j^*)). \tag{16}$$

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

Example 2. *Suppose that* $n = 1,000$ *and* $P(Y_i = i) = 1/5,\ i = 1, \ldots, 5$. *The following table, based on the results from* 10,000 *simulation runs, gives for various values of* r, *the values (as determined by the simulation) of* $p = P(W = 0)$, $p(1-p)$ *(equal to the variance of the indicator estimator),* $Var(\mathcal{E}_{\text{CBSE}})$, $Var(\mathcal{E})$, *and* $Var(\mathcal{E}^*)$, *where* $\mathcal{E}^*$ *is as given in (16). The simulation also gave the value of* $Cov(\mathcal{E}, \mathcal{E}_{\text{CBSE}})$, *which enabled us to determine* V_b, *the variance of the best linear combination of these two unbiased estimators (equal to the variance obtained when using* $\mathcal{E}$ *along with* $\mathcal{E} - \mathcal{E}_{\text{CBSE}}$ *as a control variable). That is, it gave the value of*

$$V_b = \min_{\alpha} Var(\alpha\mathcal{E} + (1-\alpha)\mathcal{E}_{\text{CBSE}}) = Var(\mathcal{E})\left(1 - Corr^2(\mathcal{E}, \mathcal{E} - \mathcal{E}_{\text{CBSE}})\right).$$

As indicated in the table, V_b *is only marginally less than* $Var(\mathcal{E})$, *and much larger than* $Var(\mathcal{E}^*)$.

$$\begin{pmatrix} r & p & p(1-p) & Var(\mathcal{E}_{\text{CBSE}}) & Var(\mathcal{E}) & V_b & Var(\mathcal{E}^*) \\ 5 & 0.276018 & 0.199832 & 0.194959 & 0.028545 & 0.026579 & 0.006636 \\ 6 & 0.773515 & 0.175190 & 0.009301 & 0.005185 & 0.005072 & 0.000444 \\ 7 & 0.950120 & 0.047392 & 3.1819 \times 10^{-4} & 2.7736 \times 10^{-4} & 2.7643 \times 10^{-4} & 8.7166 \times 10^{-6} \\ 8 & 0.989804 & 0.010092 & 1.2331 \times 10^{-5} & 1.1964 \times 10^{-5} & 1.1958 \times 10^{-5} & 2.3745 \times 10^{-7} \end{pmatrix}.$$

Remark. *Because of our coupling of* V *and* W, *the simulation effort needed to obtain our estimator is basically the same as needed to obtain* W.

5.2. A pattern problem with Markov chain generated data

Consider a stationary Markov chain $Y_m,\ m \geq 1$, with transition probabilities $P_{u,v}$ and stationary probabilities π_u. Let $Q_{u,v} = \pi_v P_{v,u}/\pi_u$ be the transition probabilities of the reverse chain. We are interested in the probability that the pattern $y_1, \ldots, y_r$ does not appear within the first $n + r - 1$ data values. To use our method, let X_i be the indicator of the event that $Y_i = y_1, \ldots, Y_{i+r-1} = y_r$ and note that $\lambda_i = P(X_i = 1) = \pi_{y_1} P_{y_1,y_2} \cdots P_{y_{r-1},y_r}$. With $W = \sum_{i=1}^{n} X_i$, we are interested in $P(W = 0)$.

To estimate $P(W = 0)$, first generate I, equally likely to be any of $1, \ldots, n$. Suppose $I = i$.

1. Set $Y_i = y_1, \ldots, Y_{i+r-1} = y_r$.
2. For $j \geq i + r$, if $Y_{j-1} = u$, then let $Y_j = v$ with probability $P_{u,v}$. (In other words, starting at time $i + r - 1$, generate the remaining states in sequence by using the transition probability $P_{u,v}$.)
3. For $j < i$, if $Y_{j+1} = u$, then let $Y_j = v$ with probability $Q_{u,v}$. (So going backwards from time i we generate the states using the transition probabilities of the reversed chain).

