Variance Reduction Techniques

Introduction

In a typical scenario for a simulation study, one is interested in determining θ, a parameter connected with some stochastic model. To estimate θ, the model is simulated to obtain, among other things, the output datum X which is such that $\theta = E[X]$. Repeated simulation runs, the ith one yielding the output variable X_i, are performed. The simulation study is then terminated when n runs have been performed and the estimate of θ is given by $\overline{X} = \sum_{i=1}^{n} X_i/n$. Because this results in an unbiased estimate of θ, it follows that its mean square error is equal to its variance. That is,

$$\text{MSE} = E[(\overline{X} - \theta)^2] = \text{Var}(\overline{X}) = \frac{\text{Var}(X)}{n}$$

Hence, if we can obtain a different unbiased estimate of θ having a smaller variance than does $\overline{X}$, we would obtain an improved estimator.

In this chapter we present a variety of different methods that one can attempt to use so as to reduce the variance of the (so-called raw) simulation estimate $\overline{X}$.

However, before presenting these variance reduction techniques, let us illustrate the potential pitfalls, even in quite simple models, of using the raw simulation estimator.

Example 9a Quality Control Consider a process that produces items sequentially. Suppose that these items have measurable values attached to them and that when the process is "in control" these values (suitably normalized) come from a standard normal distribution. Suppose further that when the process goes "out of control" the distribution of these values changes from the standard normal to some other distribution.

Simulation. DOI: http://dx.doi.org/10.1016/B978-0-12-415825-2.00009-7

To help detect when the process goes out of control the following type of procedure, called an exponentially weighted moving-average control rule, is often used. Let $X_1, X_2, \ldots$ denote the sequence of data values. For a fixed value $\alpha, 0 \leqslant \alpha \leqslant 1$, define the sequence $S_n, n \geqslant 0$, by

$$\begin{aligned} S_0 &= 0 \\ S_n &= \alpha S_{n-1} + (1-\alpha)X_n, \quad n \geqslant 1, \end{aligned}$$

Now when the process is in control, all the X_n have mean 0, and thus it is easy to verify that, under this condition, the exponentially weighted moving-average values S_n also have mean 0. The moving-average control rule is to fix a constant B, along with the value of α, and then to declare the process "out of control" when $|S_n|$ exceeds B. That is, the process is declared out of control at the random time N, where

$$N = \text{Min}\{n : |S_n| > B\}$$

Now it is clear that eventually $|S_n|$ will exceed B and so the process will be declared out of control even if it is still working properly—that is, even when the data values are being generated by a standard normal distribution. To make sure that this does not occur too frequently, it is prudent to choose α and B so that, when the $X_n, n \geqslant 1$, are indeed coming from a standard normal distribution, $E[N]$ is large. Suppose that it has been decided that, under these conditions, a value for $E[N]$ of 800 is acceptable. Suppose further that it is claimed that the values $\alpha = 0.9$ and $B = 0.8$ achieve a value of $E[N]$ of around 800. How can we check this claim?

One way of verifying the above claim is by simulation. Namely, we can generate standard normals $X_n, n \geqslant 1$, until $|S_n|$ exceeds 0.8 (where $\alpha = 0.9$ in the defining equation for S_n). If N_1 denotes the number of normals needed until this occurs, then, for our first simulation run, we have the output variable N_1. We then generate other runs, and our estimate of $E[N]$ is the average value of the output data obtained over all runs.

However, let us suppose that we want to be 99 percent confident that our estimate of $E[N]$, under the in-control assumption, is accurate to within ± 0.1. Hence, since 99 percent of the time a normal random variable is within ± 2.58 standard deviations of its mean (i.e., $z_{.005} = 2.58$), it follows that the number of runs needed—call it n—is such that

$$\frac{2.58\sigma_n}{\sqrt{n}} \approx 0.1$$

where σ_n is the sample standard deviation based on the first n data values. Now σ_n will approximately equal $\sigma(N)$, the standard deviation of N, and we now argue that this is approximately equal to $E[N]$. The argument runs as follows: Since we are assuming that the process remains in control throughout, most of the time the value of the exponentially weighted moving average is near the origin. Occasionally, by chance, it gets large and approaches, in absolute value, B. At such times it may go beyond B and the run ends, or there may be a string of normal data values

which, after a short time, eliminate the fact that the moving average had been large (this is so because the old values of S_i are continually multiplied by 0.9 and so lose their effect). Hence, if we know that the process has not yet gone out of control by some fixed time k, then, no matter what the value of k, it would seem that the value of S_k is around the origin. In other words, it intuitively appears that the distribution of time until the moving average exceeds the control limits is approximately memoryless; that is, it is approximately an exponential random variable. But for an exponential random variable Y, $\text{Var}(Y) = (E[Y])^2$. Since the standard deviation is the square root of the variance, it thus seems intuitive that, when in control throughout, $\sigma(N) \approx E[N]$. Hence, if the original claim that $E[N] \approx 800$ is correct, the number of runs needed is such that

$$\sqrt{n} \approx 25.8 \times 800$$

or

$$n \approx (25.8 \times 800)^2 \approx 4.26 \times 10^8$$

In addition, because each run requires approximately 800 normal random variables (again assuming the claim is roughly correct), we see that to do this simulation would require approximately $800 \times 4.26 \times 10^8 \approx 3.41 \times 10^{11}$ normal random variables—a formidable task. □

9.1 The Use of Antithetic Variables

Suppose we are interested in using simulation to estimate $\theta = E[X]$ and suppose we have generated X_1 and X_2, identically distributed random variables having mean θ. Then

$$\text{Var}\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}\text{Var}(X_1) + [\text{Var}(X_2) + 2\text{Cov}(X_1, X_2)]$$

Hence it would be advantageous (in the sense that the variance would be reduced) if X_1 and X_2 rather than being independent were negatively correlated.

To see how we might arrange for X_1 and X_2 to be negatively correlated, suppose that X_1 is a function of m random numbers: that is, suppose that

$$X_1 = h(U_1, U_2, \ldots, U_m)$$

where $U_1, \ldots, U_m$ are m independent random numbers. Now if U is a random number—that is, U is uniformly distributed on (0, 1)—then so is $1 - U$. Hence the random variable

$$X_2 = h(1 - U_1, 1 - U_2, \ldots, 1 - U_m)$$

has the same distribution as X_1. In addition, since $1 - U$ is clearly negatively correlated with U, we might hope that X_2 might be negatively correlated with X_1;

and indeed that result can be proved in the special case where h is a monotone (either increasing or decreasing) function of each of its coordinates. [This result follows from a more general result which states that two increasing (or decreasing) functions of a set of independent random variables are positively correlated. Both results are presented in the Appendix to this chapter.] Hence, in this case, after we have generated $U_1, \ldots, U_m$ so as to compute X_1, rather than generating a new independent set of m random numbers, we do better by just using the set $1 - U_1, \ldots, 1 - U_m$ to compute X_2. In addition, it should be noted that we obtain a double benefit: namely, not only does our resulting estimator have smaller variance (at least when h is a monotone function), but we are also saved the time of generating a second set of random numbers.

Example 9b Simulating the Reliability Function Consider a system of n components, each of which is either functioning or failed. Letting

$$s_i = \begin{cases} 1 & \text{if component } i \text{ works} \\ 0 & \text{otherwise} \end{cases}$$

we call $\mathbf{s} = (s_1, \ldots, s_n)$ the state vector. Suppose also that there is a nondecreasing function $\phi(s_1, \ldots, s_n)$ such that

$$\phi(s_1, \ldots, s_n) = \begin{cases} 1 & \text{if the system works under state vector } s_1, \ldots, s_n \\ 0 & \text{otherwise} \end{cases}$$

The function $\phi(s_1, \ldots, s_n)$ is called the structure function.

Some common structure functions are the following:

(a) *The series structure*: For the series structure

$$\phi(s_1, \ldots, s_n) = \min_i s_i$$

The series system works only if all its components function.

(b) *The parallel structure*: For the parallel structure

$$\phi(s_1, \ldots, s_n) = \max_i s_i$$

Hence the parallel system works if at least one of its components works.

(c) *The k-of-n system*: The structure function

$$\phi(s_1, \ldots, s_n) = \begin{cases} 1 & \text{if } \sum_{i=1}^n s_i \geqslant k \\ 0 & \text{otherwise} \end{cases}$$

is called a k-of-n structure function. Since $\sum_{i=1}^n s_i$ represents the number of functioning components, a k-of-n system works if at least k of the n components are working.

It should be noted that a series system is an n-of-n system, whereas a parallel system is a 1-of-n system.

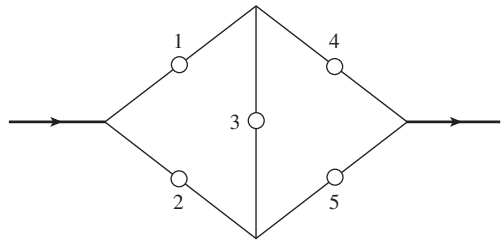

Figure 9.1. The bridge structure.

(d) *The bridge structure*: A five-component system for which

$$\phi(s_1, s_2, s_3, s_4, s_5) = \text{Max}(s_1 s_3 s_5, s_2 s_3 s_4, s_1 s_4, s_2 s_5)$$

is said to have a bridge structure. Such a system can be represented schematically by Figure 9.1. The idea of the diagram is that the system functions if a signal can go, from left to right, through the system. The signal can go through any given node i provided that component i is functioning. We leave it as an exercise for the reader to verify the formula given for the bridge structure function.

Let us suppose now that the states of the components—call them S_i—$i = 1, \ldots n$, are independent random variables such that

$$P\{S_i = 1\} = p_i = 1 - P\{S_i = 0\} \quad i = 1, \ldots, n$$

Let

$$\begin{aligned} r(p_1, \ldots, p_n) &= P\{\phi(S_1, \ldots, S_n) = 1\} \\ &= E[\phi(S_1, \ldots, S_n)] \end{aligned}$$

The function $r(p_1, \ldots, p_n)$ is called the *reliability* function. It represents the probability that the system will work when the components are independent with component i functioning with probability $p_i, i = 1, \ldots, n$.

For a series system

$$\begin{aligned} r(p_1, \ldots, p_n) &= P\{S_i = 1 \quad \text{for all } i = 1, \ldots, n\} \\ &= \prod_{i=1}^{n} P\{S_i = 1\} \\ &= \prod_{i=1}^{n} p_i \end{aligned}$$

and for a parallel system

$$\begin{aligned} r(p_1, \ldots, p_n) &= P\{S_i = 1 \quad \text{for at least one } \ i, i = 1, \ldots, n\} \\ &= 1 - P\{S_i = 0 \quad \text{for all } i = 1, \ldots, n\} \\ &= 1 - \prod_{i=1}^{n} P(S_i = 0) \\ &= 1 - \prod_{i=1}^{n} (1 - p_i) \end{aligned}$$

However, for most systems it remains a formidable problem to compute the reliability function (even for such small systems as a 5-of-10 system or the bridge system it can be quite tedious to compute). So let us suppose that for a given nondecreasing structure function ϕ and given probabilities $p_1, \ldots, p_n$, we are interested in using simulation to estimate

$$r(p_1, \ldots, p_n) = E[\phi(S_1, \ldots, S_n)]$$

Now we can simulate the S_i by generating uniform random numbers $U_1, \ldots, U_n$ and then setting

$$S_i = \begin{cases} 1 & \text{if } U_i < p_i \\ 0 & \text{otherwise} \end{cases}$$

Hence we see that

$$\phi(S_1, \ldots, S_m) = h(U_1, \ldots, U_n)$$

where h is a decreasing function of $U_1, \ldots, U_n$. Therefore

$$\text{Cov}(h(\mathbf{U}), h(\mathbf{1} - \mathbf{U})) \leqslant 0$$

and so the antithetic variable approach of using $U_1, \ldots, U_n$ to generate both $h(U_1, \ldots, U_n)$ and $h(1 - U_1, \ldots, 1 - U_n)$ results in a smaller variance than if an independent set of random numbers were used to generate the second value of h. □

Oftentimes the relevant output of a simulation is a function of the input random variables $Y_1, \ldots, Y_m$. That is, the relevant output is $X = h(Y_1, \ldots, Y_m)$. Suppose Y_i has distribution F_i, $i = 1, \ldots, m$. If these input variables are generated by the inverse transform technique, we can write

$$X = h\left(F_1^{-1}(U_1), \ldots, F_m^{-1}(U_m)\right)$$

where $U_1, \ldots, U_m$ are independent random numbers. Since a distribution function is increasing, it follows that its inverse is also increasing and thus if $h(y_1, \ldots, y_m)$ were a monotone function of its coordinates, then it follows that $h(F_1^{-1}(U_1), \ldots, F_m^{-1}(U_m))$ will be a monotone function of the U_i. Hence the method of antithetic variables, which would first generate $U_1, \ldots, U_m$ to compute X_1 and then use $1 - U_1, \ldots, 1 - U_m$ to compute X_2, would result in an estimator having a smaller variance than would have been obtained if a new set of random numbers were used for X_2.

Example 9c Simulating a Queueing System Consider a given queueing system, let D_i denote the delay in queue of the ith arriving customer, and suppose we are interested in simulating the system so as to estimate $\theta = E[\mathrm{X}]$, where

$$X = D_1 + \cdots + D_n$$

is the sum of the delays in queue of the first n arrivals. Let $I_1, \ldots, I_n$ denote the first n interarrival times (i.e., I_j is the time between the arrivals of customers $j - 1$ and j), and let $S_1, \ldots, S_n$ denote the first n service times of this system, and suppose that these random variables are all independent. Now in many systems X is a function of the $2n$ random variables $I_1, \ldots, I_n, S_1, \ldots, S_n$, say,

$$X = h(I_1, \ldots, I_n, S_1, \ldots, S_n)$$

Also, as the delay in queue of a given customer usually increases (depending of course on the specifics of the model) as the service times of other customers increase and usually decreases as the times between arrivals increase, it follows that, for many models, h is a monotone function of its coordinates. Hence, if the inverse transform method is used to generate the random variables $I_1, \ldots, I_n, S_1, \ldots, S_n$, then the antithetic variable approach results in a smaller variance. That is, if we initially use the $2n$ random numbers $U_i, i = 1, \ldots, 2n$, to generate the interarrival and service times by setting $I_i = F_i^{-1}(U_i)$, $S_i = G_i^{-1}(U_{n+i})$, where F_i and G_i are, respectively, the distribution functions of I_i and S_i, then the second simulation run should be done in the same fashion, but using the random numbers $1 - U_i, i = 1, \ldots, 2n$. This results in a smaller variance than if a new set of $2n$ random numbers were generated for the second run. □

The following example illustrates the sort of improvement that can sometimes be gained by the use of antithetic variables.

Example 9d Suppose we were interested in using simulation to estimate

$$\theta = E[e^U] = \int_0^1 e^x \, dx$$

(Of course, we know that $\theta = e - 1$; however, the point of this example is to see what kind of improvement is possible by using antithetic variables.) Since the function $h(u) = e^u$ is clearly a monotone function, the antithetic variable approach leads to a variance reduction, whose value we now determine. To begin, note that

$$\begin{aligned} \text{Cov}(e^U, e^{1-U}) &= E[e^U e^{1-U}] - E[e^U]E[e^{1-U}] \\ &= e - (e-1)^2 = -0.2342 \end{aligned}$$

Also, because

$$\begin{aligned} \text{Var}(e^U) &= E[e^{2U}] - (E[e^U])^2 \\ &= \int_0^1 e^{2x} \, dx - (e-1)^2 \\ &= \frac{e^2 - 1}{2} - (e-1)^2 = 0.2420 \end{aligned}$$

we see that the use of independent random numbers results in a variance of

$$\text{Var}\left(\frac{\exp\{U_1\} + \exp\{U_2\}}{2}\right) = \frac{\text{Var}(e^U)}{2} = 0.1210$$

whereas the use of the antithetic variables U and $1 - U$ gives a variance of

$$\text{Var}\left(\frac{e^U + e^{1-U}}{2}\right) = \frac{\text{Var}(e^U)}{2} + \frac{\text{Cov}(e^U, e^{1-U})}{2} = 0.0039$$

a variance reduction of 96.7 percent. □

Example 9e Estimating e Consider a sequence of random numbers and let N be the first one that is greater than its immediate predecessor. That is,

$$N = \min(n : n \geqslant 2, U_n > U_{n-1})$$

Now,

$$\begin{aligned} P\{N > n\} &= P\{U_1 \geqslant U_2 \geqslant \cdots \geqslant U_n\} \\ &= 1/n! \end{aligned}$$

where the final equality follows because all possible orderings of $U_1, \ldots, U_n$ are equally likely. Hence,

$$P\{N = n\} = P\{N > n-1\} - P\{N > n\} = \frac{1}{(n-1)!} - \frac{1}{n!} = \frac{n-1}{n!}$$

and so

$$E[N] = \sum_{n=2}^{\infty} \frac{1}{(n-2)!} = e$$

Also,

$$E[N^2] = \sum_{n=2}^{\infty} \frac{n}{(n-2)!} = \sum_{n=2}^{\infty} \frac{2}{(n-2)!} + \sum_{n=2}^{\infty} \frac{n-2}{(n-2)!}$$
$$= 2e + \sum_{n=3}^{\infty} \frac{1}{(n-3)!} = 3e$$

and so

$$\text{Var}(N) = 3e - e^2 \approx 0.7658$$

Hence, e can be estimated by generating random numbers and stopping the first time one exceeds its immediate predecessor.

If we employ antithetic variables, then we could also let

$$M = \min(n : n \geqslant 2, 1 - U_n > 1 - U_{n-1}) = \min(n : n \geqslant 2, U_n < U_{n-1})$$

Since one of the values of N and M will equal 2 and the other will exceed 2, it would seem, even though they are not monotone functions of the U_n, that the estimator $(N + M)/2$ should have a smaller variance than the average of two independent random variables distributed according to N. Before determining $\text{Var}(N + M)$, it is useful to first consider the random variable N_a, whose distribution is the same as the conditional distribution of the number of additional random numbers that must be observed until one is observed greater than its predecessor, given that $U_2 \leqslant U_1$. Therefore, we may write

$$N = 2, \quad \text{with probability } \frac{1}{2}$$
$$N = 2 + N_a, \quad \text{with probability } \frac{1}{2}$$

Hence,

$$E[N] = 2 + \frac{1}{2}E[N_a]$$
$$E[N^2] = \frac{1}{2}4 + \frac{1}{2}E[(2 + N_a)^2]$$
$$= 4 + 2E[N_a] + \frac{1}{2}E\left[N_a^2\right]$$

Using the previously obtained results for $E[N]$ and $\text{Var}(N)$ we obtain, after some algebra, that

$$E[N_a] = 2e - 4$$
$$E\left[N_a^2\right] = 8 - 2e$$

implying that

$$\text{Var}(N_a) = 14e - 4e^2 - 8 \approx 0.4997$$

Now consider the random variable N and M. It is easy to see that after the first two random numbers are observed, one of N and M will equal 2 and the other will equal 2 plus a random variable that has the same distribution as N_a. Hence,

$$\text{Var}(N + M) = \text{Var}(4 + N_a) = \text{Var}(N_a)$$

Hence,

$$\frac{\text{Var}(N_1 + N_2)}{\text{Var}(N + M)} \approx \frac{1.5316}{0.4997} \approx 3.065$$

Thus, the use of antithetic variables reduces the variance of the estimator by a factor of slightly more than 3. □

In the case of a normal random variable having mean μ and variance σ^2, we can use the antithetic variable approach by first generating such a random variable Y and then taking as the antithetic variable $2\mu - Y$, which is also normal with mean μ and variance σ^2 and is clearly negatively correlated with Y. If we were using simulation to compute $E[h(Y_1, \ldots, Y_n)]$, where the Y_i are independent normal random variables with means $\mu_i, i = 1, \ldots, n$, and h is a monotone function of its coordinates, then the antithetic approach of first generating the n normals $Y_1, \ldots, Y_n$ to compute $h(Y_1, \ldots, Y_n)$ and then using the antithetic variables $2\mu_i - Y_i, i = 1, \ldots, n$, to compute the next simulated value of h would lead to a reduction in variance as compared with generating a second set of n normal random variables.