Let $V = 1 + V_I$ be the number of times the pattern $y_1, \ldots, y_r$ appears in $Y_1, \ldots, Y_{n+r-1}$.

To generate W, we will define a new Markov chain with transition probabilities $P_{u,v}$ and let W be the number of times the pattern appears. However, we will do it in a way so that it is related to the Y-chain above. It is defined as follows. To begin, recall that the generated value of I was $I = i$. We define the new chain—let its states be $W_1, \ldots, W_{n+r-1}$—as follows:

1. Simulate W_i by using that $P(W_i = k) = \pi_k$.
2. For $j < i$,

   if $W_{j+1} = Y_{j+1}$, set $W_k = Y_k$ for all $k = 1, \ldots, j$.
   2. if $W_{j+1} \neq Y_{j+1}$, then if $W_{j+1} = u$, then let $W_j = v$ with probability $Q_{u,v}$.

3. For $j > i$. Starting with the simulated value of W_i, generate the states W_{i+1} up to W_{i+r-1} by using the transition probabilities $P_{u,v}$. For the states at times $j \geq i + r$ do the following:

   if $W_{j-1} = Y_{j-1}$, set $W_k = Y_k$ for all $k \geq j$.
   3. if $W_{j-1} \neq Y_{j-1}$, then if $W_{j-1} = u$, let $W_j = v$ with probability $P_{u,v}$.

Let W be the number of times the pattern appears in $W_1, \ldots, W_{n+r-1}$.

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

So in generating the chain for determining W we start by generating the value of W_i and using its value we then use the transition probabilities until we have generated the values of $W_i, \ldots, W_{i+r-1}$. For times larger than $i + r - 1$, we will continue generating according to the transition probabilities $P_{u,v}$, except that if at some time we are in the same state as the Y-chain was at that time then we just let the W chain's value equal the Y-chain's value from then on. We do the same thing going backwards in time, except that we use the reverse transition probabilities. (If it is easier to generate from the original chain than from the reversed chain, we can reverse the procedure by first generating the states of the forward chain to determine W and then generate the chain that determines the value of V, coupling its values with those of the first chain when appropriate.)

The estimator of $P(W = 0)$ is

$$\mathcal{E} = e^{-\lambda} + \lambda(f_0(W + 1) - f_0(V)).$$

Example 3. *The following example will consider* 4 *Markov chains, all with states* 0, 1. *The transitions probabilities for these chains are*

Case 1: $P_{00} = 0.3,\ P_{10} = 0.2$
Case 2: $P_{00} = 0.3,\ P_{10} = 0.8$
Case 3: $P_{00} = 0.5,\ P_{10} = 0.6$
Case 4: $P_{00} = 0.5,\ P_{10} = 0.8$

For each chain, we use the preceding approach to estimate p, the probability that the pattern 0000011111 does not occur in the first 5,009 data values. The following table, based on 10,000 simulation runs, gives (as determined by the simulation) the variances of $\mathcal{E}$ and of $\mathcal{E}_{\text{CBSE}} = 1 - \lambda/V$, as well as V_b, the variance of the best convex linear combination of these estimators.

Case	p	$p(1-p)$	$\text{Var}(\mathcal{E}_{\text{CBSE}})$	$\text{Var}(\mathcal{E})$	V_b
1	0.074597	0.069033	0.309118	0.000458	0.000456
2	0.976097	0.023332	3.61369×10^{-6}	1.43963×10^{-8}	1.43420×10^{-8}
3	0.111967	0.099430	0.268122	0.0004628	0.0004627
4	0.857370	0.122287	7.95539×10^{-4}	2.2507×10^{-6}	2.2445×10^{-6}

Remarks. *1. Because* $0 \le 1 - \mathcal{E}_{\text{CBSE}} = \lambda/V \le \lambda$, *it follows that* $Var(\mathcal{E}_{\text{CBSE}}) \le \lambda^2/4$, *and so always has a small variance when* λ *is small. However, its variance can be large when* $\lambda > 1$, *and that is the reason why it is large in Cases* 1 *and* 3, *which have respective values* $\lambda = 2.58048$ *and* $\lambda = 2.18182$. *The estimator* $\mathcal{E}$ *has a small variance in all cases.*

2. *Because the pattern* 0000011111 *has "no overlap" (in the sense that no part of an occurring pattern can be utilized in the next occurrence of the pattern), the Poisson approximation* $P(W = 0) \approx e^{-\lambda}$ *would be expected to be quite accurate, and indeed it yields the following estimates of p in the four cases: 0.075738, 0.976098, 0.112836, and 0.857404. On the other hand, if the pattern were* 1111111111, *then it would not be expected to be so accurate. Indeed, when the transition probabilities are as given in Case 3, the Poisson estimator of the probability that this pattern does not occur within the first 5,009 data values is* 0.55172, *whereas simulation shows that* $p = 0.69673,\ p(1-p) = 0.21130,\ Var(\mathcal{E}) = 0.021106$. *Thus, once again* $\mathcal{E}$ *has a very small variance.*

Example 4. *Consider the Markov chain with states* 1, 2, 3 *and transition probability matrix*

$$\begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.6 & 0.2 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

and suppose we are interested in estimating p, the probability, starting in steady state, that the pattern $1, 2, 3, 1, 3, 2, 2, 1, 2, 3$ does not occur within $5,009$ data values. In this case, a simulation based on 10,000 runs, yielded the estimates

$$p = 0.9893264, \quad Var(\mathcal{E}_{\text{CBSE}}) = 3.18819 \times 10^{-7}, \quad Var(\mathcal{E}) = 1.97215 \times 10^{-31}.$$

Remark. *Because we need to generate two Markov chains to obtain our estimator, the simulation effort could be twice what is needed to generate W. However, because of the coupling of these two chains, if the number of states of the Markov chain is not too large, then the number of additional simulations needed beyond generating one chain is minimal. In any case, in the preceding examples, the variance of our estimator is much smaller than $Var(I\{W = 0\}) = p(1 - p)$.*

5.3. Additional examples

Example 5 (A Generalized Coupon Collecting Problem)**.** *Suppose that 1,000 balls are independently distributed into 10 urns with each ball going into urn i with probability p_i, $i = 1, \ldots, 10$. Let N_i denote the number of balls that go into urn i and set $X_i = I\{N_i < r_i\}$. With $W = \sum_{i=1}^{10} X_i$, we are interested in estimating $P(W = 0)$.*

To determine the value of our estimator $\mathcal{E} = e^{-\lambda} + \lambda(f_0(W+1) - f_0(V))$, first generate the multinomial vector $(N_1, \ldots, N_{10})$ and use it to determine W. We will also use these N_i to determine $V (= V_I + 1)$. Let $\lambda_i = E[X_i] = P(\text{Bin}(1,000, p_i) < r_i)$, where $\text{Bin}(n, p)$ is binomial with parameters (n, p); let $\lambda = \sum_{i=1}^{10} \lambda_i$, and do the following:

1. Generate I where $P(I = i) = \lambda_i/\lambda$. Suppose $I = j$.
2. If $N_j < r_j$ let $V = W$.
3. If $N_j \geq r_j$ generate Y, a binomial $(1,000, p_j)$ random variable conditioned to be less than r_j. Suppose $Y = k$. Then, remove $N_j - k$ balls from box j, putting each of these balls into one of the urns $i, i \neq j$ with prob $p_i/(1 - p_j)$. Let $N_1^*, \ldots, N_{10}^*$ be the new number of balls in each urn after the preceding is done. Let $V = \sum_{k=1}^{10} I\{N_k^* < r_k\}$.

A simulation of $10,000$ runs, when $p_i = (10 + i)/155$, $r_i = 60 + 4i$, $i = 1, \ldots, 10$, yielded the result

$$P(W = 0) = 0.572859, \quad \text{Var}(\mathcal{E}) = 0.006246.$$

Because $\lambda = 0.493024$, the Poisson approximation of $P(W = 0)$ is $e^{-\lambda} = 0.610779$.