9.2 The Use of Control Variates

Again suppose that we want to use simulation to estimate $\theta = E[X]$, where X is the output of a simulation. Now suppose that for some other output variable Y, the expected value of Y is known—say, $E[Y] = \mu_y$. Then for any constant c, the quantity

$$X + c(Y - \mu_y)$$

is also an unbiased estimator of θ. To determine the best value of c, note that

$$\begin{aligned}\text{Var}(X + c(Y - \mu_y)) &= \text{Var}(X + cY) \\ &= \text{Var}(X) + c^2\,\text{Var}(Y) + 2c\,\text{Cov}(X, Y)\end{aligned}$$

Simple calculus now shows that the above is minimized when $c = c^*$, where

$$c^* = -\frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \tag{9.1}$$

and for this value the variance of the estimator is

$$\text{Var}(X + c^*(Y - \mu_y)) = \text{Var}(X) - \frac{[\text{Cov}(X, Y)]^2}{\text{Var}(Y)} \tag{9.2}$$

The quantity Y is called a *control variate* for the simulation estimator X. To see why it works, note that c^* is negative (positive) when X and Y are positively (negatively) correlated. So suppose that X and Y were positively correlated, meaning, roughly, that X is large when Y is large and vice versa. Hence, if a simulation run results in a large (small) value of Y—which is indicated by Y being larger (smaller) than its known mean μ_y—then it is probably true that X is also larger (smaller) than its mean θ, and so we would like to correct for this by lowering (raising) the value of the estimator X, and this is done since c^* is negative (positive). A similar argument holds when X and Y are negatively correlated.

Upon dividing Equation (9.2) by $\text{Var}(X)$, we obtain that

$$\frac{\text{Var}(X + c^*(Y - \mu_y))}{\text{Var}(X)} = 1 - \text{Corr}^2(X, Y)$$

where

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

is the correlation between X and Y. Hence, the variance reduction obtained in using the control variate Y is $100\,\text{Corr}^2(X, Y)$ percent.

The quantities $\text{Cov}(X, Y)$ and $\text{Var}(Y)$ are usually not known in advance and must be estimated from the simulated data. If n simulation runs are performed, and the output data $X_i, Y_i, i = 1, \ldots, n$, result, then using the estimators

$$\widehat{\text{Cov}}(X, Y) = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})/(n - 1)$$

and

$$\widehat{\text{Var}}(Y) = \sum_{i=1}^{n} (Y_i - \overline{Y})^2/(n - 1),$$

we can approximate c^* by $\hat{c}^*$, where

$$\hat{c}^* = -\frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}.$$

The variance of the controlled estimator

$$\text{Var}(\overline{X} + c^*(\overline{Y} - \mu_y)) = \frac{1}{n}\left(\text{Var}(X) - \frac{\text{Cov}^2(X, Y)}{\text{Var}(Y)}\right)$$

can then be estimated by using the estimator of $\text{Cov}(X, Y)$ along with the sample variance estimators of $\text{Var}(X)$ and $\text{Var}(Y)$.

Remark Another way of doing the computations is to make use of a standard computer package for simple linear regression models. For if we consider the simple linear regression model

$$X = a + bY + e$$

where e is a random variable with mean 0 and variance σ^2, then $\hat{a}$ and $\hat{b}$, the least squares estimators of a and b based on the data $X_i, Y_i, i = 1, \ldots, n$, are

$$\hat{b} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$
$$\hat{a} = \overline{X} - \hat{b}\overline{Y}$$

Therefore, $\hat{b} = -\hat{c}^*$. In addition, since

$$\overline{X} + \hat{c}^*(\overline{Y} - \mu_y) = \overline{X} - \hat{b}(\overline{Y} - \mu_y)$$
$$= \hat{a} + \hat{b}\mu_y$$

it follows that the control variate estimate is the evaluation of the estimated regression line at the value $Y = \mu_y$. Also, because $\hat{\sigma}^2$, the regression estimate of σ^2, is the estimate of $\mathrm{Var}(X - \hat{b}Y) = \mathrm{Var}(X + \hat{c}^*Y)$, it follows that the estimated variance of the control variate estimator $\overline{X} + \hat{c}^*(\overline{Y} - \mu_y)$ is $\hat{\sigma}^2/n$. □

Example 9f Suppose, as in Example 9b, that we wanted to use simulation to estimate the reliability function

$$r(p_1, \ldots, p_n) = E[\phi(S_1, \ldots, S_n)]$$

where

$$S_i = \begin{cases} 1 & \text{if } U_i < p_i \\ 0 & \text{otherwise} \end{cases}$$

Since $E[S_i] = p_i$, it follows that

$$E\left[\sum_{i=1}^{n} S_i\right] = \sum_{i=1}^{n} p_i$$

Hence, we can use the number of working components, $Y \equiv \sum S_i$, as a control variate of the estimator $X \equiv \phi(S_1, \ldots, S_n)$. Since $\sum_{i=1}^{n} S_i$ and $\phi(S_1, \ldots, S_n)$ are both increasing functions of the S_i, they are positively correlated, and thus the sign of c^* is negative. □

Example 9g Consider a queueing system in which customers arrive in accordance with a nonhomogeneous Poisson process with intensity function $\lambda(s), s > 0$. Suppose that the service times are independent random variables

having distribution G and are also independent of the arrival times. Suppose we were interested in estimating the total time spent in the system by all customers arriving before time t. That is, if we let W_i denote the amount of time that the ith entering customer spends in the system, then we are interested in $\theta = E[X]$, where

$$X = \sum_{i=1}^{N(t)} W_i$$

and where $N(t)$ is the number of arrivals by time t. A natural quantity to use as a control in this situation is the total of the service times of all these customers. That is, let S_i denote the service time of the ith customer and set

$$Y = \sum_{i=1}^{N(t)} S_i$$

Since the service times are independent of $N[t]$, it follows that

$$E[Y] = E[S]E[N(t)]$$

where $E[S]$, the mean service time, and $E[N(t)]$, the mean number of arrivals by t, are both known quantities. □

Example 9h As in Example 9d, suppose we were interested in using simulation to compute $\theta = E[e^U]$. Here, a natural variate to use as a control is the random number U. To see what sort of improvement over the raw estimator is possible, note that

$$\begin{aligned}\text{Cov}(e^U, U) &= E[Ue^U] - E[U]E[e^U] \\ &= \int_0^1 xe^x\,dx - \frac{(e-1)}{2} \\ &= 1 - \frac{(e-1)}{2} = 0.14086\end{aligned}$$

Because $\text{Var}(U) = \frac{1}{12}$ it follows from (9.2) that

$$\begin{aligned}\text{Var}\left(e^U + c^*\left(U - \frac{1}{2}\right)\right) &= \text{Var}(e^U) - 12(0.14086)^2 \\ &= 0.2420 - 0.2380 = 0.0039\end{aligned}$$

where the above used, from Example 9d, that $\text{Var}(e^U) = 0.2420$. Hence, in this case, the use of the control variate U can lead to a variance reduction of up to 98.4 percent. □

Example 9i A List Recording Problem Suppose we are given a set of n elements, numbered 1 through n, which are to be arranged in an ordered list. At each unit of time a request is made to retrieve one of these elements, with the request being for element i with probability $p(i)$, $\sum_{i=1}^{n} p(i) = 1$. After being requested, the element is put back in the list but not necessarily in the same position. For example, a common reordering rule is to interchange the requested element with the one immediately preceding it. Thus, if $n = 4$ and the present ordering is 1, 4, 2, 3, then under this rule a request for element 2 would result in the reorder 1, 2, 4, 3. Starting with an initial ordering that is equally likely to be any of the $n!$ orderings and using this interchange rule, suppose we are interested in determining the expected sum of the positions of the first N elements requested. How can we efficiently accomplish this by simulation?

One effective way is as follows. The "natural" way of simulating the above is first to generate a random permutation of $1, 2, \ldots, n$ to establish the initial ordering, and then at each of the next N periods determine the element requested by generating a random number U and then letting the request be for element j if $\sum_{k=1}^{j-1} p(k) < U \leqslant \sum_{k=1}^{j} p(k)$. However, a better technique is to generate the element requested in such a way that small values of U correspond to elements close to the front. Specifically, if the present ordering is $i_1, i_2, \ldots, i_n$, then generate the element requested by generating a random number U and then letting the selection be for i_j if $\sum_{k=1}^{j-1} p(i_k) < U \leqslant \sum_{k=1}^{j} p(i_k)$. For example, if $n = 4$ and the present ordering is 3, 1, 2, 4, then we should generate U and let the selection be for 3 if $U \leqslant p(3)$, let it be for 1 if $p(3) < U \leqslant p(3) + p(1)$, and so on. As small values of U thus correspond to elements near the front, we can use $\sum_{r=1}^{N} U_r$ as a control variable, where U_r is the random number used for the rth request in a run. That is, if P_r is the position of the rth selected element in a run, then rather than just using the raw estimator $\sum_{r=1}^{N} P_r$ we should use

$$\sum_{r=1}^{N} P_r + c^* \left(\sum_{r=1}^{N} U_r - \frac{N}{2} \right)$$

where

$$c^* = -\frac{\text{Cov}\left(\sum_{r=1}^{N} P_r, \sum_{r=1}^{N} U_r \right)}{\frac{N}{12}}$$

and where the above covariance should be estimated using the data from all the simulated runs.

Although the variance reduction obtained will, of course, depend on the probabilities $p(i), i = 1, \ldots, n$, and the value of N, a small study indicates that when $n = 50$ and the $p(i)$ are approximately equal, then for $15 \leqslant N \leqslant 50$ the variance of the controlled estimator is less than $\frac{1}{2400}$ the variance of the raw simulation estimator. □

Of course, one can use more than a single variable as a control. For example, if a simulation results in output variables $Y_i, i = 1, \ldots, k$, and $E[Y_i] = \mu_i$ is known, then for any constants $c_i, i = 1, \ldots, k$, we may use

$$X + \sum_{i=1}^{k} c_i(Y_i - \mu_i)$$

as an unbiased estimator of $E[X]$.

Example 9j Blackjack The game of blackjack is often played with the dealer shuffling multiple decks of cards, putting aside used cards, and finally reshuffling when the number of remaining cards is below some limit. Let us say that a new round begins each time the dealer reshuffles, and suppose we are interested in using simulation to estimate $E[X]$, a player's expected winnings per round, where we assume that the player is employing some fixed strategy which might be of the type that "counts cards" that have already been played in the round and stakes different amounts depending on the "count." We will assume that the game consists of a single player against the dealer.

The randomness in this game results from the shuffling of the cards by the dealer. If the dealer uses k decks of 52 cards, then we can generate the shuffle by generating a random permutation of the numbers 1 through $52k$; let $I_1, \ldots, I_{52k}$ denote this permutation. If we now set

$$u_j = I_j \bmod 13 + 1$$

and let

$$v_j = \min(u_j, 10)$$

then $v_j, j = 1, \ldots, 52k$ represents the successive values of the shuffled cards, with 1 standing for an ace.

Let N denote the number of hands played in a round, and let B_j denote the amount bet on hand j. To reduce the variance, we can use a control variable that is large when the player is dealt more good hands than the dealer, and is small in the reverse case. Since being dealt 19 or better is good, let us define

$$W_j = 1 \text{ if the player's two dealt cards on deal } j \text{ add to atleast } 19$$

and let W_j be 0 otherwise. Similarly, let

$$Z_j = 1 \text{ if the dealer's two dealt cards on deal } j \text{ add to atleast } 19$$

and let Z_j be 0 otherwise. Since W_j and Z_j clearly have the same distribution it follows that $E[W_j - Z_j] = 0$, and it is not difficult to show that

$$E\left[\sum_{j=1}^{N} B_j(W_j - Z_j)\right] = 0$$

Thus, we recommend using $\sum_{j=1}^{N} B_j(W_j - Z_j)$ as a control variable. Of course, it is not clear that 19 is the best value, and one should experiment on letting 18 or even 20 be the critical value. However, some preliminary work indicates that 19 works best, and it has resulted in variance reductions of 15 percent or more depending on the strategy employed by the player. An even greater variance reduction should result if we use two control variables. One control variable is defined as before, with the exception that the W_j and Z_j are defined to be 1 if the hand is either 19 or 20. The second variable is again similar, but this time its indicators are 1 when the hands consist of blackjacks. □

When multiple control variates are used, the computations can be performed by using a computer program for the multiple linear regression model

$$X = a + \sum_{i=1}^{k} b_i Y_i + e$$

where e is a random variable with mean 0 and variance σ^2. Letting $\hat{c}_i^*$ be the estimate of the best c_i, for $i = 1, \ldots, k$, then

$$\hat{c}_i^* = -\hat{b}_i, \quad i = 1, \ldots, k$$

where $\hat{b}_i, i = 1, \ldots, k$, are the least squares regression estimates of $b_i, i = 1, \ldots, k$. The value of the controlled estimate can be obtained from

$$\overline{X} + \sum_{i=1}^{k} \hat{c}_i^*(\overline{Y}_i - \mu_i) = \hat{a} + \sum_{i=1}^{k} \hat{b}_i \mu_i$$

That is, the controlled estimate is just the estimated multiple regression line evaluated at the point $(\mu_1, \ldots, \mu_k)$.

The variance of the controlled estimate can be obtained by dividing the regression of σ^2 by the number of simulation runs.

Remarks

1. Since the variance of the controlled estimator is not known in advance, one often performs the simulation in two stages. In the first stage a small number of runs are performed so as to give a rough estimate of $\text{Var}(X + c^*(Y - \mu_y))$. (This estimate can be obtained from a simple linear regression program, where Y is the independent and X is the dependent variable, by using the estimate of σ^2.) We can then fix the number of trials needed in the second run so that the variance of the final estimator is within an acceptable bound.
2. A valuable way of interpreting the control variable approach is that it combines estimators of θ. That is, suppose the values of X and W are both determined by the simulation, and suppose $E[X] = E[W] = \theta$. Then we may consider any unbiased estimator of the form

$$\alpha X + (1 - \alpha)W$$

The best such estimator, which is obtained by choosing α to minimize the variance, is given by letting $\alpha = \alpha^*$, where

$$\alpha^* = \frac{\text{Var}(W) - \text{Cov}(X, W)}{\text{Var}(X) + \text{Var}(W) - 2\,\text{Cov}(X, W)} \tag{9.3}$$

Now if $E[Y] = \mu_y$ is known, we have the two unbiased estimators X and $X + Y - \mu_y$. The combined estimator can then be written as

$$(1 - c)X + c(X + Y - \mu_y) = X + c(Y - \mu_y)$$

To go the other way in the equivalence between control variates and combining estimators, suppose that $E[X] = E[W] = \theta$. Then if we use X, controlling with the variable $Y = X - W$, which is known to have mean 0, we then obtain an estimator of the form

$$X + c(X - W) = (1 + c)X - cW$$

which is a combined estimator with $\alpha = 1 + c$.

3. With the interpretation given in Remark 2, the antithetic variable approach may be regarded as a special case of control variables. That is, if $E[X] = \theta$, where $X = h(U_1, \ldots, U_n)$, then also $E[W] = \theta$, where $W = h(1 - U_1, \ldots, 1 - U_n)$. Hence, we can combine to get an estimator of the form $\alpha X + (1 - \alpha)W$. Since $\text{Var}(X) = \text{Var}(W)$, as X and W have the same distribution, it follows from Equation (9.3) that the best value of α is $\alpha = \frac{1}{2}$, and this is the antithetic variable estimator.
4. Remark 3 indicates why it is not usually possible to effectively combine antithetic variables with a control variable. If a control variable Y has a large positive (negative) correlation with $h(U_1, \ldots, U_n)$ then it probably has a large negative (positive) correlation with $h(1 - U_1, \ldots, 1 - U_n)$. Consequently, it is unlikely to have a large correlation with the antithetic estimator $\frac{h(U_1,\ldots,U_n)+h(1-U_1,\ldots,1-U_n)}{2}$. □

9.3 Variance Reduction by Conditioning

Recall the conditional variance formula proved in Section 2.10 of Chapter 2.

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

Since both terms on the right are nonnegative, because a variance is always nonnegative, we see that

$$\text{Var}(X) \geqslant \text{Var}(E[X|Y]) \tag{9.4}$$

Now suppose we are interested in performing a simulation study so as to ascertain the value of $\theta = E[X]$, where X is an output variable of a simulation run. Also,

suppose there is a second variable Y, such that $E[X|Y]$ is known and takes on a value that can be determined from the simulation run. Since

$$E[E[X|Y]] = E[X] = \theta$$

it follows that $E[X|Y]$ is also an unbiased estimator of θ; thus, from (9.4) it follows that as an estimator of θ, $E[X|Y]$ is superior to the (raw) estimator X.

Remarks To understand why the conditional expectation estimator is superior to the raw estimator, note first that we are performing the simulation to estimate the unknown value of $E[X]$. We can now imagine that a simulation run proceeds in two stages: First, we observe the simulated value of the random variable Y and then the simulated value of X. However, if after observing Y we are now able to compute the (conditional) expected value of X, then by using this value we obtain an estimate of $E[X]$, which eliminates the additional variance involved in simulating the actual value of X. □

At this point one might consider further improvements by using an estimator of the type $\alpha X + (1 - \alpha) E[X|Y]$. However, by Equation (9.3) the best estimator of this type has $\alpha = \alpha^*$, where

$$\alpha^* = \frac{\text{Var}(E[X|Y]) - \text{Cov}(X, E[X|Y])}{\text{Var}(X) + \text{Var}(E[X|Y]) - 2\,\text{Cov}(X, E[X|Y])}$$

We now show that $\alpha^* = 0$, showing that combining the estimators X and $E[X|Y]$ does not improve on just using $E[X|Y]$.

First note that

$$\begin{aligned}\text{Var}(E[X|Y]) &= E[(E[X|Y])^2] - (E[E[X|Y]])^2 \\ &= E[(E[X|Y])^2] - (E[X])^2 \end{aligned} \tag{9.5}$$

On the other hand,

$$\begin{aligned}\text{Cov}(X, E[X|Y]) &= E[X\,E[X|Y]] - E[X]E[E[X|Y]] \\ &= E[X\,E[X|Y]] - (E[X])^2 \\ &= E[E[X\,E[X|Y]|Y]] - (E[X])^2 \\ &\quad \text{(conditioning on } Y) \\ &= E[E[X|Y]E[X|Y]] - (E[X])^2 \\ &\quad \text{(since given } Y,\ E[X|Y] \text{ is a constant)} \\ &= \text{Var}(E[X|Y]) \quad \text{[from (9.5)]}\end{aligned}$$

Thus, we see that no additional variance reduction is possible by combining the estimators X and $E[X|Y]$.

We now illustrate the use of "conditioning" by a series of examples.

Example 9k Let us reconsider our use of simulation to estimate π. In Example 3a of Chapter 3, we showed how we can estimate π by determining how often a randomly chosen point in the square of area 4 centered around the origin falls within the inscribed circle of radius 1. Specifically, if we let $V_i = 2U_i - 1$, where U_i, $i = 1, 2$, are random numbers, and set

$$I = \begin{cases} 1 & \text{if } V_1^2 + V_2^2 \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

then, as noted in Example 3a, EI = /4.