Remark. *Because of our coupling of V and W, the simulation needed beyond obtaining W is minimal. Either no additional simulation, except for I, is needed or, no matter how many urns, one would need to then simulate a binomial conditioned to be smaller than some value (easily done by the discrete inverse transform algorithm) and then generate a small number of additional random variables that are used to modify the number of balls in urns j, $j \neq I$.*

Example 6 (A Partial Sums Example)**.** *Suppose $Z_1, \ldots, Z_{20}$ are independent and identically distributed standard normals. Let*

$$S_i = \sum_{j=1}^{i} Z_j, \quad X_i = I\{|S_i| > c\sqrt{i}\}, \quad W = \sum_{i=1}^{20} X_i.$$

Suppose we want to estimate $P(W = 0)$ and $P(W = 2)$. To obtain our simulation estimators, we use that the joint distribution of Z_k and S_k is bivariate normal with correlation $1/\sqrt{k}$. From this, it follows that the conditional distribution of Z_k given that $S_k = y$ is normal with mean y/k and variance

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

$(k-1)/k$. *Using the preceding and that* $\lambda_i = 2P(Z_1 > c)$, *the simulation estimator is obtained as follows:*

- *Generate* I *such that* $P(I = i) = 1/20,\ i = 1, \ldots, 20$.
- *If* $I = i$, *generate a normal with mean* 0 *and variance* i *that is conditioned to be greater than* $c\sqrt{i}$. *(This is efficiently done either by using the reject procedure with an exponential distribution—see [6] pp. 218–219—or, with* Φ *being the standard normal distribution function and* U *being uniform on* $(0, 1)$, *by letting* $U^* = \Phi(c) + (1 - \Phi(c))U$ *and using* $\sqrt{i}\Phi^{-1}(U^*)$.*) Let the value of this generated random variable be* s_i.
- *Generate* Z_i *conditional on* $S_i = s_i$. *Let* z_i *be the generated value.*
- *Generate* Z_{i-1} *conditional on* $S_{i-1} = s_i - z_i$. *Let* z_{i-1} *be the generated value.*
- *Generate* Z_{i-2} *conditional on* $S_{i-2} = s_i - z_i - z_{i-1}$. *Let* z_{i-2} *be the generated value.*
- *Continue the preceding until you have generated* Z_1.
- *Generate independent standard normals* $Z_{i+1}, \ldots, Z_{20}$ *and let their values be* $z_{i+1}, \ldots, z_{20}$.
- *With* $s_j = \sum_{k=1}^{j} z_k,\ j = 1, \ldots, 20$, *let* $V =$ *number* $j : |s_j| > c\sqrt{j}$.
- *Generate independent standard normals* $Z_1, \ldots, Z_i$. *Using these along with the previously generated* $Z_{i+1}, \ldots, Z_{20}$ *determine the value of* W.
- *Using the preceding generated values of* $V = 1 + V_I$ *and* W *along with* $\lambda = 40P(Z > c)$, *obtain the values of the estimators.*

Note that because, given $I = i$, *the distribution of the estimator will be the same whether* $S_i > c\sqrt{i}$ *or* $S_i < -c\sqrt{i}$, *we have arbitrarily assumed the former.*

The following table, based on 10,000 *simulation runs, gives the simulation estimates of the variances of* $\mathcal{E}_0$ *and* $\mathcal{E}_2$, *the proposed estimators for* $P(W = 0)$ *and for* $P(W = 2)$, *for* $c = 2$ *and* $c = 2.5$.

$$\begin{pmatrix} Case & P(W=0) & Var(\mathcal{E}_0) & P(W=2) & Var(\mathcal{E}_2) \\ c=2 & 0.76980 & 3.842\times10^{-3} & 0.03788 & 4.644\times10^{-4} \\ c=2.5 & 0.92279 & 3.519\times10^{-5} & 0.01347 & 3.177\times10^{-6} \end{pmatrix}.$$

Remark. *If* $I = i$, *both* V *and* W *use the values* $Z_{i+1}, \ldots, Z_{20}$. *Consequently, for a given* c, *the simulation effort to obtain our estimators is about* 1.5 *times that needed to generate* $Z_1, \ldots, Z_{20}$.