The use of the average of successive values of I to estimate /4 can be improved upon by using EIV_1 rather than I. Now

$$\begin{aligned} E[I|V_1 = v] &= P\left\{V_1^2 + V_2^2 \leqslant 1 | V_1 = v\right\} \\ &= P\left\{v^2 + V_2^2 \leqslant 1 | V_1 = v\right\} \\ &= P\left\{V_2^2 \leqslant 1 - v^2\right\} \quad \text{by the independence of } V_1 \text{ and } V_2 \\ &= P\{-(1-v^2)^{1/2} \leqslant V_2 \leqslant (1-v^2)^{1/2}\} \\ &= \int_{-(1-v^2)^{1/2}}^{(1-v^2)^{1/2}} \left(\frac{1}{2}\right) dx \quad \text{since } V_2 \text{ is uniform over } (-1, 1) \\ &= (1-v^2)^{1/2} \end{aligned}$$

Hence,

$$E[I|V_1] = \left(1 - V_1^2\right)^{1/2}$$

and so the estimator $(1 - V_1^2)^{1/2}$ also has mean $\pi/4$ and has a smaller variance than I. Since

$$P(V_1^2 \leqslant x) = P(-\sqrt{x} \leqslant V_1 \leqslant \sqrt{x}) = \sqrt{x} = P(U^2 \leqslant x)$$

it follows that V_1^2 and U^2 have the same distribution, and so we can simplify somewhat by using the estimator $(1 - U^2)^{1/2}$, where U is a random number.

The improvement in variance obtained by using the estimator $(1 - U^2)^{1/2}$ over the estimator I is easily determined.

$$\begin{aligned} \mathrm{Var}\left[(1-U^2)^{1/2}\right] &= E[1-U^2] - \left(\frac{\pi}{4}\right)^2 \\ &= \frac{2}{3} - \left(\frac{\pi}{4}\right)^2 \approx 0.0498 \end{aligned}$$

where the first equality used the identity $\mathrm{Var}(W) = E[W^2] - (E[W])^2$. On the other hand, because I is a Bernoulli random variable having mean $\pi/4$, we have

$$\text{Var}(I) = \left(\frac{\pi}{4}\right)\left(1 - \frac{\pi}{4}\right) \approx 0.1686$$

thus showing that conditioning results in a 70.44 percent reduction in variance. (In addition, only one rather than two random numbers is needed for each simulation run, although the computational cost of having to compute a square root must be paid.)

Since the function $(1 - u^2)^{1/2}$ is clearly a monotone decreasing function of u in the region $0 < u < 1$, it follows that the estimator $(1 - U^2)^{1/2}$ can be improved upon by using antithetic variables. That is, the estimator

$$\frac{1}{2}[(1 - U^2)^{1/2} + (1 - (1 - U)^2)^{1/2}]$$

has smaller variance than $\frac{1}{2}[(1 - U_1^2)^{1/2} + (1 - U_2^2)^{1/2}]$.

Another way of improving the estimator $(1-U^2)^{1/2}$ is by using a control variable. A natural control variable in this case is U^2 and, because $E[U^2] = \frac{1}{3}$, we could use an estimator of the type

$$(1 - U^2)^{1/2} + c\left(U^2 - \frac{1}{3}\right)$$

The best c—namely, $c^* = -\text{Cov}\,[(1 - U^2)^{1/2}, U^2)/\text{Var}(U^2)]$—can be estimated by using the simulation to estimate the covariance term. (We could also have tried to use U as a control variable; it makes a difference because a correlation between two random variables is only a measure of their "linear dependence" rather than of their total dependence. But the use of U^2 leads to a greater improvement; see Exercise 15.) □

Example 9l Suppose there are r types of coupons and that every new coupon collected is, independently of those previously collected, type i with probability p_i, $\sum_{i=1}^{r} p_i = 1$. Assume that coupons are collected one at a time, and that we continue to collect coupons until we have collected n_i or more type i coupons, for all $i = 1, \ldots, r$. With N denoting the number of coupons needed, we are interested in using simulation to estimate both $E[N]$ and $P(N > m)$.

To obtain efficient estimates, suppose that the times at which coupons are collected constitute the event times of a Poisson process with rate $\lambda = 1$. That is, an event of the Poisson process occurs whenever a new coupon is collected. Say that the Poisson event is of type i if the coupon collected is of type i. If we let $N_i(t)$ denote the number of type i events by time t (that is, $N_i(t)$ is the number of type i coupons that have been collected by time t), then it follows from results on Poisson random variables presented in Section 2.8 that the processes $\{N_i(t), t \geqslant 0\}$ are, for $i = 1, \ldots, r$, *independent* Poisson processes with respective rates p_i. Hence, if we let T_i be the time until there have been n_i type i coupons collected, then $T_1, \ldots, T_r$ are *independent* gamma random variables, with respective parameters (n_i, p_i). (It is to gain this independence that we supposed that coupons were collected at times

distributed as a Poisson process. For suppose we had defined M_i as the number of coupons one needs collect to obtain n_i type i coupons. Then, whereas M_i would have a negative binomial distribution with parameters (n_i, p_i), the random variables $M_1, \ldots, M_r$ would not be independent.)

To obtain estimates of $E[N]$, generate the random variables $T_1, \ldots, T_r$ and let $T = \max_i T_i$. Thus, T is the moment at which we have reached the goal of having collected at least n_i type i coupons for each $i = 1, \ldots, r$. Now, at time T_i a total of n_i type i coupons would have been collected. Because the additional number of type i coupons collected between times T_i and T would have a Poisson distribution with mean $p_i(T - T_i)$, it follows that $N_i(T)$, the total number of type i coupons collected, is distributed as n_i plus a Poisson random variable with mean $p_i(T - T_i)$. As the Poisson arrival processes are independent, it follows upon using that the sum of independent Poisson random variables is itself Poisson distributed that the conditional distribution of N given the values $T_1, \ldots, T_r$ is that of $\sum_i n_i$ plus a Poisson random variable with mean $\sum_{i=1}^r p_i(T - T_i)$. Thus, with $n = \sum_{i=1}^r n_i$ and $\mathbf{T} = (T_1, \ldots, T_r)$, we have that

$$\begin{aligned} E[N|\mathbf{T}] &= n + \sum_{i=1}^{r} p_i(T - T_i) \\ &= T + n - \sum_{i=1}^{r} p_i T_i \end{aligned} \tag{9.6}$$

In addition, because T is the time of event N, it follows that

$$T = \sum_{i=1}^{N} X_i$$

where $X_1, X_2, \ldots$are the interarrival times of the Poisson process, and are thus independent exponentials with mean 1. Because N is independent of the X_i the preceding identity gives that

$$E[T] = E[E[T|N]] = E[N E[X_i]] = E[N]$$

Hence, T is also an unbiased estimator of $E[N]$, suggesting a weighted average estimator:

$$\alpha E[N|\mathbf{T}] + (1 - \alpha)T = T + \alpha(n - \sum_{i=1}^{r} p_i T_i)$$

Because $E\left[\sum_{i=1}^r p_i T_i\right] = n$, this is equivalent to estimating $E[N]$ by using the unbiased estimator T along with the control variable $\sum_{i=1}^r p_i T_i$. That is, it is equivalent to using an estimator of the form

$$T + c\,(\sum_{i=1}^{r} p_i T_i - n)$$

with the value of c that minimizes the variance of the preceding, namely $c = -\frac{\text{Cov}(T, \sum_{i=1}^{r} p_i T_i)}{\text{Var}(\sum_{i=1}^{r} p_i T_i)}$, being estimated from the simulation data.

To estimate $P(N > m)$, again use that conditional on $\mathbf{T}$, N is distributed as $n + X$ where X is Poisson with mean $\lambda(\mathbf{T}) \equiv \sum_{i=1}^{r} p_i(T - T_i)$. This yields that

$$P(N > m|\mathbf{T}) = P(X > m - n) = 1 - \sum_{i=0}^{m-n} e^{-\lambda(\mathbf{T})}(\lambda(\mathbf{T}))^i / i!, \quad m \geqslant n$$

The preceding conditional probability $P(N > m|\mathbf{T})$should be used as the estimator of $P(N > m)$. □

In our next example we use the conditional expectation approach to efficiently estimate the probability that a compound random variable exceeds some fixed value.

Example 9m Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed positive random variables that are independent of the nonnegative integer valued random variable N. The random variable

$$S = \sum_{i=1}^{N} X_i$$

is said to be a *compound* random variable. In an insurance application, X_i could represent the amount of the ith claim made to an insurance company, and N could represent the number of claims made by some specified time t; S would be the total claim amount made by time t. In such applications, N is often assumed to be either a Poisson random variable (in which case S is called a *compound Poisson random variable*) or a mixed Poisson random variable, where we say that N is a mixed Poisson random variable if there is another random variable Λ, such that the conditional distribution of N, given that $\Lambda = \lambda$, is Poisson with mean λ. For instance, if Λ has a probability density function $g(\lambda)$, then the probability mass function of the mixed Poisson random variable N is

$$P\{N = n\} = \int_0^\infty \frac{e^{-\lambda}\lambda^n}{n!} g(\lambda)\, d\lambda$$

Mixed Poisson random variables arise when there is a randomly determined "environmental state" that determines the mean of the (Poisson) number of events that occur in the time period of interest. The distribution function of Λ is called the mixing distribution.

Suppose that we want to use simulation to estimate

$$p = P\left\{\sum_{i=1}^{N} X_i > c\right\}$$

for some specified positive constant c. The raw simulation approach would first generate the value of N, say $N = n$, then generate the values of $X_1, \ldots, X_n$ and use them to determine the value of the raw simulation estimator

$$I = \begin{cases} 1, & \text{if } \sum_{i=1}^{N} X_i > c \\ 0, & \text{otherwise} \end{cases}$$

The average value of I over many such runs would then be the estimator of p.

We can improve upon the preceding by a conditional expectation approach that starts by generating the values of the X_i in sequence, stopping when the sum of the generated values exceeds c. Let M denote the number that is needed; that is,

$$M = \min\left(n : \sum_{i=1}^{n} X_i > c\right)$$

If the generated value of M is m, then we use $P\{N \geqslant m\}$ as the estimate of p from this run. To see that this results in an estimator having a smaller variance than does the raw simulation estimator I, note that because the X_i are positive

$$I = 1 \Longleftrightarrow N \geqslant M$$

Hence,

$$E[I|M] = P\{N \geqslant M|M\}$$

Now,

$$P\{N \geqslant M|M = m\} = P\{N \geqslant m|M = m\} = P\{N \geqslant m\}$$

where the final equality used the independence of N and M. Consequently, if the value of M obtained from the simulation is $M = m$, then the value $E[I|M]$ obtained is $P\{N \geqslant m\}$.

The preceding conditional expectation estimator can be further improved by using a control variable. Let $\mu = E[X_i]$, and define

$$Y = \sum_{i=1}^{M} (X_i - \mu)$$

It can be shown that $E[Y] = 0$. To intuitively see why Y and the conditional expectation estimator $P\{N \geqslant M|M\}$ are strongly correlated, note first that the conditional expectation estimator will be small when M is large. But, because M is the number of the X_i that needs to be summed to exceed c, it follows that M will be large when the X_i are small, which would make Y small. That is, both $E[I|M]$ and Y tend to be small at the same time. A similar argument shows that if $E[I|M]$ is large then Y also tends to be large. Thus, it is clear that $E[I|M]$ and Y are strongly positively correlated, indicating that Y should be an effective control variable. □

Even though $E[X_i - \mu] = 0$ in Example 9m, because the number of terms in the sum $\sum_{i=1}^{M}(X_i - \mu)$ is random rather than fixed, it is not immediate that $E[\sum_{i=1}^{M}(X_i - \mu)] = 0$. That it is zero is a consequence of a result known as *Wald's equation*. To state this result we first need the concept of a stopping time for a sequence of random variables,

Definition: The nonnegative integer valued random variable N is said to be a *stopping time* for the sequence of random variables $X_1, X_2, \ldots$ if the event that $\{N = n\}$ is determined by the values of $X_1, \ldots, X_n$.

The idea behind a stopping time is that the random variables $X_1, X_2, \ldots$ are observed in sequence and at some point, depending on the values so far observed but not on future values, we stop. We now have

Wald's Equation: If N is a stopping time for a sequence of independent and identically distributed random variables $X_1, X_2, \ldots$ with finite mean $E[X]$ then

$$E\left[\sum_{n=1}^{N} X_n\right] = E[N]E[X]$$

provided that $E[N] < \infty$.

Example 9n A Finite Capacity Queueing Model Consider a queueing system in which arrivals enter only if there are fewer than N other customers in the system when they arrive. Any customer encountering N others upon arrival is deemed to be lost to the system. Suppose further that potential customers arrive in accordance with a Poisson process having rate λ; and suppose we are interested in using simulation to estimate the expected number of lost customers by some fixed time t.

A simulation run would consist of simulating the above system up to time t. If, for a given run, we let L denote the number of lost customers, then the average value of L, over all simulation runs, is the (raw) simulation estimator of the desired quantity $E[L]$. However, we can improve upon this estimator by conditioning upon the amount of time that the system is at capacity. That is, rather than using L, the actual number of lost customers up to time t, we consider $E[L|T_C]$, where T_C is the total amount of time in the interval $(0, t)$ that there are N customers in the system. Since customers are always arriving at the Poisson rate λ no matter what is happening within the system, it follows that

$$E[L|T_C] = \lambda T_C$$

Hence an improved estimator is obtained by ascertaining, for each run, the total time in which there are N customers in the system—say, $T_{C,i}$ is the time at capacity during the ith run. Then the improved estimator of $E[L]$ is $\lambda \sum_{i=1}^{k} T_{C,i}/k$, where k is the number of simulation runs. (In effect, since the expected number of lost customers given the time at capacity T_C is just λT_C, what this estimator does is

use the actual conditional expectation rather than simulating—and increasing the variance of the estimator—a Poisson random variable having this mean.)

If the arrival process were a nonhomogeneous Poisson process having intensity function $\lambda(s), 0 \leqslant s \leqslant t$, then we would not be able to compute the conditional expected number of lost customers if we were given only the total time at capacity. What we now need is the actual times at which the system was at capacity. So let us condition on the intervals during which the system was at capacity. Now letting N_C denote the number of intervals during $(0, t)$ during which the system is at capacity, and letting those intervals be designated by $I_1, \ldots, I_{N_C}$, then

$$E[L|N_C, I_1, \ldots, I_{N_C}] = \sum_{i=1}^{N_C} \int_{I_i} \lambda(s)\, ds$$

The use of the average value, over all simulation runs, of the above quantity leads to a better estimator—in the sense of having a smaller mean square error—of $E[L]$ than the raw simulation estimator of the average number lost per run.

One can combine the preceding with other variance reduction techniques in estimating $E[L]$. For instance, if we let M denote the number of customers that actually enter the system by time t, then with $N(t)$ equal to the number of arrivals by time t we have that

$$N(t) = M + L$$

Taking expectations gives that

$$\int_0^t \lambda(s)\, ds = E[M] + E[L]$$

Therefore, $\int_0^t \lambda(s)\, ds - M$ is also an unbiased estimator of $E[L]$, which suggests the use of the combined estimator

$$\alpha \sum_{i=1}^{N_C} \int_{I_i} \lambda(s)\, ds + (1 - \alpha) \left(\int_0^t \lambda(s)\, ds - M \right)$$

The value of α to be used is given by Equation (9.3) and can be estimated from the simulation. □

Example 9o Suppose we wanted to estimate the expected sum of the times in the system of the first n customers in a queueing system. That is, if W_i is the time that the ith customer spends in the system, we are interested in estimating

$$\theta = E\left[\sum_{i=1}^{n} W_i\right]$$

Let S_i denote the "state of the system" at the moment that the ith customer arrives, and consider the estimator

$$\sum_{i=1}^{n} E[W_i|S_i]$$

Since

$$E\left[\sum_{i=1}^{n} E[W_i|S_i]\right] = \sum_{i=1}^{n} E[E[W_i|S_i]] = \sum_{i=1}^{n} E[W_i] = \theta$$

it follows that this is an unbiased estimator of θ. It can be shown[1] that, in a wide class of models, this estimator has a smaller variance than the raw simulation estimator $\sum_{i=1}^{n} W_i$. (It should be noted that whereas it is immediate that $E[W_i|S_i]$ has smaller variance than W_i, this does not imply, because of the covariance terms, that $\sum_{i=1}^{n} E[W_i|S_i]$ has smaller variance than $\sum_{i=1}^{n} W_i$.)

The quantity S_i, which refers to the state of the system as seen by the ith customer upon its arrival, is supposed to represent the least amount of information that enables us to compute the conditional expected time that the customer spends in the system. For example, if there is a single server and the service times are all exponential with mean μ, then S_i would refer to N_i, the number of customers in the system encountered by the ith arrival. In this case,

$$E[W_i|S_i] = E[W_i|N_i] = (N_i + 1)\mu$$

which follows because the ith arrival will have to wait for N_i service times (one of which is the completion of service of the customer presently being served when customer i arrives—but, by the memoryless property of the exponential, that remaining time will also be exponential with mean μ) all having mean μ, and then to this we must add its own service time. Thus, the estimator that takes the average value, over all simulation runs, of the quantity $\sum_{i=1}^{n}(N_i + 1)\mu$ is a better estimator than the average value of $\sum_{i=1}^{n} W_i$. □

Our next example refers to the distribution of the number of nontarget cells that are not accidentally killed before a set of target cells have been destroyed.

Example 9p Consider a set of $n + m$ cells, with cell i having weight $w_i, i = 1, \ldots, n + m$. Imagine that cells $1, \ldots, n$ are cancerous and that cells $n + 1, \ldots, n + m$ are normal, and suppose cells are killed one at a time in the following fashion. If, at any time, S is the current set of cells that are still alive then, independent of the order in which the cells that are not in S have been killed, the next cell to be killed is cell $i, i \in S$, with probability $\frac{w_i}{\sum_{j \in S} w_j}$. Therefore, with probability $\frac{w_i}{\sum_{j=1}^{n+m} w_j}$ cell i is the first cell killed; given that cell i is the first cell killed, the next cell killed will be cell $k, k \neq i$, with probability $\frac{w_k}{\sum_{j \neq i} w_j}$, and so on. This process of killing cells continues until all of the first n cells (the cancer cells) have been killed. Let N denote the number of the normal cells that are still alive at the time when all the cancer cells have been killed. We are interested in determining $P\{N \geqslant k\}$.

[1] S. M. Ross, "Simulating Average Delay—Variance Reduction by Conditioning," *Probability Eng. Informational Sci.* **2**(3), 1988.

Before attempting to develop an efficient simulation procedure, let us consider a related model in which cell i, $i = 1, \ldots, n + m$, is killed at the random time T_i, where $T_1, \ldots, T_{n+m}$ are independent exponential random variables with respective rates $w_1, \ldots, w_{n+m}$. By the lack of memory property of exponential random variables, it follows that if S is the set of cells that are currently alive then, as in the original model, cell $i, i \in S$, will be the next cell killed with probability $\frac{w_i}{\sum_{j \in S} w_j}$, showing that the order in which cells are killed in this related model has the same probability distribution as it does in the original model. Let N represent the number of cells that are still alive when all cells $1, \ldots, n$ have been killed. Now, if we let $T^{(k)}$ be the k^{th} largest of the values $T_{n+1}, \ldots, T_{n+m}$, then $T^{(k)}$ is the first time at which there are fewer than k normal cells alive. Thus, in order for N to be at least k, all the cancer cells must have been killed by time $T^{(k)}$. That is,

$$P\{N \geqslant k\} = P\{\underset{i \leqslant n}{\text{Max}}\, T_i < T^{(k)}\}$$

Therefore,

$$\begin{aligned} P\{N \geqslant k | T^{(k)}\} &= \{P \underset{i \leqslant n}{\text{Max}}\, T_i < T^{(k)} | T^{(k)}\} \\ &= \prod_{i=1}^{n} (1 - e^{-w_i T^{(k)}}) \end{aligned}$$

where the final equality used the independence of the T_i. Hence, we obtain an unbiased, conditional expectation estimator of $P\{N \geqslant k\}$ by generating the m exponential random variables $T_{m+1}, \ldots, T_{n+m}$. Then letting $T^{(k)}$ be the k^{th} largest of these values gives the estimator $\prod_{i=1}^{n}(1 - e^{-w_i T^{(k)}})$. Because this estimator is an increasing function of the generated $T_{n+1}, \ldots, T_{n+m}$, further variance reduction is possible provided the T_i are obtained from the inverse transform method. For then the estimator will be an increasing function of the m random numbers used, indicating that antithetic variables will lead to further variance reduction. Putting it all together, the following gives a single run of the algorithm for estimating $P\{N \geqslant k\}$.

STEP 1: Generate random numbers $U_1, \ldots, U_m$.
STEP 2: Let $T^{(k)}$ be the k^{th} largest of the m values $-\frac{1}{w_{n+i}} \log(U_i)$, $i = 1, \ldots, m$.
STEP 3: Let $S^{(k)}$ be the k^{th} largest of the m values $-\frac{1}{w_{n+i}} \log(1 - U_i)$, $i = 1, \ldots, m$.
STEP 4: The estimator from this run is

$$\frac{1}{2}\left[\prod_{i=1}^{n}(1 - e^{-w_i T^{(k)}}) + \prod_{i=1}^{n}(1 - e^{-w_i S^{(k)}})\right]$$

Estimating the Expected Number of Renewals by Time *t*

Suppose that "events" are occurring randomly in time. Let T_1 denote the time of the first event, T_2 the time between the first and second event, and, in general, T_i the time between the $(i-1)$th and the ith event, $i \geqslant 1$. If we let

$$S_n = \sum_{i=1}^{n} T_i$$

the first event occurs at time S_1, the second at time S_2, and, in general, the nth event occurs at time S_n (see Figure 9.2). Let $N(t)$ denote the number of events that occur by time t; that is, $N(t)$ is the largest n for which the nth event occurs by time t, or, equivalently,

$$N(t) = \text{Max}\{n : S_n \leqslant t\}$$

If the interevent times $T_1, T_2 \ldots$ are independent and identically distributed according to some distribution function F, then the process $\{N(t), t \geqslant 0\}$ is called a *renewal process*.

A renewal process is easily simulated by generating the interarrival times. Suppose now that we wanted to use simulation to estimate $\theta = E[N(t)]$, the mean number of events by some fixed time t. To do so we would successively simulate the interevent times, keeping track of their sum (which represent the times at which events occur) until that sum exceeds t. That is, we keep on generating interevent times until we reach the first event time after t. Letting $N(t)$—the raw simulation estimator—denote the number of simulated events by time t, we find that a natural quantity to use as a control variable is the sequence of $N(t)+1$ interevent times that were generated. That is, if we let μ denote the mean interevent time, then as the random variables $T_i - \mu$ have mean 0 it follows from Wald's equation that

$$E\left[\sum_{i=1}^{N(t)+1} (T_i - \mu)\right] = 0$$

Hence, we can control by using an estimator of the type

$$\begin{aligned} N(t) + c\left[\sum_{i=1}^{N(t)+1} (T_i - \mu)\right] &= N(t) + c\left[\sum_{i=1}^{N(t)+1} T_i - \mu(N(t)+1)\right] \\ &= N(t) + c[S_{N(t)+1} - \mu N(t) - \mu] \end{aligned}$$

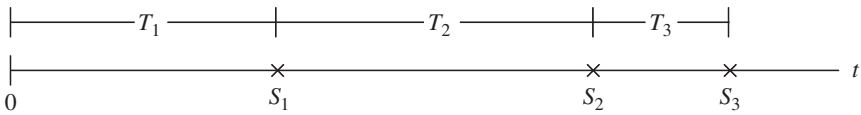

Figure 9.2. x = event.

Now since S_n represents the time of the nth event and $N(t) + 1$ represents the number of events by time t plus 1, it follows that $S_{N(t)+1}$ represents the time of the first event after time t. Hence, if we let $Y(t)$ denote the time from t until the next event [$Y(t)$ is commonly called the excess life at t], then

$$S_{N(t)+1} = t + Y(t)$$

and so the above controlled estimator can be written as

$$N(t) + c[t + Y(t) - \mu N(t) - \mu]$$

The best c is given by

$$c^* = -\frac{\text{Cov}[N(t), Y(t) - \mu N(t)]}{\text{Var}[Y(t) - \mu N(t)]}$$

Now for t large, it can be shown that the terms involving $N(t)$ dominate—because their variance will grow linearly with t, whereas the other terms will remain bounded—and so for t large

$$c^* \approx -\frac{\text{Cov}[N(t), -\mu N(t)]}{\text{Var}[-\mu N(t)]} = \frac{\mu \,\text{Var}[N(t)]}{\mu^2 \,\text{Var}[N(t)]} = \frac{1}{\mu}$$

Thus, for t large, the best controlled estimator of the above type is close to

$$N(t) + \frac{1}{\mu}(t + Y(t) - \mu N(t) - \mu) = \frac{Y(t)}{\mu} + \frac{t}{\mu} - 1 \tag{9.7}$$

In other words, for t large, the critical value to be determined from the simulation is $Y(t)$, the time from t until the next renewal.

The above estimator can further be improved upon by the use of "conditioning." Namely, rather than using the actual observed time of the first event after t, we can condition on $A(t)$, the time at t since the last event (see Figure 9.3). The quantity $A(t)$ is often called the age of the renewal process at t. [If we imagine a system consisting of a single item that functions for a random time having distribution F and then fails and is immediately replaced by a new item, then we have a renewal process with each event corresponding to the failure of an item. The variable $A(t)$ would then refer to the age of the item in use at time t, where by age we mean the amount of time it has already been in use.]

Now if the age of the process at time t is x, the expected remaining life of the item is just the expected amount by which an interevent time exceeds x given that it is greater than x. That is,

$$\begin{aligned} E[Y(t)|A(t) = x] &= E[T - x|T > x] \\ &= \int_x^\infty (y - x)\frac{f(y)\,dy}{1 - F(x)} \\ &\equiv \mu[x] \end{aligned}$$

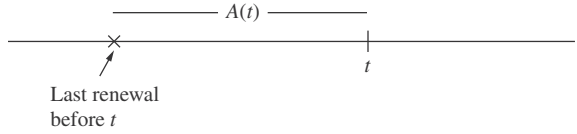

Figure 9.3. Age at t.

where the above supposes that F is a continuous distribution with density function f. Hence, with $\mu[x]$ defined as above to equal $E[T - x|T > x]$, we see that

$$E[Y(t)|A(t)] = \mu[A(t)]$$

Thus, for large t, a better estimator of $E[N(t)]$ than the one given in Equation (9.7) is

$$\frac{\mu[A(t)]}{\mu} + \frac{t}{\mu} - 1 \tag{9.8}$$

9.4 Stratified Sampling

Suppose we want to estimate $\theta = E[X]$, and suppose there is some discrete random variable Y, with possible values $y_1, \ldots, y_k$, such that

(a) the probabilities $p_i = P\{Y = y_i\}, i = 1, \ldots, k$, are known; and
(b) for each $i = 1, \ldots, k$, we can simulate the value of X conditional on $Y = y_i$.

Now if we are planning to estimate $E[X]$ by n simulation runs, then the usual approach would be to generate n independent replications of the random variable X and then use $\bar{X}$, their average, as the estimate of $E[X]$. The variance of this estimator is

$$\text{Var}(\bar{X}) = \frac{1}{n}\text{Var}(X)$$

However, writing

$$E[X] = \sum_{i=1}^{k} E[X|Y = y_i]p_i$$

we see that another way of estimating $E[X]$ is by estimating the k quantities $E[X|Y = y_i], i = 1, \ldots, k$. For instance, suppose rather than generating n independent replications of X, we do np_i of the simulations conditional on the event that $Y = y_i$ for each $i = 1, \ldots, k$. If we let $\bar{X}_i$ be the average of the np_i observed values of X generated conditional on $Y = y_i$, then we would have the unbiased estimator

$$\mathcal{E} = \sum_{i=1}^{k} \bar{X}_i p_i$$

The estimator $\mathcal{E}$ is called a *stratified sampling* estimator of $E[X]$.

Because $\bar{X}_i$ is the average of np_i independent random variables whose distribution is the same as the conditional distribution of X given that $Y = y_i$, it follows that

$$\text{Var}(\bar{X}_i) = \frac{\text{Var}(X|Y = y_i)}{np_i}$$

Consequently, using the preceding and that the $\bar{X}_i$, $i = 1, \ldots, k$, are independent, we see that

$$\begin{aligned}\text{Var}(\mathcal{E}) &= \sum_{i=1}^{k} p_i^2 \text{Var}(\bar{X}_i) \\ &= \frac{1}{n} \sum_{i=1}^{k} p_i \text{Var}(X|Y = y_i) \\ &= \frac{1}{n} E[\text{Var}(X|Y)]\end{aligned}$$

Because $\text{Var}(\bar{X}) = \frac{1}{n}\text{Var}(X)$, whereas $\text{Var}(\mathcal{E}) = \frac{1}{n}E[\text{Var}(X|Y)]$, we see from the conditional variance formula

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

that the variance savings in using the stratified sampling estimator $\mathcal{E}$ over the usual raw simulation estimator is

$$\text{Var}(\bar{X}) - \text{Var}(\mathcal{E}) = \frac{1}{n}\text{Var}(E[X|Y])$$

That is, the variance savings per run is $\text{Var}(E[X|Y])$ which can be substantial when the value of Y strongly affects the conditional expectation of X.

Remark The variance of the stratified sampling estimator can be estimated by letting S_i^2 be the sample variance of the np_i runs done conditional on $Y = y_i$, $i = 1, \ldots, k$. Then S_i^2 is an unbiased estimator of $\text{Var}(X|Y = y_i)$, yielding that $\frac{1}{n}\sum_{i=1}^{k} p_i S_i^2$ is an unbiased estimator of $\text{Var}(\mathcal{E})$. □

Example 9q On good days customers arrive at an infinite server queue according to a Poission process with rate 12 per hour, whereas on other days they arrive according to a Poisson process with rate 4 per hour. The service times, on all days, are exponentially distributed with rate 1 per hour. Every day at time 10 hours the system is shut down and all those presently in service are forced to leave without completing service. Suppose that each day is, independently, a good day with probability 0.5 and that we want to use simulation to estimate θ, the mean number of customers per day that do not have their services completed.

Let X denote the number of customers whose service is not completed on a randomly selected day; let Y equal 0 if the day is ordinary, and let it equal 1 if the day is good. Then it can be shown that the conditional distributions of X given that $Y = 0$ and that $Y = 1$ are, respectively, both Poisson with respective means

$$E[X|Y = 0] = 4(1 - e^{-10}), \quad E[X|Y = 1] = 12(1 - e^{-10})$$

Because the variance of a Poisson random variable is equal to its mean, the preceding shows that

$$\begin{aligned}\text{Var}(X|Y = 0) &= E[X|Y = 0] \approx 4\\ \text{Var}(X|Y = 1) &= E[X|Y = 1] \approx 12\end{aligned}$$

Thus,

$$E[\text{Var}(X|Y)] \approx \frac{1}{2}(4 + 12) = 8$$

and

$$\text{Var}(E[X|Y]) = E[(E[X|Y])^2] - (E[X])^2 \approx \frac{4^2 + (12)^2}{2} - 8^2 = 16$$

Consequently,

$$\text{Var}(X) \approx 8 + 16 = 24$$

which is about 3 times as large as $E[\text{Var}(X|Y)]$, the variance of the stratified sampling estimator that simulates exactly half the days as good days and the other half as ordinary days. □

Again suppose that the probability mass function $p_i = P\{Y = y_i\}, i = 1, \ldots, k$ is known, that we can simulate X conditional on $Y = i$, and that we plan to do n simulation runs. Although performing np_i of the n runs conditional on $Y = y_i, i = 1, \ldots, k$, (the so-called *proportional stratified sampling* strategy) is better than generating n independent replications of X, these are not necessarily the optimal numbers of conditional runs to perform. Suppose we plan to do n_i runs conditional on $Y = y_i$, where $n = \sum_{i=1}^{k} n_i$. Then, with $\bar{X}_i$ equal to the average of the n_i runs conditional on $Y = y_i$, the stratified sampling estimator is

$$\hat{\theta} = \sum_{i=1}^{k} p_i \bar{X}_i$$

with its variance given by

$$\text{Var}(\hat{\theta}) = \sum_{i=1}^{k} p_i^2 \text{Var}(X|Y = i)/n_i$$

Whereas the quantities $\text{Var}(X|Y = i), i = 1, \ldots, k$, will be initially unknown, we could perform a small simulation study to estimate them—say we use the estimators s_i^2. We could then choose the n_i by solving the following optimization problem:

$$\begin{aligned} &\text{choose} \quad n_1, \ldots, n_k \\ &\text{such that} \quad \sum_{i=1}^{k} n_i = n \\ &\text{to minimize} \quad \sum_{i=1}^{k} p_i^2 s_i^2 / n_i \end{aligned}$$

Using Lagrange multipliers, it is easy to show that the optimal values of the n_i in the preceding optimization problem are

$$n_i = n \frac{p_i s_i}{\sum_{j=1}^{k} p_j s_j}, \quad i = 1, \ldots, k$$

Once the n_i are determined and the simulations performed, we would estimate $E[X]$ by $\sum_{i=1}^{k} p_i \bar{X}_i$, and we would estimate the variance of this estimator by $\sum_{i=1}^{k} p_i^2 S_i^2 / n_i$, where S_i^2 is the sample variance of the n_i runs done conditional on $Y = y_i, i = 1, \ldots, k$.

For another illustration of stratified sampling, suppose that we want to use n simulation runs to estimate

$$\theta = E[h(U)] = \int_0^1 h(x)\, dx$$

If we let

$$S = j \quad \text{if } \frac{j-1}{n} \leqslant U < \frac{j}{n}, \quad j = 1, \ldots, n$$

then

$$\begin{aligned} \theta &= \frac{1}{n} \sum_{j=1}^{n} E[h(U)|S = j] \\ &= \frac{1}{n} \sum_{j=1}^{n} E[h(U_{(j)})] \end{aligned}$$

where $U_{(j)}$ is uniform on $((j-1)/n, j/n)$. Hence, by the preceding, it follows that rather than generating $U_1, \ldots, U_n$ and then using $\sum_{j=1}^{n} h(U_j)/n$ to estimate θ, a better estimator is obtained by using

$$\hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} h\left(\frac{U_j + j - 1}{n} \right)$$

Example 9r In Example 9k we showed that

$$\frac{\pi}{4} = E\left[\sqrt{(1-U^2)}\right]$$

Hence, we can estimate π by generating $U_1, \ldots, U_n$ and using the estimator

$$\text{est} = \frac{4}{n}\sum_{j=1}^{n}\sqrt{1-[(U_j+j-1)/n]^2}$$

In fact, we can improve the preceding by making use of antithetic variables to obtain the estimator

$$\hat{\pi} = \frac{2}{n}\sum_{j=1}^{n}\left(\sqrt{1-[(U_j+j-1)/n]^2}+\sqrt{1-[(j-U_j)/n]^2}\right)$$

A simulation using the estimator $\hat{\pi}$ yielded the following results:

n	$\hat{\pi}$
5	3.161211
10	3.148751
100	3.141734
500	3.141615
1000	3.141601
5000	3.141593

When $n = 5000$, the estimator $\hat{\pi}$ is correct to six decimal places. □

Remarks

1. Suppose we want to use simulation to estimate $E[X]$ by stratifying on the values of Y, a continuous random variable having distribution function G. To perform the stratification, we would first generate the value of Y and then simulate X conditional on this value of Y. Say we use the inverse transform method to generate Y; that is, we obtain Y by generating a random number U and setting $Y = G^{-1}(U)$. If we plan to do n simulation runs, then rather than using n independent random numbers to generate the successive values of Y, one could stratify by letting the i^{th} random number be the value of a random variable that is uniform in the region $\left(\frac{i-1}{n}, \frac{i}{n}\right)$. In this manner we obtain the value of Y in run i—call it Y_i—by generating a random number U_i and setting $Y_i = G^{-1}\left(\frac{U_i+i-1}{n}\right)$. We would then obtain X_i, the value of X in run i, by simulating X conditional on Y equal to the observed value of Y_i. The random variable X_i would then be an unbiased estimator of

$E[X|G^{-1}(\frac{i-1}{n}) < Y \leqslant G^{-1}(\frac{i}{n})]$, yielding that $\frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} E\left[X|G^{-1}\left(\frac{i-1}{n}\right) < Y \leqslant G^{-1}\left(\frac{i}{n}\right)\right] \\
&= \sum_{i=1}^{n} E\left[X|\frac{i-1}{n} < G(Y) \leqslant \frac{i}{n}\right]\frac{1}{n} \\
&= \sum_{i=1}^{n} E\left[X|\frac{i-1}{n} < G(Y) \leqslant \frac{i}{n}\right] P\left\{\frac{i-1}{n} < G(Y) \leqslant \frac{i}{n}\right\} \\
&= E[X]
\end{aligned}$$

where the penultimate equation used that $G(Y)$ is uniform on (0, 1).

2. Suppose that we have simulated n independent replications of X without doing any stratification, and that in n_i of the simulation runs, the resulting value of Y was y_i, $\sum_{i=1}^{k} n_i = n$. If we let $\bar{X}_i$ denote the average of the n_i runs in which $Y = y_i$, then $\bar{X}$, the average value of X over all n runs, can be written as

$$\begin{aligned}
\bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
&= \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_i \\
&= \sum_{i=1}^{k} \frac{n_i}{n}\bar{X}_i
\end{aligned}$$

When written this way, it is clear that using $\bar{X}$ to estimate $E[X]$ is equivalent to estimating $E[X|Y = i]$ by $\bar{X}_i$ and estimating p_i by n_i/n for each $i = 1, \ldots, k$. But since the p_i are known, and so need not be estimated, it would seem that a better estimator of $E[X]$ than $\bar{X}$ would be the estimator $\sum_{i=1}^{k} p_i \bar{X}_i$. In other words, we should act as if we had decided in advance to do stratified sampling, with n_i of our simulation runs to be done conditional on $Y = y_i, i = 1, \ldots, k$. This method of stratifying after the fact is called *poststratification*. □

At this point one might question how stratifying on the random variable Y compares with using Y as a control variable. The answer is that stratifying always results in an estimator having a smaller variance than is obtained by the estimator that uses Y as a control variable Because, from (9.9) and (9.2), the variance of the stratified estimator based on n runs is $\frac{1}{n}E[\text{Var}(X|Y)]$, whereas the variance from n runs of

using Y as a control variable is $\frac{1}{n}\left(\text{Var}(X) - \frac{\text{Cov}^2(X,Y)}{\text{Var}(Y)}\right)$, that this answer is true is shown by the following proposition.

Proposition

$$E[\text{Var}(X|Y)] \leqslant \text{Var}(X) - \frac{\text{Cov}^2(X, Y)}{\text{Var}(Y)}$$

To prove the following proposition we will first need a lemma..