Example 7 (A Reliability Application)**..** *Consider a system of five independent components, and suppose that the system is failed if and only if all of the components of any of the component sets* $C_1 = \{1, 2\},\ C_2 = \{4, 5\},\ C_3 = \{1, 3, 5\},\ C_4 = \{2, 3, 4\}$ *are all failed. Suppose further that component* i *is failed with probability* q_i *where* $q_1 = c/10,\ q_2 = c/12,\ q_3 = c/15,\ q_4 = c/20,\ q_5 = c/30$. *We want to use simulation to estimate the probability that the system is not failed for various values of* c. *Thus, with* X_i *being the indicator variable that all components in* C_i *are failed, and* $W = \sum_{i=1}^{4} X_i$ *we are interested in* $P(W = 0)$. *For various values of* c, *the following table gives, as determined by the simulation,* $P(W = 0)$ *as well as* $Var(\mathcal{E}_{\text{CBSE}})$ *and* $Var(\mathcal{E})$ *where, with* $\lambda = E[W]$,

$$\mathcal{E} = e^{-\lambda} + \lambda(f_0(W+1) - f_0(V)), \quad \mathcal{E}_{\text{CBSE}} = 1 - \frac{\lambda}{V}.$$

The table also gives V_b, *the variance of the best unbiased linear combination of these two estimators. The table is based on the results of* 50,000 *simulation runs.*

$$\begin{pmatrix} c & p & Var(\mathcal{E}_{\text{CBSE}}) & Var(\mathcal{E}) & V_b \\ 1 & 0.989579 & 4.2032\times10^{-7} & 4.8889\times10^{-7} & 3.7883\times10^{-7} \\ 2 & 0.95721 & 2.6329\times10^{-5} & 3.0573\times10^{-5} & 2.3200\times10^{-5} \\ 3 & 0.90248 & 0.000293 & 0.000315 & 0.000243 \\ 5 & 0.73230 & 0.005685 & 0.005228 & 0.00422 \end{pmatrix}.$$

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

Thus, interestingly, it appears that except for when $c = 5$, the conditional Bernoulli sampling estimator appears to have a smaller variance than does the one based on the Chen-Stein identity.

Remark. *Because W is obtained by simulating the state—either failed or not—of each component, and V is then obtained by simulating I and using the earlier simulated states for all components not in C_I, there is basically no additional simulation effort needed to obtain our estimator.*

6. A post-simulation approach

There are many models where it is very difficult to simulate the system conditional on $X_i = 1$. (For instance, suppose a customer has just arrived at a stationary $M/M/1$ queueing system and we are interested in the distribution of number of the next n customers that have to spend longer than x in the system.) If none of $\lambda_1, \ldots, \lambda_n$ is very small, one possibility is to unconditionally simulate the system to obtain $X_1, \ldots, X_n$, and then use the resulting data to yield realizations of the quantities V_i for those i for which $X_i = 1$. In this way, a single simulation run would yield the values of $W = \sum_{i=1}^n X_i$ and of W of the random variables $V_1, \ldots, V_n$. We could then use the identity

$$P(W \in A) = P_\lambda(A) + \lambda E[f_A(W+1)] - \sum_{i=1}^n \lambda_i E[f_A(V_i+1)]$$

to obtain a simulation estimator of $P(W \in A)$. Namely, if we have r simulation runs, with $X_1^t, \ldots, X_n^t$ being the simulated vector in run t, then we can estimate $E[f_A(W+1)]$ by $(1/r)\sum_{t=1}^r f_A(1+\sum_{i=1}^n X_i^t)$, and $E[f_A(V_i+1)]$ by $\sum_{t=1}^r I\{X_i^t = 1\} f_A(1+\sum_{j\neq i} X_j^t)/\sum_{t=1}^r I\{X_i^t = 1\}$.