Lemma

$$\text{Cov}(X,\ Y) = \text{Cov}(E[X|Y],\ Y)$$

Proof

$$\begin{aligned}
\text{Cov}(X,\ Y) &= E[XY] - E[X]\,E[Y] \\
&= E\,[E[XY|Y]] - E\,[E[X|Y]]\ E[Y] \\
&= E\,[Y\,E[X|Y]] - E\,[E[X|Y]]\ E[Y] \\
&= \text{Cov}(E[X|Y],\ Y)
\end{aligned}$$

where the preceding used that $E[XY|Y] = Y\,E[X|Y]$, which follows because conditional on Y the random variable Y can be treated as a constant. □

Proof of Proposition: By the conditional variance formula

$$E[\text{Var}(X|Y)] = \text{Var}(X) - \text{Var}(E[X|Y])$$

Hence, we must show that

$$\text{Var}(E[X|Y]) \geqslant \frac{\text{Cov}^2(X, Y)}{\text{Var}(Y)}$$

which, by the Lemma, is equivalent to showing that

$$1 \geqslant \frac{\text{Cov}^2(E[X|Y], Y)}{\text{Var}(Y)\text{Var}(E[X|Y])}$$

The result now follows because the right side of the preceding is $\text{Corr}^2(E[X|Y], Y)$, and the square of a correlation is always less than or equal to 1. □

Suppose again that we are interested in estimating $\theta = E[X]$, where X is dependent on the random variable S, which takes on one of the values $1, 2, \ldots k$ with respective probabilities $p_i, i = 1, \ldots, k$. Then

$$E[X] = p_1 E[X|S = 1] + p_2 E[X|S = 2] + \cdots + p_k E[X|S = k]$$

If all of the quantities $E[X|S = i]$ are known (that is, if $E[X|S]$ is known), but the p_i are not, then we can estimate θ by generating the value of S and then using the conditional expectation estimator $E[X|S]$. On the other hand, if it is the p_i that are known and we can generate from the conditional distribution of X given the value of S, then we can use simulation to obtain estimators $\widehat{E}[X|S = i]$ of the quantities $E[X|S = i]$ and then use the stratified sampling estimator $\sum_{i=1}^{k} p_i \widehat{E}[X|S = i]$ to estimate $E[X]$. When some of the p_i and some of the $E[X|S = i]$ are known, we can use a combination of these approaches.

Example 9s In the game of video poker a player inserts one dollar into a machine, which then deals the player a random hand of five cards. The player is then allowed to discard certain of these cards, with the discarded cards replaced by new ones from the remaining 47 cards. The player is then returned a certain amount depending on the makeup of her or his final cards. The following is a typical payoff scheme:

Hand	Payoff
Royal flush	800
Straight flush	50
Four of a kind	25
Full house	8
Flush	5
Straight	4
Three of a kind	3
Two pair	2
High pair (jacks or better)	1
Anything else	0

In the preceding, a hand is characterized as being in a certain category if it is of that type and not of any higher type. That is, for instance, by a flush we mean five cards of the same suit that are not consecutive.

Consider a strategy that never takes any additional cards (that is, the player stands pat) if the original cards constitute a straight or higher, and that always retains whatever pairs or triplets it is dealt. For a given strategy of this type let X denote the player's winnings on a single hand, and suppose we are interested in estimating $\theta = E[X]$. Rather than just using X as the estimator, let us start by conditioning on the type of hand that is initially dealt to the player. Let R represent a royal flush, S represent a straight flush, 4 represent four of a kind, 3 represent three of a kind, 2 represent two pair, 1 represent a high pair, 0 represent a low pair,

and "other" represent all other hands not mentioned. We then have

$$\begin{aligned}E[X] = {} & E[X|\mathrm{R}]P\{\mathrm{R}\} + E[X|\mathrm{S}]P\{\mathrm{S}\} + E[X|4]P\{4\} + E[X|\text{full}]P\{\text{full}\} \\ & + E[X|\text{flush}]P\{\text{flush}\} + E[X|\text{straight}]P\{\text{straight}\} + E[X|3]P\{3\} \\ & + E[X|2]P\{2\} + E[X|1]P\{1\} + E[X|0]P\{0\} + E[X|\text{other}]P\{\text{other}\}\end{aligned}$$

Now, with $C = \binom{52}{5}^{-1}$, we have

$$P\{\mathrm{R}\} = 4C = 1.539 \times 10^{-6}$$

$$P\{\mathrm{S}\} = 4 \cdot 9 \cdot C = 1.3851 \times 10^{-4}$$

$$P\{4\} = 13 \cdot 48 \cdot C = 2.40096 \times 10^{-4}$$

$$P\{\text{full}\} = 13 \cdot 12 \binom{4}{3}^{-1} \binom{4}{2}^{-1} C = 1.440576 \times 10^{-3}$$

$$P\{\text{flush}\} = 4\left(\binom{13}{5}^{-1} - 10\right) C = 1.965402 \times 10^{-3}$$

$$P\{\text{straight}\} = 10(4^5 - 4)C = 3.924647 \times 10^{-3}$$

$$P\{3\} = 13 \binom{12}{2}^{-1} 4^3 C = 2.1128451 \times 10^{-2}$$

$$P\{2\} = \binom{13}{2}^{-1} 44 \binom{4}{2}^{-1} \binom{4}{2}^{-1} C = 4.7539016 \times 10^{-2}$$

$$P\{1\} = 4 \binom{4}{2}^{-1} \binom{12}{3}^{-1} 4^3 C = 0.130021239$$

$$P\{0\} = 9 \binom{4}{2}^{-1} \binom{12}{3}^{-1} 4^3 C = 0.292547788$$

$$\begin{aligned}P\{\text{other}\} = {} & 1 - P\{\mathrm{R}\} - P\{\mathrm{S}\} - P\{\text{full}\} - P\{\text{flush}\} \\ & - P\{\text{straight}\} - \sum_{i=0}^{4} P\{i\} = 0.5010527\end{aligned}$$

Therefore, we see that

$$E[X] = 0.0512903 + \sum_{i=0}^{3} E[X|i]P\{i\} + E[X|\text{other}]\ 0.5010527$$

Now, $E[X|3]$ can be analytically computed by noting that the 2 new cards will come from a subdeck of 47 cards that contains 1 card of one denomination (namely the

denomination to which your three of a kind belong), 3 cards of two denominations, and 4 cards of the other 10 denominations. Thus, letting F be the final hand, we have that

$$P\{F = 4|\text{dealt 3}\} = \frac{46}{\binom{47}{2}} = 0.042553191$$

$$P\{F = \text{full}|\text{dealt 3}\} = \frac{2 \cdot 3 + 10 \cdot 6}{\binom{47}{2}} = 0.061054579$$

$$P\{F = 3|\text{dealt 3}\} = 1 - 0.042553191 - 0.061054579 = 0.89639223$$

Hence,

$$\begin{aligned} E[X|3] &= 25(0.042553191) + 8(0.061054579) + 3(0.89639223) \\ &= 4.241443097 \end{aligned}$$

Similarly, we can analytically derive (and the derivation is left as an exercise) $E[X|i]$ for $i = 0, 1, 2$.

In running the simulation, we should thus generate a hand. If it contains at least one pair or a higher hand then it should be discarded and the process begun again. When we are dealt a hand that does not contain a pair (or any higher hand), we should use whatever strategy we are employing to discard and receive new cards. If X_o is the payoff on this hand, then X_o is the estimator of $E[X|\text{other}]$, and the estimator of $\theta = E[X]$ based on this single run is

$$\begin{aligned} \hat{\theta} = \; & 0.0512903 + 0.021128451(4.241443097) + 0.047539016E[X|2] \\ & + 0.130021239E[X|1] + 0.292547788E[X|0] + 0.5010527X_o \end{aligned}$$

Note that the variance of the estimator is

$$\text{Var}(\hat{\theta}) = (0.5010527)^2\text{Var}(X_o) \qquad \square$$

Remarks

1. We have supposed that the strategy employed always sticks with a pat hand and always keeps whatever pairs it has. However, for the payoffs given this is not an optimal strategy. For instance, if one is dealt 2, 10, jack, queen, king, all of spades, then rather than standing with this flush it is better to discard the 2 and draw another card (why is that?). Also, if dealt 10, jack, queen, king, all of spades, along with the 10 of hearts, it is better to discard the 10 of hearts and draw 1 card than it is to keep the pair of 10s.

2. We could have made further use of stratified sampling by breaking up the "other" category into, say, those "other" hands that contain four cards of the same suit, and those that do not. It is not difficult to analytically compute the probability that a hand will be without a pair and with four cards of the same suit. We could then use simulation to estimate the conditional expected payoffs in these two "other" cases. □

9.5 Applications of Stratified Sampling

In the following subsections, we show how to use ideas of stratified sampling when analyzing systems having Poisson arrivals, monotone functions of many variables, and compound random vectors.

In 9.5.1 we consider a model in which arrivals occur according to a Poisson process, and then we present an efficient way to estimate the expected value of a random variable whose mean depends on the arrival process only through arrivals up to some specified time. In 9.5.2 we show how to use stratified sampling to efficiently estimate the expected value of a nondecreasing function of random numbers. In 9.5.3 we define the concept of a compound random vector and show how to efficiently estimate the expectation of a function of this vector.

9.5.1 Analyzing Systems Having Poisson Arrivals

Consider a system in which arrivals occur according to a Poisson process and suppose we are interested in using simulation to compute $E[D]$, where the value of D depends on the arrival process only through those arrivals before time t. For instance, D might be the sum of the delays of all arrivals by time t in a parallel multiserver queueing system. We suggest the following approach to using simulation to estimate $E[D]$. First, with $N(t)$ equal to the number of arrivals by time t, note that for any specified integral value m

$$E[D] = \sum_{j=0}^{m} E[D|N(t) = j]e^{-\lambda t}(\lambda t)^j/j! + E[D|N(t) > m] \times \left(1 - \sum_{j=0}^{m} e^{-\lambda t}(\lambda t)^j/j!\right) \tag{9.9}$$

Let us suppose that $E[D|N(t) = 0]$ can be easily computed and also that D can be determined by knowing the arrival times along with the service time of each arrival.

Each run of our suggested simulation procedure will generate an independent estimate of $E[D]$. Moreover, each run will consist of $m + 1$ stages, with stage j producing an unbiased estimator of $E[D|N(t) = j]$, for $j = 1, \ldots, m$, and with stage $m+1$ producing an unbiased estimator of $E[D|N(t) > m]$. Each succeeding

stage will make use of data from the previous stage along with any additionally needed data, which in stages $2, \ldots, m$ will be another arrival time and another service time. To keep track of the current arrival times, each stage will have a set S whose elements are arranged in increasing value and which represents the set of arrival times. To go from one stage to the next, we make use of the fact that conditional on there being a total of j arrivals by time t, the set of j arrival times are distributed as j independent uniform $(0, t)$ random variables. Thus, the set of arrival times conditional on $j+1$ events by time t is distributed as the set of arrival times conditional on j events by time t along with a new independent uniform $(0, t)$ random variable.

A run is as follows:

STEP 1: Let $N = 1$. Generate a random number U_1, and let $S = \{tU_1\}$.

STEP 2: Suppose $N(t) = 1$, with the arrival occurring at time tU_1. Generate the service time of this arrival, and compute the resulting value of D. Call this value D_1.

STEP 3: Let $N = N + 1$.

STEP 4: Generate a random number U_N, and add tU_N in its appropriate place to the set S so that the elements in S are in increasing order.

STEP 5: Suppose $N(t) = N$, with S specifying the N arrival times; generate the service time of the arrival at time tU_N and, using the previously generated service times of the other arrivals, compute the resulting value of D. Call this value D_N.

STEP 6: If $N < m$ return to Step 3. If $N = m$, use the inverse transform method to generate the value of $N(t)$ conditional on it exceeding m. If the generated value is $m + k$, generate k additional random numbers, multiply each by t, and add these k numbers to the set S. Generate the service times of these k arrivals and, using the previously generated service times, compute D. Call this value $D_{>m}$.

With $D_0 = E[D|N(t) = 0]$, the estimate from this run is

$$\mathcal{E} = \sum_{j=0}^{m} D_j e^{-\lambda t}(\lambda t)^j/j! + D_{>m}\left(1 - \sum_{j=0}^{m} e^{-\lambda t}(\lambda t)^j/j!\right) \tag{9.10}$$

Because the set of unordered arrival times, given that $N(t) = j$, is distributed as a set of j independent uniform $(0, t)$ random variables, it follows that

$$E[D_j] = E[D|N(t) = j], \qquad E[D_{>m}] = E[D|N(t) > m]$$

thus showing that $\mathcal{E}$ is an unbiased estimator of $E[D]$. Generating multiple runs and taking the average value of the resulting estimates yields the final simulation estimator.

Remarks

1. It should be noted that the variance of our estimator $\sum_{j=0}^{m} D_j e^{-\lambda t}(\lambda t)^j/j! + D_{>m}(1 - \sum_{j=0}^{m} e^{-\lambda t}(\lambda t)^j/j!)$ is, because of the positive correlations introduced by reusing the same data, larger than it would be if the D_j were independent estimators. However, the increased speed of the simulation should more than make up for this increased variance.
2. When computing D_{j+1}, we can make use of quantities used in computing D_j. For instance, suppose $D_{i,j}$ was the delay of arrival i when $N(t) = j$. If the new arrival time tU_{j+1} is the k^{th} smallest of the new set S, then $D_{i,j+1} = D_{i,j}$ for $i < k$.
3. Other variance reduction ideas can be used in conjunction with our approach. For instance, we can improve the estimator by using a linear combination of the service times as a control variable. □

It remains to determine an appropriate value of m. A reasonable approach might be to choose m to make

$$E[D|N(t) > m]P\{N(t) > m\} = E[D|N(t) > m]\left(1 - \sum_{j=0}^{m} e^{-\lambda t}(\lambda t)^j/j!\right)$$

sufficiently small. Because $\text{Var}(N(t)) = \lambda t$, a reasonable choice would be of the form

$$m = \lambda t + k\sqrt{\lambda t}$$

for some positive number k.

To determine the appropriate value of k, we can try to bound $E[D|N(t) > m]$ and then use this bound to determine the appropriate value of k (and m). For instance, suppose D is the sum of the delays of all arrivals by time t in a single server system with mean service time 1. Then because this quantity will be maximized when all arrivals come simultaneously, we see that

$$E[D|N(t)] \leqslant \sum_{i=1}^{N(t)-1} i$$

Because the conditional distribution of $N(t)$ given that it exceeds m will, when m is at least 5 standard deviations greater than $E[N(t)]$, put most of its weight near $m + 1$, we see from the preceding that one can reasonably assume that, for $k \geqslant 5$,

$$E[D|N(t) > m] \leqslant (m + 1)^2/2$$

Using that, for a standard normal random variable Z (see Sec. 4.3 of Ross, S., and E. Pekoz, *A Second Course in Probability*, 2007)

$$P(Z > x) \leqslant (1 - 1/x^2 + 3/x^4)\frac{e^{-x^2}/2}{x\sqrt{2\pi}}, \quad x > 0$$

we see, upon using the normal approximation to the Poisson, that for $k \geqslant 5$ and $m = \lambda t + k\sqrt{\lambda t}$, we can reasonably assume that

$$E[D|N(t) > m]P\{N(t) > m\} \leqslant (m+1)^2 \frac{e^{-k^2/2}}{2k\sqrt{2\pi}}$$

For instance, with $\lambda t = 10^3$ and $k = 6$, the preceding upper bound is about .0008.

We will end this subsection by proving that the estimator $\mathcal{E}$ has a smaller variance than does the raw simulation estimator D.

Theorem

$$Var(\mathcal{E}) \leqslant Var(D)$$

Proof We will prove the result by showing that $\mathcal{E}$ can be expressed as a conditional expectation of D given some random vector. To show this, we will utilize the following approach for simulating D:

STEP 1: Generate the value of N', a random variable whose distribution is the same as that of $N(t)$ conditioned to exceed m. That is,

$$P\{N' = k\} = \frac{(\lambda t)^k/k!}{\sum_{k=m+1}^{\infty}(\lambda t)^k/k!}, \quad k > m$$

STEP 2: Generate the values of $A_1, \ldots, A_{N'}$, independent uniform $(0, t)$ random variables.

STEP 3. Generate the values of $S_1, \ldots, S_{N'}$, independent service time random variables.

STEP 4. Generate the value of $N(t)$, a Poisson random variable with mean λt.

STEP 5. If $N(t) = j \leqslant m$, use the arrival times $A_1, \ldots, A_j$ along with their service times $S_1, \ldots, S_j$ to compute the value of $D = D_j$.

STEP 6. If $N(t) > m$, use the arrival times $A_1, \ldots, A_{N'}$ along with their service times $S_1, \ldots, S_{N'}$ to compute the value of $D = D_{>m}$.

Nothing that,

$$\begin{aligned}
&E[D|N', A_1, \ldots, A_{N'}, S_1, \ldots, S_{N'}] \\
&= \sum_j E[D|N', A_1, \ldots, A_{N'}, S_1, \ldots, S_{N'}, N(t) = j] \\
&\quad \times P\{N(t) = j|N', A_1, \ldots, A_{N'}, S_1, \ldots, S_{N'}\} \\
&= \sum_j E[D|N', A_1, \ldots, A_{N'}, S_1, \ldots, S_{N'}, N(t) = j]P\{N(t) = j\} \\
&= \sum_{j=0}^{m} D_j P\{N(t) = j\} + \sum_{j>m} D_{>m} P\{N(t) = j\} \\
&= \mathcal{E}
\end{aligned}$$

we see that $\mathcal{E}$ is the conditional expectation of D given some data. Consequently, the result follows from the conditional variance formula. □

9.5.2 Computing Multidimensional Integrals of Monotone Functions

Suppose that we want to use simulation to estimate the n dimensional integral

$$\theta = \int_0^1 \int_0^1 \cdots \int_0^1 g(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

With $U_1, \ldots, U_n$ being independent uniform (0,1) random variables, the preceding can be expressed as

$$\theta = E[g(U_1, \ldots, U_n)]$$

Suppose that g is a nondecreasing function of each of its variables. That is, for fixed values $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, the function $g(x_1, \ldots, x_i, \ldots, x_n)$ is increasing in x_i, for each $i = 1, \ldots, n$. If we let $Y = \prod_{i=1}^n U_i$, then because both Y and $g(U_1, \ldots, U_n)$ are increasing functions of the U_i, it would seem that $E[\text{Var}(g(U_1, \ldots, U_n)|Y)]$ might often be relatively small. Thus, we should consider estimating θ by stratifying on $\prod_{i=1}^n U_i$. To accomplish this, we need to first determine

(a) the probability distribution of $\prod_{i=1}^n U_i$;
(b) how to generate the value of $\prod_{i=1}^n U_i$ conditional that it lies in some interval;
(c) how to generate $U_1, \ldots, U_n$ conditional on the value of $\prod_{i=1}^n U_i$.

To accomplish the preceding objectives, we relate the U_i to a Poisson process. Recall that $-\log(U)$ is exponential with rate 1, and interpret $-\log(U_i)$ as the time between the $(i-1)^{th}$ and the i^{th} event of a Poisson process with rate 1. With this interpretation, the j^{th} event of the Poisson process will occur at time T_j, where

$$T_j = \sum_{i=1}^{j} -\log(U_i) = -\log(U_1 \ldots U_j)$$

Because the sum of n independent exponential random variables with rate 1 is a gamma $(n, 1)$ random variable, we can generate the value of $T_n = -\log(U_1 \cdots U_n)$ by generating (in a stratified fashion to be discussed) a gamma $(n, 1)$ random variable. This results in a generated value for $\prod_{i=1}^n U_i$, namely

$$\prod_{i=1}^{n} U_i = e^{-T_n}$$

To generate the individual random variables $U_1, \ldots, U_n$ conditional on the value of their product, we use the Poisson process result that conditional on the n^{th} event

of the Poisson process occurring at time t, the sequence of the first $n-1$ event times is distributed as an ordered sequence of $n-1$ independent uniform $(0, t)$ random variables. Thus, once the value of T_n has been generated, the individual U_i can be obtained by first generating $n-1$ random numbers $V_1, \ldots, V_{n-1}$, and then ordering them to obtain their ordered values $V_{(1)} < V_{(2)} < \ldots < V_{(n-1)}$. As $T_n V_{(j)}$ represents the time of event j, this yields

$$\begin{aligned} T_n V_{(j)} &= -\log(U_1 \ldots U_j) \\ &= -\log(U_1 \ldots U_{j-1}) - \log(U_j) \\ &= T_n V_{(j-1)} - \log(U_j) \end{aligned}$$

Therefore, with $V_{(0)} = 0$, $V_{(n)} = 1$,

$$U_j = e^{-T_n[V_{(j)} - V_{(j-1)}]}, \quad j = 1, \ldots, n \tag{9.11}$$

Thus we see how to generate $U_1, \ldots, U_n$ conditional on the value of $\Pi_{i=1}^n U_i$. To perform the stratification, we now make use of the fact that $T_n = -\log(\Pi_{i=1}^n U_i)$ is a gamma $(n, 1)$ random variable. Let G_n be the gamma $(n, 1)$ distribution function. If we plan to do m simulation runs, then on the k^{th} run a random number U should be generated and T_n should be taken to equal $G_n^{-1}\left(\frac{U+k-1}{m}\right)$. For this value of T_n, we then use the preceding to simulate the values of $U_1, \ldots, U_n$ and calculate $g(U_1, \ldots, U_n)$. [That is, we generate $n-1$ random numbers, order them to obtain $V_{(1)} < V_{(2)} < \ldots < V_{(n-1)}$, and let the U_j be given by (9.11)]. The average of the values of g obtained in the m runs is the stratified sampling estimator of $E[g(U_1, \ldots, U_n)]$.

Remarks

1. A gamma random variable with parameters n, 1 has the same distribution as does $\frac{1}{2}\chi^2_{2n}$, where χ^2_{2n}, is a chi-squared random variable with $2n$ degrees of freedom. Consequently,

$$G_n^{-1}(x) = \frac{1}{2} F^{-1}_{\chi^2_{2n}}(x)$$

   where $F^{-1}_{\chi^2_{2n}}(x)$ is the inverse of the distribution function of a chi-squared random variable with $2n$ degrees of freedom. Approximations for the inverse of the chi-squared distribution are readily available in the literature.
2. With a slight modification, we can apply the preceding stratification idea even when the underlying function is monotone increasing in some of its coordinates and monotone decreasing in the others. For instance, suppose we want to evaluate $E[h(U_1, \ldots, U_n)]$, where h is monotone decreasing in its first coordinate and monotone increasing in the others. Using that $1 - U_1$ is also uniform on (0,1), we can write

$$E[h(U_1, U_2, \ldots, U_n)] = E[h(1-U_1, U_2, \ldots, U_n)] = E[g(U_1, U_2, \ldots, U_n)]$$

where $g(x_1, x_2, \ldots, x_n) \equiv h(1 - x_1, x_2, \ldots, x_n)$ is monotone increasing in each of its coordinates. □

9.5.3 Compound Random Vectors

Let N be a nonnegative integer valued random variable with probability mass function

$$p(n) = P\{N = n\}$$

and suppose that N is independent of the sequence of independent and identically distributed random variables $X_1, X_2, \ldots$, having the common distribution function F. Then the random vector $(X_1, \ldots, X_N)$ is called a *compound random vector*. (When $N = 0$, call the compound random vector the null vector.)

For a family of functions $g_n(x_1, \ldots, x_n)$, $n \geqslant 0$, with $g_0 \equiv 0$, suppose we are interested in using simulation to estimate $E[g_N(X_1, \ldots, X_N)]$, for a specified compound random vector $(X_1, \ldots, X_N)$. Some functional families of interest are as follows.

- If

$$g_n(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } \sum_{i=1}^n x_i > a \\ 0, & \text{if otherwise} \end{cases}$$

  then $E[g_N(X_1, \ldots, X_N)]$ is the probability that a compound random variable exceeds a.
- A generalization of the preceding example is, for $0 < \alpha < 1$, to take

$$g_n(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } \sum_{i=1}^n \alpha^i x_i > a \\ 0, & \text{if otherwise} \end{cases}$$

  Now $E[g_N(X_1, \ldots, X_N)]$ is the probability that the discounted sum of a compound random vector exceeds a.
- Both of the previous examples are special cases of the situation where, for a specified sequence a_i, $i \geqslant 1$,

$$g_n(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } \sum_{i=1}^n a_i x_i > a \\ 0, & \text{if otherwise} \end{cases}$$

- One is sometimes interested in a function of a weighted sum of the k largest values of the random vector, leading to the consideration of the functions

$$g_n(x_1, \ldots, x_n) = g\left(\sum_{i=1}^{\min(k,n)} a_i x_{(i:n)}\right)$$

  where $x_{(i:n)}$ is the i^{th} largest of the values $x_1, \ldots, x_n$, and where g is a specified function with $g(0) = 0$.

To use simulation to estimate $\theta \equiv E[g_N(X_1, \ldots, X_N)]$, choose a value of m such that $P\{N > m\}$ is small, and suppose that we are able to simulate N conditional on it exceeding m. With $p_n = P\{N = n\}$, conditioning on the mutually exclusive and exhaustive possibilities that N is 0 or 1, . . ., or m, or that it exceeds m, yields

$$\begin{aligned}\theta &= \sum_{n=0}^{m} E[g_N(X_1, \ldots, X_N)|N = n]p_n + E[g_N(X_1, \ldots, X_N)|N > m]P\{N > m\} \\ &= \sum_{n=0}^{m} E[g_n(X_1, \ldots, X_n)|N = n]p_n + E[g_N(X_1, \ldots, X_N)|N > m]P\{N > m\} \\ &= \sum_{n=0}^{m} E[g_n(X_1, \ldots, X_n)]p_n + E[g_N(X_1, \ldots, X_N)|N > m]\left[1 - \sum_{n=0}^{m} p_n\right]\end{aligned}$$

where the final equality made use of the independence of N and $X_1, \ldots, X_n$.

To effect a simulation run to estimate $E[g_N(X_1, \ldots, X_N)]$, first generate the value of N conditional on it exceeding m. Suppose the generated value is m'. Then generate m' independent random variables $X_1, \ldots, X_{m'}$ having distribution function F. That completes a simulation run, with the estimator from that run being

$$\mathcal{E} = \sum_{n=1}^{m} g_n(X_1, \ldots, X_n)p_n + g_{m'}(X_1, \ldots, X_{m'})\left[1 - \sum_{n=0}^{m} p_n\right]$$

Remarks

1. If it is relatively easy to compute the values of the functions g_n, we recommend that one also use the data $X_1, \ldots, X_{m'}$ in the reverse order to obtain a second estimator, and then average the two estimators. That is, use the run estimator

$$\mathcal{E}^* = \frac{1}{2}\left(\mathcal{E} + \sum_{n=1}^{m} g_n(X_{m'}, \ldots, X_{m'-n+1})p_n + g_{m'}(X_{m'}, \ldots, X_1)\left[1 - \sum_{n=0}^{m} p_n\right]\right)$$

2. If if is difficult to generate N conditional on its value exceeding m, it is often worthwhile to try to bound $E[g_N(X_1, \ldots, X_N)|N > m]P\{N > m\}$ and then determine an appropriately large value of m that makes the bound negligibly small. (For instance, if the functions g_n are indicator—that is, 0 or 1—functions then $E[g_N(X_1, \ldots, X_N)|N > m]P\{N > m\} \leqslant P\{N > m\}$.) The result from a simulation that ignores the term $E[g_N(X_1, \ldots, X_N)|N > m]P\{N > m\}$ will often be sufficiently accurate.
3. If $E[N|N > m]$ can be computed then it can be used as a control variable. □

9.5.4 The Use of Post-Stratification

Post-stratification is a powerful but underused variance reduction technique. For instance, suppose we want to estimate $E[X]$ and are thinking of using Y as a control variable. However, if the probability distribution of Y rather than just its mean is known, then it it is better to post-stratify on Y. Moreover, if one is planning to stratify on Y by using proportional sampling - that is, doing $nP(Y = i)$ of the total of n runs conditional on $Y = i$ - as opposed to trying to estimate the optimal number of runs in each strata, then generating the data unconditionally and then post-stratifying is usually as good as stratifying.

As examples of the preceding, suppose we want to estimate $\theta = E[h(X_1, \ldots, X_k)]$, where h is a monotone increasing function of $\mathbf{X} = (X_1, \ldots, X_k)$. If the distribution of $\sum_{i=1}^{k} X_i$ is known, then one can effectively post-stratify on this sum. Consider the following instances where this would be the case.

1. Suppose that $(X_1, \ldots, X_k)$ has a multivariate normal distribution. Hence, $S = \sum_{i=1}^{k} X_i$ will be normal, say with mean μ and variance σ^2. Breaking the possible values of S into m groupings, say by choosing $-\infty = a_1 < a_2 < \cdots < a_m < a_{m+1} = \infty$ and letting $J = i$ if $a_i < S < a_{i+1}$, then

$$\theta = \sum_{i=1}^{m} E[h(\mathbf{X})|J = i]P(J = i)$$

   If n_i of the n runs result in $J = i$ then, with $\bar{h}_i$ equal to the average of the values of $h(\mathbf{X})$ over those n_i runs, the post-stratification estimate of θ, call it $\hat{\theta}$, is

$$\hat{\theta} = \sum_{i=1}^{m} \bar{h}_i P(J = i)$$

   where $P(J = i) = \Phi(\frac{a_{i+1}-\mu}{\sigma}) - \Phi(\frac{a_i-\mu}{\sigma})$, with Φ being the standard normal distribution function.
2. If $X_1, \ldots, X_k$ are independent Poisson random variables with means $\lambda_1, \ldots, \lambda_k$, then $S = \sum_{i=1}^{k} X_i$ will be Poisson with mean $\lambda = \sum_{i=1}^{k} \lambda_i$. Hence, we can choose m and write

$$\theta = \sum_{i=0}^{m} E[h(\mathbf{X})|S = i]e^{-\lambda}\lambda^i/i! + E[h(\mathbf{X})|S > m]P(S > m).$$

   The unconditionally generated data can then be used to estimate the quantities $E[h(\mathbf{X})|S = i]$and $E[h(\mathbf{X})|S > m]$.
3. Suppose $X_1, \ldots, X_k$ are independent Bernoulli random variables with parameters $p_1, \ldots, p_k$. The distribution of $S = \sum_{i=1}^{k} X_i$ can be computed

using the following recursive idea. For $1 \leqslant r \leqslant k$, let

$$P_r(j) = P(S_r = j)$$

where $S_r = \sum_{i=1}^{r} X_i$. Now,with $q_i = 1 - p_i$, we have

$$P_r(r) = \prod_{i=1}^{r} p_i, \quad P_r(0) = \prod_{i=1}^{r} q_i$$

For $0 < j < r$, conditioning on X_r yields the recursion:

$$\begin{aligned} P_r(j) &= P(S_r = j|X_r = 1)p_r + P(S_r = j|X_r = 0)q_r \\ &= P_{r-1}(j-1)p_r + P_{r-1}(j)q_r \end{aligned}$$

Starting with $P_1(1) = p_1,\ P_1(0) = q_1$, these equations can be recursively solved to obtain the function $P_k(j)$. After this initial calculation we can do an unconditional simulation and then estimate θ by

$$\hat{\theta} = \sum_{j=0}^{k} \bar{h}_j P_k(j)$$

where $\bar{h}_j$is the average of the values of h over all simulation runs that result in $\sum_{i=1}^{k} X_i = j$.

9.6 Importance Sampling

Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote a vector of random variables having a joint density function (or joint mass function in the discrete case) $f(\mathbf{x}) = f(x_1, \ldots, x_n)$, and suppose that we are interested in estimating

$$\theta = E[h(\mathbf{X})] = \int h(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}$$

where the preceding is an n-dimensional integral over all possible values of $\mathbf{x}$. (If the X_i are discrete, then interpret the integral as an n-fold summation.)

Suppose that a direct simulation of the random vector $\mathbf{X}$, so as to compute values of $h(\mathbf{X})$, is inefficient, possibly because (a) it is difficult to simulate a random vector having density function $f(\mathbf{x})$, or (b) the variance of $h(\mathbf{X})$ is large, or (c) a combination of (a) and (b).

Another way in which we can use simulation to estimate θ is to note that if $g(\mathbf{x})$ is another probability density such that $f(\mathbf{x}) = 0$ whenever $g(\mathbf{x}) = 0$, then we can express θ as

$$\begin{aligned} \theta &= \int \frac{h(\mathbf{x}) f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\, d\mathbf{x} \\ &= E_g\left[\frac{h(\mathbf{X}) f(\mathbf{X})}{g(\mathbf{X})}\right] \end{aligned} \tag{9.12}$$

where we have written E_g to emphasize that the random vector $\mathbf{X}$ has joint density $g(\mathbf{x})$.

It follows from Equation (9.12) that θ can be estimated by successively generating values of a random vector $\mathbf{X}$ having density function $g(\mathbf{x})$ and then using as the estimator the average of the values of $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$. If a density function $g(\mathbf{x})$ can be chosen so that the random variable $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ has a small variance, then this approach—referred to as *importance sampling*—can result in an efficient estimator of θ.

Let us now try to obtain a feel for why importance sampling can be useful. To begin, note that $f(\mathbf{X})$ and $g(\mathbf{X})$ represent the respective likelihoods of obtaining the vector $\mathbf{X}$ when $\mathbf{X}$ is a random vector with respective densities f and g. Hence, if $\mathbf{X}$ is distributed according to g, then it will usually be the case that $f(\mathbf{X})$ will be small in relation to $g(\mathbf{X})$, and thus when $\mathbf{X}$ is simulated according to g the likelihood ratio $f(\mathbf{X})/g(\mathbf{X})$ will usually be small in comparison to 1. However, it is easy to check that its mean is 1:

$$E_g\left[\frac{f(\mathbf{X})}{g(\mathbf{X})}\right] = \int \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\, d\mathbf{x} = \int f(\mathbf{x})\, d\mathbf{x} = 1$$

Thus we see that even though $f(\mathbf{X})/g(\mathbf{X})$ is usually smaller than 1, its mean is equal to 1, thus implying that it is occasionally large and so will tend to have a large variance. So how can $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ have a small variance? The answer is that we can sometimes arrange to choose a density g such that those values of $\mathbf{x}$ for which $f(\mathbf{x})/g(\mathbf{x})$ is large are precisely the values for which $h(\mathbf{x})$ is exceedingly small, and thus the ratio $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ is always small. Since this will require that $h(\mathbf{x})$ is sometimes small, importance sampling seems to work best when estimating a small probability, for in this case the function $h(\mathbf{x})$ is equal to 1 when $\mathbf{x}$ lies in some set and is equal to 0 otherwise.

We will now consider how to select an appropriate density g. We will find that the so-called tilted densities are useful. Let $M(t) = E_f[e^{tX}] = \int e^{tx} f(x)\, dx$ be the moment generating function corresponding to a one-dimensional density f.

Definition *A density function*

$$f_t(x) = \frac{e^{tx} f(x)}{M(t)}$$

is called a tilted density of f, $-\infty < t < \infty$.

A random variable with density f_t tends to be larger than one with density f when $t > 0$ and tends to be smaller when $t < 0$.

In certain cases the tilted densities f_t have the same parametric form as does f.

Example 9t If f is the exponential density with rate λ, then

$$f_t(x) = Ce^{tx}\lambda e^{-\lambda x} = Ce^{-(\lambda - t)x}$$

where $C = 1/M(t)$ does not depend on x. Therefore, for $t < \lambda$, f_t is an exponential density with rate $\lambda - t$.

If f is a Bernoulli probability mass function with parameter p, then

$$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Hence, $M(t) = E_f[e^{tX}] = pe^t + 1 - p$, and so

$$\begin{aligned} f_t(x) &= \frac{1}{M(t)}(pe^t)^x(1-p)^{1-x} \\ &= \left(\frac{pe^t}{pe^t + 1 - p}\right)^x \left(\frac{1-p}{pe^t + 1 - p}\right)^{1-x} \end{aligned}$$

That is, f_t is the probability mass function of a Bernoulli random variable with parameter $p_t = (pe^t)/(pe^t + 1 - p)$.

We leave it as an exercise to show that if f is a normal density with parameters μ and σ^2 then f_t is a normal density having mean $\mu + \sigma^2 t$ and variance σ^2. □

In certain situations the quantity of interest is the sum of the independent random variables $X_1, \ldots, X_n$. In this case the joint density f is the product of one-dimensional densities. That is,

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$$

where f_i is the density function of X_i. In this situation it is often useful to generate the X_i according to their tilted densities, with a common choice of t employed.

Example 9u Let $X_1, \ldots, X_n$ be independent random variables having respective probability density (or mass) functions f_i, for $i = 1, \ldots, n$. Suppose we are interested in approximating the probability that their sum is at least as large as a, where a is much larger than the mean of the sum. That is, we are interested in

$$\theta = P\{S \geqslant a\}$$

where $S = \sum_{i=1}^n X_i$, and where $a > \sum_{i=1}^n E[X_i]$. Letting $I\{S \geqslant a\}$ equal 1 if $S \geqslant a$ and letting it be 0 otherwise, we have that

$$\theta = E_{\mathbf{f}}[I\{S \geqslant a\}]$$

where $\mathbf{f} = (f_1, \ldots, f_n)$. Suppose now that we simulate X_i according to the tilted mass function $f_{i,t}, i = 1, \ldots, n$, with the value of $t, t > 0$, left to be determined. The importance sampling estimator of θ would then be

$$\hat{\theta} = I\{S \geqslant a\} \prod \frac{f_i(X_i)}{f_{i,t}(X_i)}$$

Now,

$$\frac{f_i(X_i)}{f_{i,t}(X_i)} = M_i(t)e^{-tX_i}$$

and so,

$$\hat{\theta} = I\{S \geqslant a\}M(t)e^{-tS}$$

where $M(t) = \Pi M_i(t)$ is the moment generating function of S. Since $t > 0$ and $I\{S \geqslant a\}$ is equal to 0 when $S < a$, it follows that

$$I\{S \geqslant a\}e^{-tS} \leqslant e^{-ta}$$

and so

$$\hat{\theta} \leqslant M(t)e^{-ta}$$

To make the bound on the estimator as small as possible we thus choose t, $t > 0$, to minimize $M(t)e^{-ta}$. In doing so, we will obtain an estimator whose value on each iteration is between 0 and $\min_t M(t)e^{-ta}$. It can be shown that the minimizing t—call it t^*—is such that

$$E_{t^*}[S] = E_{t^*}\left[\sum_{i=1}^{n} X_i\right] = a$$

where, in the preceding, we mean that the expected value is to be taken under the assumption that the distribution of X_i is f_{i,t^*} for $i = 1, \ldots, n$.

For instance, suppose that $X_1, \ldots, X_n$ are independent Bernoulli random variables having respective parameters p_i, for $i = 1, \ldots, n$. Then, if we generate the X_i according to their tilted mass functions $p_{i,t}$, $i = 1, \ldots, n$, the importance sampling estimator of $\theta = P\{S \geqslant \alpha\}$ is

$$\hat{\theta} = I\{S \geqslant a\}e^{-tS}\prod_{i=1}^{n}(p_i e^t + 1 - p_i)$$

Since $p_{i,t}$ is the mass function of a Bernoulli random variable with parameter $(p_i e^t)/(p_i e^t + 1 - p_i)$, it follows that

$$E_t\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \frac{p_i e^t}{p_i e^t + 1 - p_i}$$

The value of t that makes the preceding equal to a can be numerically approximated and the t utilized in the simulation.