Example 8. *Each of* 200 *balls independently goes into box i with probability $p_i = (10+i)/155$, $i = 1, \ldots, 10$. With N_i denoting the number of balls that go into box i, let $X_i = I\{N_i < 12+i\}$, $i = 1, \ldots, 10$, $W = \sum_{i=1}^{10} X_i$, and let $p = P(W \leq 3)$. We want to determine p by simulation.*

Using our technique, we let $r = 100$ and repeated this procedure 300 times. With $\mathcal{E}$ being the estimator of p, based on $r = 100$ simulation runs, the simulation gave that

$$E[\mathcal{E}] = 0.852839, \quad \mathrm{Var}(\mathcal{E}) = 0.000663.$$

The preceding variance can be compared to $p(1-p)/100 = 0.001255$, the variance of the raw simulation estimator based on 100 runs. (If we would have done the simulation as in Example 5, then the resulting variance based on a single simulation run is 0.036489, giving that the variance of the average of 100 runs is 0.000365.)

7. Estimating probabilities of positive linear combinations

As before let $X_1, \ldots, X_n$ be Bernoulli random variables, and let $\lambda_i = E[X_i]$, $i = 1, \ldots, n$. With $S = \sum_{i=1}^n a_i X_i$, where $a_1, \ldots, a_n$ are arbitrary positive constants, suppose we are interested in estimating $p = P(S \geq k)$. We now develop a simulation procedure that yields a nonnegative unbiased estimator that is always less than or equal to the Markov inequality bound $E[S]/k$.

To obtain our estimator, note that for any random variable R

$$E[SR] = \sum_{i=1}^n a_i E[X_i R] = \sum_{i=1}^n a_i E[R \mid X_i = 1]\lambda_i. \tag{17}$$

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

Now, let I be independent of $R, X_1, \ldots, X_n$ and be such that

$$P(I = i) = a_i/a, \quad i = 1, \ldots, n \quad \text{where } a = \sum_{i=1}^{n} a_i.$$

Noting that

$$P(I = i \mid X_I = 1) = \frac{a_i \lambda_i}{\sum_{i=1}^{n} a_i \lambda_i} = \frac{a_i \lambda_i}{E[S]}$$

yields that

$$E[R \mid X_I = 1] = \sum_{i=1}^{n} E[R \mid X_I = 1,\ I = i] a_i \lambda_i / E[S] = \sum_{i=1}^{n} E[R \mid X_i = 1] a_i \lambda_i / E[S].$$

Hence, from (17), we obtain the following result.

Lemma 2. *For any random variable R*

$$E[SR] = E[S] E[R \mid X_I = 1].$$

This yields

Proposition 1.

$$P(S \geq k) = E[S] E\left[\frac{I\{S \geq k\}}{S} \mid X_I = 1 \right].$$

Proof. Letting $R = 0$ if $S = 0$ and $R = I\{S \geq k\}/S$ if $S > 0$, and applying Lemma 2 yields the result. □ □

It follows from Proposition 1 that we can estimate $P(S \geq k)$ by first simulating I conditional on $X_I = 1$. If $I = i$, then simulate $X_j,\ j \neq i$, conditional on $X_i = 1$, let $S^* = a_i + \sum_{j \neq i} a_j X_j$ and return the estimate $E[S] I\{S^* \geq k\}/S^*$. Because $0 \leq I\{S^* \geq k\}/S^* \leq 1/k$, this estimator is always nonnegative and at most the Markov inequality bound of $P(S \geq k)$; thus, it should have a small variance when this bound is small.