As an illustration, suppose that $n = 20$, $p_i = 0.4$, $a = 16$. Then

$$E_t[S] = 20\frac{0.4e^t}{0.4e^t + 0.6}$$

Setting this equal to 16 yields after a little algebra that

$$e^{t^*} = 6$$

Thus, if we generate the Bernoullis using the parameter $(0.4e^{t^*})/(0.4e^{t^*} + 0.6) = 0.8$, then as

$$M(t^*) = (0.4e^{t^*} + 0.6)^{20} \quad \text{and} \quad e^{-t^*S} = (1/6)^S$$

we see that the importance sampling estimator is

$$\hat{\theta} = I\{S \geqslant 16\}(1/6)^S 3^{20}$$

It follows from the preceding that

$$\hat{\theta} \leqslant (1/6)^{16} 3^{20} = 81/2^{16} = 0.001236$$

That is, on each iteration the value of the estimator is between 0 and 0.001236. Since, in this case, θ is the probability that a binomial random variable with parameters 20, 0.4 is at least 16, it can be explicitly computed with the result θ=0.000317. Hence, the raw simulation estimator I, which on each iteration takes the value 0 if the sum of the Bernoullis with parameter 0.4 is less than 16 and takes the value 1 otherwise, will have variance

$$\text{Var}(I) = \theta(1 - \theta) = 3.169 \times 10^{-4}$$

On the other hand, it follows from the fact that $0 \leqslant \theta \leqslant 0.001236$ that (see Exercise 29) □

$$\text{Var}(\hat{\theta}) \leqslant 2.9131 \times 10^{-7}$$

Example 9v Consider a single server queue in which the times between successive customer arrivals have density function f and the service times have density g. Let D_n denote the amount of time that the nth arrival spends waiting in queue and suppose we are interested in estimating $\alpha = P\{D_n \geqslant a\}$ when a is much larger than $E[D_n]$. Rather than generating the successive interarrival and service times according to f and g, respectively, we should generate them according to the densities f_{-t} and g_t, where t is a positive number to be determined. Note that using these distributions as opposed to f and g will result in smaller interarrival times (since $-t < 0$) and larger service times. Hence, there will be a greater chance that $D_n > a$ than if we had simulated using the densities f and g. The importance sampling estimator of α would then be

$$\hat{\alpha} = I\{D_n > a\}e^{t(S_n - Y_n)}[M_f(-t)M_g(t)]^n$$

where S_n is the sum of the first n interarrival times, Y_n is the sum of the first n service times, and M_f and M_g are the moment generating functions of the densities f and g, respectively. The value of t used should be determined by experimenting with a variety of different choices. □

Example 9w Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed normal random variables having mean μ and variance 1, where $\mu < 0$. An important problem in the theory of quality control (specifically in the analysis of cumulative sum charts) is to determine the probability that the partial sums of these values exceed B before going below $-$A. That is, let

$$S_n = \sum_{i=1}^{n} X_i$$

and define

$$N = \text{Min}\{n : \text{either } S_n < -A, \text{ or } S_n > B\}$$

where A and B are fixed positive numbers. We are now interested in estimating

$$\theta = P\{S_N > B\}$$

An effective way of estimating θ is by simulating the X_i as if they were normal with mean $-\mu$ and variance 1, stopping again when their sum either exceeds B or falls below $-A$. (Since $-\mu$ is positive, the stopped sum is greater than B more often than if we were simulating with the original negative mean.) If $X_1, \ldots, X_N$ denote the simulated variables (each being normal with mean $-\mu$ and variance 1) and

$$I = \begin{cases} 1 & \text{if} \sum_{i=1}^{N} X_i > B \\ 0 & \text{otherwise} \end{cases}$$

then the estimate of θ from this run is

$$I \prod_{i=1}^{N} \left[\frac{f_\mu(X_i)}{f_{-\mu}(X_i)} \right] \tag{9.13}$$

where f_c is the normal density with mean c and variance 1. Since

$$\frac{f_\mu(x)}{f_{-\mu}(x)} = \frac{\exp\left\{-\frac{(x-\mu)^2}{2}\right\}}{\exp\left\{-\frac{(x+\mu)^2}{2}\right\}} = e^{2\mu x}$$

it follows from (9.13) that the estimator of θ based on this run is

$$I \exp\left\{2\mu \sum_{i=1}^{N} X_i\right\} = I \exp\{2\mu S_N\}$$

When I is equal to 1, S_N exceeds B and, since $\mu < 0$, the estimator in this case is less than $e^{2\mu B}$. That is, rather than obtaining from each run either the value 0 or 1—as would occur if we did a straight simulation—we obtain in this case either the value

0 or a value that is less than $e^{2\mu B}$, which strongly indicates why this importance sampling approach results in a reduced variance. For example, if $\mu = -0.1$ and $B = 5$, then the estimate from each run lies between 0 and $e^{-1} = 0.3679$. In addition, the above is theoretically important because it shows that

$$P\{\text{cross } B \text{ before} - A\} \leqslant e^{2\mu B}$$

Since the above is true for all positive A, we obtain the interesting result

$$P\{\text{ever cross } B\} \leqslant e^{2\mu B} \qquad \square$$

Example 9x Let $\mathbf{X} = (X_1, \ldots, X_{100})$ be a random permutation of $(1, 2, \ldots, 100)$. That is, $\mathbf{X}$ is equally likely to be any of the (100)! permutations. Suppose we are interested in using simulation to estimate

$$\theta = P\left\{\sum_{j=1}^{100} jX_j > 290,000\right\}$$

To obtain a feel for the magnitude of θ, we can start by computing the mean and standard deviation of $\sum_{j=1}^{100} jX_j$. Indeed, it is not difficult to show that

$$E\left[\sum_{j=1}^{100} jX_j\right] = 100(101)^2/4 = 255,025$$

$$\text{SD}\left(\sum_{j=1}^{100} jX_j\right) = \sqrt{(99)(100)^2(101)^2/144} = 8374.478$$

Hence, if we suppose that $\sum_{j=1}^{100} jX_j$ is roughly normally distributed then, with Z representing a standard normal random variable, we have that

$$\begin{aligned}\theta &\approx P\left\{Z > \frac{290,000 - 255,025}{8374.478}\right\} \\ &= P\{Z > 4.1764\} \\ &= 0.00001481\end{aligned}$$

Thus, θ is clearly a small probability and so an importance sampling estimator is worth considering.

To utilize importance sampling we would want to generate the permutation $\mathbf{X}$ so that there is a much larger probability that $\sum_{j=1}^{100} jX_j > 290,000$. Indeed, we should try for a probability of about 0.5. Now, $\sum_{j=1}^{100} jX_j$ will attain its largest value when $X_j = j$, $j = 1, \ldots, 100$, and indeed it will tend to be large when X_j tends to be large when j is large and small when j is small. One way to generate a permutation $\mathbf{X}$ that will tend to be of this type is as follows: Generate independent exponential

random variables Y_j, $j = 1, \ldots, 100$, with respective rates λ_j, $j = 1, \ldots, 100$ where λ_j, $j = 1, \ldots, 100$, is an increasing sequence whose values will soon be specified. Now, for $j = 1, \ldots, 100$, let X_j be the index of the jth largest of these generated values. That is,

$$Y_{X_1} > Y_{X_2} > \cdots > Y_{X_{100}}$$

Since, for j large, Y_j will tend to be one of the smaller Y's, it follows that X_j will tend to be large when j is large and so $\sum_{j=1}^{100} jX_j$ will tend to be larger than if X were a uniformly distributed permutation.

Let us now compute $E[\sum_{j=1}^{100} jX_j]$. To do so, let $R(j)$ denote the rank of Y_j, $j = 1, \ldots, 100$, where rank 1 signifies the largest, rank 2 the second largest, and so on until rank 100, which is the smallest. Note that since X_j is the index of the jth largest of the Y's, it follows that $R(X_j) = j$. Hence,

$$\sum_{j=1}^{100} jX_j = \sum_{j=1}^{100} R(X_j)X_j = \sum_{j=1}^{100} jR(j)$$

where the final equality follows since $X_1, \ldots, X_{100}$ is a permutation of $1, \ldots, 100$. Therefore, we see that

$$E\left[\sum_{j=1}^{100} jX_j\right] = \sum_{j=1}^{100} jE[R(j)]$$

To compute $E[R_j]$, let $I(i, j) = 1$ if $Y_j < Y_i$ and let it be 0 otherwise, and note that

$$R_j = 1 + \sum_{i:i\neq j} I(i, j)$$

In words, the preceding equation states that the rank of Y_j is 1 plus the number of the Y_i that are larger than it. Hence, taking expectations and using the fact that

$$P\{Y_j < Y_i\} = \frac{\lambda_j}{\lambda_i + \lambda_j},$$

we obtain that

$$E[R_j] = 1 + \sum_{i:i\neq j} \frac{\lambda_j}{\lambda_i + \lambda_j}$$

and thus

$$E\left[\sum_{j=1}^{100} jX_j\right] = \sum_{j=1}^{100} j\left(1 + \sum_{i:i\neq j} \frac{\lambda_j}{\lambda_i + \lambda_j}\right)$$

If we let $\lambda_j = j^{0.7}, j = 1, \ldots, 100$, then a computation shows that $E[\sum_{j=1}^{100} jX_j] = 290,293.6$, and so when $\mathbf{X}$ is generated using these rates it would seem that

$$P\left\{\sum_{j=1}^{100} jX_j > 290{,}000\right\} \approx 0.5$$

Thus, we suggest that the simulation estimator should be obtained by first generating independent exponentials Y_j with respective rates $j^{0.7}$, and then letting X_j be the index of the jth largest, $j = 1, \ldots, 100$. Let $I = 1$ if $\sum_{j=1}^{100} jX_j > 290,000$ and let it be 0 otherwise. Now, the outcome will be $\mathbf{X}$ when $Y_{X_{100}}$ is the smallest Y, $Y_{X_{99}}$ is the second smallest, and so on. The probability of this outcome is 1/(100)! when $\mathbf{X}$ is equally likely to be any of the permutations, whereas its probability when the simulation is as performed is

$$\frac{(X_{100})^{0.7}}{\sum_{j=1}^{100}(X_j)^{0.7}} \frac{(X_{99})^{0.7}}{\sum_{j=1}^{99}(X_j)^{0.7}} \cdots \frac{(X_2)^{0.7}}{\sum_{j=1}^{2}(X_j)^{0.7}} \frac{(X_1)^{0.7}}{(X_1)^{0.7}}$$

Therefore, the importance sampling estimator from a single run is

$$\hat{\theta} = \frac{I}{(100)!} \frac{\prod_{j=1}^{100}\left(\sum_{n=1}^{n}(X_j)^{0.7}\right)}{\left(\prod_{n=1}^{100} n\right)^{0.7}} = \frac{I \prod_{n=1}^{100}\left(\sum_{j=1}^{n}(X_j)^{0.7}\right)}{\left(\prod_{n=1}^{100} n\right)^{1.7}}$$

Before the simulation is begun, the values of $C = 1.7\sum_{n=1}^{100}\log(n)$ and $a(j) = -j^{-0.7}, j = 1, \ldots, 100$ should be computed. A simulation run can then be obtained as follows:

For $j = 1$ to 100
Generate a random number U
$Y_j = a(j)\log U$
Next
Let $X_j, j = 1, \ldots, 100$, be such that Y_{X_j} is the jth largest Y
If $\sum_{j=1}^{n} jX_j \leqslant 290{,}000$ set $\hat{\theta} = 0$ and stop
$S = 0, P = 0$
For $n = 1$ to 100
$S = S + (X_n)^{0.7}$
$P = P + \log(S)$
Next
$\hat{\theta} = e^{P-C}$

A sample of 50,000 simulation runs yielded the estimate $\hat{\theta} = 3.77 \times 10^{-6}$, with a sample variance 1.89×10^{-8}. Since the variance of the raw simulation estimator, which is equal to 1 if $\sum_{j=1}^{100} jX_j > 290,000$ and is equal to 0 otherwise,

is $\text{Var}(I) = \theta(1-\theta) \approx 3.77 \times 10^{-6}$, we see that

$$\frac{\text{Var}(I)}{\text{Var}(\hat{\theta})} \approx 199.47$$

□

Importance sampling is also quite useful in estimating a conditional expectation when one is conditioning on a rare event. That is, suppose $\mathbf{X}$ is a random vector with density function f and that we are interested in estimating

$$\theta = E[h(\mathbf{X})|\mathbf{X} \in \mathcal{A}]$$

where $h(\mathbf{x})$ is an arbitrary real valued function and where $P\{\mathbf{X} \in \mathcal{A}\}$ is a small unknown probability. Since the conditional density of $\mathbf{X}$ given that it lies in $\mathcal{A}$ is

$$f(\mathbf{x}|\mathbf{X} \in \mathcal{A}) = \frac{f(\mathbf{x})}{P\{\mathbf{X} \in \mathcal{A}\}}, \quad \mathbf{x} \in \mathcal{A}$$

we have that

$$\begin{aligned}\theta &= \frac{\int_{\mathbf{x}\in\mathcal{A}} h(\mathbf{x}) f(\mathbf{x}) d(\mathbf{x})}{P\{\mathbf{X} \in \mathcal{A}\}} \\ &= \frac{E[h(\mathbf{X}) I(\mathbf{X} \in \mathcal{A})]}{E[I(\mathbf{X} \in \mathcal{A})]} \\ &= \frac{E[N]}{E[D]}\end{aligned}$$

where $E[N]$ and $E[D]$ are defined to equal the numerator and denominator in the preceding, and $I(\mathbf{X} \in \mathcal{A})$ is defined to be 1 if $\mathbf{X} \in \mathcal{A}$ and 0 otherwise. Hence, rather than simulating $\mathbf{X}$ according to the density f, which would make it very unlikely to be in $\mathcal{A}$, we can simulate it according to some other density g which makes this event more likely. If we simulate k random vectors $\mathbf{X}^1, \ldots, \mathbf{X}^k$ according to g, then we can estimate $E[N]$ by $\frac{1}{k}\sum_{i=1}^{k} N_i$ and $E[D]$ by $\frac{1}{k}\sum_{i=1}^{k} D_i$, where

$$N_i = \frac{h(\mathbf{X}^i) I(\mathbf{X}^i \in \mathcal{A}) f(\mathbf{X}^i)}{g(\mathbf{X}^i)}$$

and

$$D_i = \frac{I(\mathbf{X}^i \in \mathcal{A}) f(\mathbf{X}^i)}{g(\mathbf{X}^i)}$$

Thus, we obtain the following estimator of θ:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} h(\mathbf{X}^i) I(\mathbf{X}^i \in \mathcal{A}) f(\mathbf{X}^i)/g(\mathbf{X}^i)}{\sum_{i=1}^{k} I(\mathbf{X}^i \in \mathcal{A}) f(\mathbf{X}^i)/g(\mathbf{X}^i)} \tag{9.14}$$

The mean square error of this estimator can then be estimated by the bootstrap approach (see, for instance, Example 7e).

Example 9y Let X_i be independent exponential random variables with respective rates $1/(i+2), i = 1, 2, 3, 4$. Let $S = \sum_{i=1}^{4} X_i$, and suppose that we want to estimate $\theta = E[S|S > 62]$. To accomplish this, we can use importance sampling with the tilted distributions. That is, we can choose a value t and then simulate the X_i with rates $1/(i+2) - t$. If we choose $t = 0.14$, then $E_t[S] = 68.43$. So, let us generate k sets of exponential random variables X_i with rates $1/(i+2) - 0.14, i = 1, 2, 3, 4$, and let S_j be the sum of the jth set, $j = 1, \ldots, k$. Then we can estimate

$$E[SI(S > 62)]\text{by}\frac{C}{k}\sum_{j=1}^{k} S_j I(S_j > 62)e^{-0.14S_j}$$

$$E[I(S > 62)]\text{by}\frac{C}{k}\sum_{j=1}^{k} I(S_j > 62)e^{-0.14S_j}$$

where $C = \prod_{i=1}^{4} \frac{1}{1-0.14(i+2)} = 81.635$. The estimator of θ is

$$\hat{\theta} = \frac{\sum_{j=1}^{k} S_j I(S_j > 62)e^{-0.14S_j}}{\sum_{j=1}^{k} I(S_j > 62)e^{-0.14S_j}}$$ □

The importance sampling approach is also useful in that it enables us to estimate two (or more) distinct quantities in a single simulation. For example, suppose that

$$\theta_1 = E[h(\mathbf{Y})] \quad \text{and} \quad \theta_2 = E[h(\mathbf{W})]$$

where $\mathbf{Y}$ and $\mathbf{W}$ are random vectors having joint density functions f and g, respectively. If we now simulate $\mathbf{W}$, we can simultaneously use $h(\mathbf{W})$ and $h(\mathbf{W})f(\mathbf{W})/g(\mathbf{W})$ as estimators of θ_2 and θ_1, respectively. For example, suppose we simulate T, the total time in the system of the first r customers in a queueing system in which the service distribution is exponential with mean 2. If we now decide that we really should have considered the same system but with a service distribution that is gamma distributed with parameters (2, 1), then it is not necessary to repeat the simulation; we can just use the estimator

$$T\frac{\prod_{i=1}^{r} S_i \exp\{-S_i\}}{\prod_{i=1}^{r}\left(\frac{1}{2}\exp\{-S_i/2\}\right)} = 2^r T \exp\left\{-\sum_{i=1}^{r}\frac{S_i}{2}\right\}\prod_{i=1}^{r} S_i$$

where S_i is the (exponentially) generated service time of customer i. [The above follows since the exponential service time density is $g(s) = \frac{1}{2}e^{-s/2}$, whereas the gamma (2, 1) density is $f(s) = se^{-s}$.]

Importance sampling can also be used to estimate tail probabilities of a random variable X whose density f is known, but whose distribution function is difficult to

evaluate. Suppose we wanted to estimate $P_f\{X > a\}$ where the subscript f is used to indicate that X has density function f, and where a is a specified value. Letting

$$I(X > a) = \begin{cases} 1, & \text{if } X > a \\ 0, & \text{if } X \leqslant a \end{cases}$$

we have the following.

$$\begin{aligned} P_f\{X > a\} &= E_f[I(X > a)] \\ &= E_g\left[I(X > a)\frac{f(X)}{g(X)}\right] \quad \text{the importance sampling identity} \\ &= E_g\left[I(X > a)\,\frac{f(X)}{g(X)}\,\middle|\, X > a\right] P_g\{X > a\} \\ &\quad + E_g\left[I(X > a)\,\frac{f(X)}{g(X)}\,\middle|\, X \leqslant a\right] P_g\{X \leqslant a\} \\ &= E_g\left[\frac{f(X)}{g(X)}\,\middle|\, X > a\right] P_g\{X > a\} \end{aligned}$$

If we let g be the exponential density

$$g(x) = \lambda e^{-\lambda x}, \quad x > 0$$

the preceding shows that for $a > 0$

$$P_f\{X > a\} = \frac{e^{-\lambda a}}{\lambda} E_g[e^{\lambda X} f(X) | X > a]$$

Because the conditional distribution of an exponential random variable that is conditioned to exceed a has the same distribution as a plus the exponential, the preceding gives that

$$\begin{aligned} P_f\{X > a\} &= \frac{e^{-\lambda a}}{\lambda} E_g\left[e^{\lambda(X+a)} f(X + a)\right] \\ &= \frac{1}{\lambda} E_g[e^{\lambda X} f(X + a)] \end{aligned}$$

Thus, we can estimate the tail probability $P_f\{X > a\}$ by generating $X_1, \ldots, X_k$, independent exponential random variables with rate λ, and then using

$$\frac{1}{\lambda}\frac{1}{k}\sum_{i=1}^{k} e^{\lambda X_i} f(X_i + a)$$

as the estimator.