7.1. When the X_i are independent

In this case, use that

$$\begin{aligned} P(S \geq k) &= \sum_{i=1}^{n} E\left[\frac{I\{S \geq k\}}{S} \middle| X_i = 1 \right] a_i \lambda_i \\ &= E\left[\sum_{i=1}^{n} a_i \lambda_i \frac{I\{a_i + S - a_i X_i \geq k\}}{a_i + S - a_i X_i} \right]. \end{aligned}$$

So the simulation approach, in this case, is to generate $X_1, \ldots, X_n$, let $S = \sum_i a_i X_i$, and use the estimator $\sum_{i=1}^{n} a_i \lambda_i (I\{a_i + S - a_i X_i \geq k\}/(a_i + S - a_i X_i))$.

Example 9. *Suppose the X_i are independent with $\lambda_i = (i + 20)/240,\ i = 1, \ldots, 100$ and $a_i = 20 - i/10,\ i = 1, \ldots, 100$. The mean and variance of the estimator of $P(S > 410)$ are, respectively, 0.451232 and 0.173305. The mean and variance of the estimator of $P(S > 440)$ are, respectively, 0.286072 and 0.140649. (The variances in these two cases can be compared with the variances of the raw simulation estimators $I\{S > 410\}$ and $I\{S > 410\}$, which are, respectively, 0.247622 and 0.204235.)*

https://doi.org/10.1017/S0269964821000565 Published online by Cambridge University Press

7.2. The general case

In the general case, we generate I such that $P(I = i) = a_i\lambda_i/E[S],\ i = 1, \ldots, n$. If $I = i$, we generate $X_1, \ldots, X_n$ conditional on $X_i = 1$, set $S^* = a_i + \sum_{j \neq i} a_j X_j$ and use the estimator $\mathcal{E} = E[S]I\{S^* \geq k\}/S^*$.

Example 10. *In the multinomial example considered in Example 8, suppose that* $a_i = 20 - i,\ i = 1, \ldots, 10$ *and that we want the mean and variance of the estimators of* $p = P(S > k)$ *for* $k = 40, 45, 50, 55$. *Using the approach of this section, the results are given by the following table.*

$$\begin{pmatrix} k & p = P(S > k) & p(1-p) & Var(\mathcal{E}) \\ 40 & 0.460100 & 0.248408 & 0.135402 \\ 45 & 0.340522 & 0.224567 & 0.125781 \\ 50 & 0.200719 & 0.160431 & 0.089031 \\ 55 & 0.129894 & 0.113022 & 0.060730 \end{pmatrix}.$$

Thus, in all cases, the variance of the proposed estimator of $P(S > k)$ *is quite a bit less than* $Var(I\{S > k\})$.

Acknowledgments. We would like to thank Rundong Ding for carrying out the simulation computations.

Competing interest. The authors declare no conflict of interest.

References

[1] Barbour, A.D., Holst, L., & Janson, S. (1992). *Poisson approximations*, vol. 2. New York, NY: Oxford University Press.
[2] Lippert, R.A., Huang, H., & Waterman, M.S. (2002). Distributional regimes for the number of k-word matches between two random sequences. *Proceedings of the National Academy of Sciences of the United States of America* 99(22): 13980–13989.
[3] Lladser, M.E., Betterton, D., & Knight, R. (2008). Multiple pattern matching: A Markov chain approach. *Journal of Mathematical Biology* 56(1), 51–92.
[4] Regnier, M. & Szpankowski, W. (1998). On pattern frequency occurrences in a Markovian sequence. *Algorithmica* 22(4): 631–649.
[5] Ross, N. (2011). Fundamentals of Stein's method. *Probability Surveys* 8: 210–293.
[6] Ross, S.M. (2013). *Simulation*, 5th ed. San Diego, CA: Academic Press.
[7] Ross, S.M. (2016). Improved Chen-Stein bounds on the probability of a union. *Journal of Applied Probability* 53(4): 1265–1270.

Cite this article: Ross SM (2023). Using a Chen-Stein identity to obtain low variance simulation estimators. *Probability in the Engineering and Informational Sciences* **37**, 232–244. https://doi.org/10.1017/S0269964821000565