As an illustration of the preceding, suppose that f is the density function of a standard normal random variable Z, and that $a > 0$. With X being an exponential random variable with rate $\lambda = a$, the preceding yields that

$$\begin{aligned} P\{Z > a\} &= \frac{1}{a\sqrt{2\pi}} E[e^{aX-(X+a)^2/2}] \\ &= \frac{e^{-a^2/2}}{a\sqrt{2\,\pi}} E[e^{-X^2/2}] \end{aligned}$$

Thus we can estimate $P\{Z > a\}$ by generating X, an exponential random variable with rate a, and then using

$$EST = \frac{e^{-a^2/2}}{a\sqrt{2\,\pi}} e^{-X^2/2}$$

as the estimator. To compute the variance of this estimator note that

$$\begin{aligned} E[e^{-X^2/2}] &= \int_0^\infty e^{-x^2/2} a e^{-ax}\, dx \\ &= a \int_0^\infty \exp\{-(x^2 + 2ax)/2\}\, dx \\ &= a e^{a^2/2} \int_0^\infty \exp\{-(x + a)^2/2\}\, dx \\ &= a e^{a^2/2} \int_a^\infty \exp\{-y^2/2\}\, dy \\ &= a e^{a^2/2} \sqrt{2\pi}\ \overline{\Phi}(a) \end{aligned}$$

Similarly, we can show that

$$E[e^{-X^2}] = a e^{a^2/4} \sqrt{\pi}\ \overline{\Phi}(a/\sqrt{2})$$

Combining the preceding then yields Var(EST). For instance, when $a = 3$

$$E[e^{-X^2/2}] = 3e^{4.5} \sqrt{2\pi}\ \overline{\Phi}(3) \approx 0.9138$$

and

$$E[e^{-X^2}] = 3e^{2.25} \sqrt{\pi}\ \overline{\Phi}(2.1213) \approx 0.8551$$

giving that

$$\text{Var}(e^{-X^2/2}) \approx .8551 - (.9138)^2 = 0.0201$$

Because $\frac{e^{-4.5}}{3\sqrt{2\pi}} \approx 0.001477$, we obtain, when $a = 3$, that

$$\text{Var}(EST) = (0.001477)^2 \text{Var}(e^{-X^2/2}) \approx 4.38 \times 10^{-8}$$

As a comparison, the variance of the raw simulation estimator, equal to 1 if a generated standard normal exceeds 3 and to 0 otherwise, is $P\{Z > 3\}(1 - P\{Z > 3\}) \approx 0.00134$. Indeed, the variance of EST is so small that the estimate from a single exponential will, with 95 percent confidence, be within ± 0.0004 of the correct answer.

Example 9z Importance sampling and conditional expectation can sometimes be combined by using the identity

$$E_f[X] = E_f\left[E_f[X|Y]\right] = E_g\left[E_f[X|Y]\frac{f(X)}{g(X)}\right]$$

For instance, suppose we were interested in estimating $P(X_1 + X_2 > 10) = E[I\{X_1 + X_2 > 10\}]$, where X_1 and X_2 are independent exponentials with mean 1. If we estimate the preceding via importance sampling, with g being the joint density of two independent exponentials with mean 5, then X_1, X_2 is generated according to g and the estimator is

$$I\{X_1 + X_2 > 10\}\frac{e^{-(X_1+X_2)}}{\frac{1}{25}e^{-(X_1+X_2)/5}} = 25\, I\{X_1 + X_2 > 10\}e^{-\frac{4}{5}(X_1+X_2)} \leqslant 25\, e^{-8}$$

On the other hand, we could first condition on X_1 to obtain that

$$P(X_1 + X_2 > 10|X_1) = \begin{cases} 1, & \text{if } X_1 > 10 \\ e^{-(10-X_1)}, & \text{if } X_1 \leqslant 10 \end{cases}$$

That is, $P(X_1 + X_2 > 10|X_1) = e^{-(10-X_1)^+}$. Hence, if we now estimate $E[e^{-(10-X_1)^+}]$ by importance sampling, sampling X_1 from an exponential distribution with mean 10, then the estimator of $P(X_1 + X_2 > 10)$ is

$$e^{-(10-X_1)^+}\frac{e^{-X_1}}{\frac{1}{10}e^{-X_1/10}} = 10\, e^{-(10-X_1)^+}e^{-.9X_1} \leqslant 10\, e^{-9}$$

where the inequality follows because

$$X_1 \leqslant 10 \Rightarrow e^{-(10-X_1)^+}e^{-.9X_1} = e^{-(10-X_1/10)} \leqslant e^{-9}$$

and

$$X_1 > 10 \Rightarrow e^{-(10-X_1)^+}e^{-.9X_1} = e^{-.9X_1} \leqslant e^{-9}$$

□

9.7 Using Common Random Numbers

Suppose that each of n jobs is to be processed by either of a pair of identical machines. Let T_i denote the processing time for job $i, i, = 1, \ldots, n$. We are

interested in comparing the time it takes to complete the processing of all the jobs under two different policies for deciding the order in which to process jobs. Whenever a machine becomes free, the first policy, called longest job first, always chooses the remaining job having the longest processing time, whereas the second policy, called shortest job first, always selects the one having the shortest processing time. For example, if $n = 3$ and $T_1 = 2$, $T_2 = 5$, and $T_3 = 3$, then the longest job first would complete processing at time 5, whereas the shortest job first would not get done until time 7. We would like to use simulation to compare the expected difference in the completion times under these two policies when the times to process jobs, $T_1, \ldots, T_n$, are random variables having a given distribution F.

In other words, if $g(t_1, \ldots, t_n)$ is the time it takes to process the n jobs having processing times $t_1, \ldots, t_n$ when we use the longest job first policy and if $h(t_1, \ldots, t_n)$ is the time when we use the shortest first policy, then we are interested in using simulation to estimate

$$\theta = \theta_1 - \theta_2$$

where

$$\theta_1 = E[g(\mathbf{T})], \quad \theta_2 = E[h(\mathbf{T})], \quad \mathbf{T} = (T_1, \ldots, T_n)$$

If we now generate the vector $\mathbf{T}$ to compute $g(\mathbf{T})$, the question arises whether we should use those same generated values to compute $h(\mathbf{T})$ or whether it is more efficient to generate an independent set to estimate θ_2. To answer this question suppose that we used $\mathbf{T}^* = (T_1^*, \ldots, T_n^*)$, having the same distribution as $\mathbf{T}$, to estimate θ_2. Then the variance of the estimator $g(\mathbf{T}) - h(\mathbf{T}^*)$ of θ is

$$\begin{aligned} \mathrm{Var}(g(\mathbf{T}) - h(\mathbf{T}^*)) &= \mathrm{Var}(g(\mathbf{T})) + \mathrm{Var}(h(\mathbf{T}^*)) - 2\mathrm{Cov}(g(\mathbf{T}),\ h(\mathbf{T}^*)) \\ &= \mathrm{Var}(g(\mathbf{T})) + \mathrm{Var}(h(\mathbf{T})) - 2\mathrm{Cov}(g(\mathbf{T}),\ h(\mathbf{T}^*)) \end{aligned} \tag{9.15}$$

Hence, if $g(\mathbf{T})$ and $h(\mathbf{T})$ are positively correlated—that is, if their covariance is positive—then the variance of the estimator of θ is smaller if we use the same set of generated random values $\mathbf{T}$ to compute both $g(\mathbf{T})$ and $h(\mathbf{T})$ than it would be if we used an independent set $\mathbf{T}^*$ to compute $h(\mathbf{T}^*)$ [in this latter case the covariance in (9.15) would be 0].

Since both g and h are increasing functions of their arguments, it follows, because increasing functions of independent random variables are positively correlated (see the Appendix of this chapter for a proof), that in the above case it is more efficient to successively compare the policies by always using the same set of generated job times for both policies.

As a general rule of thumb when comparing different operating policies in a randomly determined environment, after the environmental state has been simulated one should then evaluate all the policies for this environment. That is, if the environment is determined by the vector $\mathbf{T}$ and $g_i(\mathbf{T})$ is the return from policy i under the environmental state $\mathbf{T}$, then after simulating the value of the random vector $\mathbf{T}$ one should evaluate, for that value of $\mathbf{T}$, all the returns $g_i(\mathbf{T})$.

9.8 Evaluating an Exotic Option

With time 0 taken to be the current time, let $P(y)$ denote the price of a stock at time y. A common assumption is that a stock's price evolves over time according to a geometric Brownian motion process. This means that, for any price history up to time y, the ratio of the price at time $t+y$ to that at time y has a lognormal distribution with mean parameter μt and variance parameter $t\sigma^2$. That is, independent of the price history up to time y, the random variable

$$\log\left(\frac{P(t+y)}{P(y)}\right)$$

has a normal distribution with mean μt and variance $t\sigma^2$. The parameters μ and σ are called, respectively, the drift and the volatility of the geometric Brownian motion.

A European call option on the stock, having expiration time t and strike K, gives its owner the right, but not the obligation, to purchase the stock at time t for a fixed price K. The option will be exercised at time t provided that $P(t) > K$. Because we are able to purchase a stock whose market price is $P(t)$ for the price K, we say that our gain in this case is $P(t) - K$. Thus, in general, the gain at time t from the option is

$$(P(t) - K)^+$$

where

$$x^+ = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leqslant 0 \end{cases}$$

For a given initial price $P(0) = v$, let $C(K, t, v)$ denote the expected value of the payoff from a K, t European call option. Using that

$$W \equiv \log(P(t)/v)$$

is a normal random variable with mean $t\mu$ and variance $t\sigma^2$, we have that

$$C(K, t, v) = E[(P(t) - K)^+] = E[(ve^W - K)^+]$$

It is not difficult to explicitly evaluate the preceding to obtain $C(K, t, v)$.

The preceding option is called a standard (or *vanilla*) call option. In recent years there has been an interest in nonstandard (or *exotic*) options. Among the nonstandard options are the barrier options; these are options that only become alive, or become dead, when a barrier is crossed. We will now consider a type of barrier option, called an up-and-in option, that is specified not only by the price K and time t, but also by an additional price b and an additional time s, $s < t$. The conditions of this option are such that its holder only has the right to purchase the stock at time t for price K if the stock's price at time s exceeds b. In other words,

the K, t option either becomes alive at time s if $P(s) > b$, or becomes dead if $P(s) \leqslant b$. We now show how we can efficiently use simulation to find the expected payoff of such an option.

Suppose that $P(0) = v$, and define X and Y by

$$X = \log\left(\frac{P(s)}{v}\right), \qquad Y = \log\left(\frac{P(t)}{P(s)}\right)$$

It follows from the properties of geometric Brownian motion that X and Y are independent normal random variables, with X having mean $s\mu$ and variance $s\sigma^2$, and Y having mean $(t-s)\mu$ and variance $(t-s)\sigma^2$. Because

$$P(s) = ve^X$$
$$P(t) = ve^{X+Y}$$

we can write the payoff from the option as

$$\text{payoff} = I(ve^X > b)(ve^{X+Y} - K)^+$$

where

$$I(ve^X > b) = \begin{cases} 1, & \text{if } ve^X > b \\ 0, & \text{if } ve^X \leqslant b \end{cases}$$

Therefore, the payoff can be be simulated by generating a pair of normal random variables. The raw simulation estimator would first generate X. If X is less than $\log(b/v)$, that run ends with payoff value 0; if X is greater than $\log(b/v)$, then Y is also generated and the payoff from that run is the value of $(ve^{X+Y} - K)^+$.

We can, however, significantly improve the efficiency of the simulation by a combination of the variance reduction techniques of stratified sampling and conditional expectation. To do so, let R denote the payoff from the option, and write

$$\begin{aligned} E[R] &= E[R|ve^X > b]P\{ve^X > b\} + E[R|ve^X \leqslant b]P\{ve^X \leqslant b\} \\ &= E[R|X > \log(b/v)]P\{X > \log(b/v)\} \\ &= E[R|X > \log(b/v)]\overline{\Phi}\left(\frac{\log(b/v) - s\mu}{\sigma\sqrt{s}}\right) \end{aligned}$$

where $\overline{\Phi} = 1 - \Phi$ is the standard normal tail distribution function. Therefore, to obtain $E[R]$ it suffices to determine its conditional expectation given that $X > \log(b/v)$, which can be accomplished by first generating X conditional on the event that it exceeds $\log(b/v)$. Suppose that the generated value is x (we will show in the following how to generate a normal conditioned to exceed some value). Now, rather than generating the value of Y to determine the simulated payoff, let us take as our estimator the conditional expected payoff given the value of X.

This conditional expectation can be computed because, as $X > \log(b/v)$, the option is alive at time s and thus has the same expected payoff as would a standard option when the initial price of the security is ve^X and the option expires after an additional time $t - s$. That is, after we simulate X conditional on it exceeding $\log(b/v)$, we should use the following estimator for the expected payoff of the barrier option:

$$\text{Estimator} = C(K, t-s, ve^X)\overline{\Phi}\left(\frac{\log(b/v) - s\mu}{\sigma\sqrt{s}}\right) \tag{9.16}$$

After k simulation runs, with X_i being the generated value of the conditioned normal on run i, the estimator is

$$\overline{\Phi}\left(\frac{\log(b/v) - s\mu}{\sigma\sqrt{s}}\right)\frac{1}{k}\sum_{i=1}^{k} C(K, t-s, ve^{X_i})$$

We now show how to generate X conditional on it exceeding $\log(b/v)$. Because X can be expressed as

$$X = s\mu + \sigma\sqrt{s}Z \tag{9.17}$$

where Z is a standard normal random variable, this is equivalent to generating Z conditional on the event that

$$Z > c \equiv \frac{\log(b/v) - s\mu}{\sigma\sqrt{s}} \tag{9.18}$$

Thus, we need to generate a standard normal conditioned to exceed c.

When $c \leqslant 0$, we can just generate standard normals until we obtain one larger than c. The more interesting situation is when $c > 0$. In this case, an efficient procedure is to use the rejection technique with g being the density function of $c + Y$, where Y is an exponential random variable whose rate λ will be determined in the following. The density function of $c + Y$ is

$$g(x) = \lambda e^{-\lambda x}e^{\lambda c} = \lambda e^{-\lambda(x-c)}, \quad x > c$$

whereas that of the standard normal conditioned to exceed c is

$$f(x) = \frac{1}{\sqrt{2\pi}\,\overline{\Phi}(c)}e^{-x^2/2}, \quad x > c$$

Consequently,

$$\frac{f(x)}{g(x)} = \frac{e^{-\lambda c}e^{\lambda x - x^2/2}}{\lambda\overline{\Phi}(c)\sqrt{2\pi}}$$

Because $e^{\lambda x - x^2/2}$ is maximized when $x = \lambda$, we obtain that

$$\max_x \frac{f(x)}{g(x)} \leqslant C(\lambda) \equiv \frac{e^{\lambda^2/2 - \lambda c}}{\lambda\overline{\Phi}(c)\sqrt{2\pi}}$$

Calculus now shows that $C(\lambda)$ is minimized when

$$\lambda = \frac{c + \sqrt{c^2 + 4}}{2}$$

Take the preceding to be the value of λ. Because

$$\frac{f(x)}{C(\lambda)g(x)} = e^{\lambda x - x^2/2 - \lambda^2/2} = e^{-(x-\lambda)^2/2}$$

we see that the following algorithm generates a standard normal random variable that is conditioned to exceed the positive value c.

1. Set $\lambda = \frac{c+\sqrt{c^2+4}}{2}$.
2. Generate U_1 and set $Y = -\frac{1}{\lambda}\log(U_1)$ and $V = c + Y$.
3. Generate U_2.
4. If $U_2 \leqslant e^{-(V-\lambda)^2/2}$ stop; otherwise return to 2.

The value of V obtained is distributed as a standard normal random variable that is conditioned to exceed $c > 0$.

Remarks

- The preceding algorithm for generating a standard normal conditioned to exceed c is very efficient, particularly when c is large. For instance, if $c = 3$ then $\lambda \approx 3.3$ and $C(\lambda) \approx 1.04$.
- The inequality in Step 4 can be rewritten as

  $$-\log(U_2) \geqslant (V - \lambda)^2/2$$

  Using that $-\log(U_2)$ is exponential with rate 1, and that conditional on an exponential exceeding a value the amount by which it exceeds it is also exponential with the same rate, it follows that not only does the preceding algorithm yield a standard normal conditioned to exceed c, but it also gives an independent exponential random variable with rate 1, which can then be used in generating the next conditioned standard normal.
- Using that $C(K, t, v)$, the expected payoff of a standard option, is an increasing function of the stock's initial price v, it follows that the estimator given by (9.16) is increasing in X. Equivalently, using the representation of Equation (9.17), the estimator (9.16) is increasing in Z. This suggests the use of Z as a control variate. Because Z is generated conditional on the inequality (9.18), its mean is

  $$E[Z|Z > c] = \frac{1}{\sqrt{2\pi}\,\overline{\Phi}(c)} \int_c^\infty x e^{-x^2/2} dx$$

  $$= \frac{e^{-c^2/2}}{\sqrt{2\pi}\,\overline{\Phi}(c)}$$

- The expected return from the barrier option can be expressed as a two-dimensional integral involving the product of normal density functions. This two-dimensional integral can then be evaluated in terms of the joint probability distribution of random variables having a bivariate normal distribution. However, for more general payoff functions than $(P(t) - K)^+$, such as power payoffs of the form $[(P(t) - K)^+]^\alpha$, such expressions are not available, and the simulation procedure described might be the most efficient way to estimate the expected payoff. □

9.9 Appendix: Verification of Antithetic Variable Approach When Estimating the Expected Value of Monotone Functions

The following theorem is the key to showing that the use of antithetic variables will lead to a reduction in variance in comparison with generating a new independent set of random numbers whenever the function h is monotone in each of its coordinates.

Theorem *If $X_1, \ldots, X_n$ are independent, then for any increasing functions f and g of n variables*

$$E[f(\mathbf{X})g(\mathbf{X})] \geqslant E[f(\mathbf{X})]E[g(\mathbf{X})] \tag{9.19}$$

where $\mathbf{X} = (X_1, \ldots, X_n)$.

Proof The proof is by induction on n. To prove it when $n = 1$, let f and g be increasing functions of a single variable. Then for any x and y

$$[f(x) - f(y)][g(x) - g(y)] \geqslant 0$$

since if $x \geqslant y$ ($x \leqslant y$) then both factors are nonnegative (nonpositive). Hence, for any random variables X and Y,

$$[f(X) - f(Y)][g(X) - g(Y)] \geqslant 0$$

implying that

$$E\{[f(X) - f(Y)][g(X) - g(Y)]\} \geqslant 0$$

or, equivalently

$$E[f(X)g(X)] + E[f(Y)g(Y)] \geqslant E[f(X)g(Y)] + E[f(Y)g(X)]$$

If we now suppose that X and Y are independent and identically distributed then, as in this case,

$$E[f(X)g(X)] = E[f(Y)g(Y)]$$
$$E[f(X)g(Y)] = E[f(Y)g(X)] = E[f(X)]E[g(X)]$$

we obtain the result when $n = 1$.

So assume that Equation (9.19) holds for $n-1$ variables, and now suppose that $X_1, \ldots, X_n$ are independent and f and g are increasing functions. Then

$$\begin{aligned}
&E[f(\mathbf{X})g(\mathbf{X})|X_n = x_n] \\
&\quad = E[f(X_1, \ldots, X_{n-1}, x_n)g(X_1, \ldots, X_{n-1}, x_n)|X_n = x] \\
&\quad = E[f(X_1, \ldots, X_{n-1}, x_n)g(X_1, \ldots, X_{n-1}, x_n)] \\
&\quad \text{by independence} \\
&\quad \geqslant E[f(X_1, \ldots, X_{n-1}, x_n)]E[g(X_1, \ldots, X_{n-1}, x_n)] \\
&\quad \text{by the induction hypothesis} \\
&\quad = E[f(\mathbf{X})|X_n = x_n]E[g(\mathbf{X})|X_n = x_n]
\end{aligned}$$

Hence,

$$E[f(\mathbf{X})g(\mathbf{X})|X_n] \geqslant E[f(\mathbf{X})|X_n]E[g(\mathbf{X})|X_n]$$

and, upon taking expectations of both sides,

$$\begin{aligned}
E[f(\mathbf{X})g(\mathbf{X})] &\geqslant E[E[f(\mathbf{X})|X_n]E[g(\mathbf{X})|X_n]] \\
&\geqslant E[f(\mathbf{X})]E[g(\mathbf{X})]
\end{aligned}$$

The last inequality follows because $E[f(\mathbf{X})|X_n]$ and $E[g(\mathbf{X})|X_n]$ are both increasing functions of X_n, and so, by the result for $n = 1$,

$$\begin{aligned}
E[E[f(\mathbf{X})|X_n]E[g(\mathbf{X})|X_n]] &\geqslant E[E[f(\mathbf{X})|X_n]]E[E[g(\mathbf{X})|X_n]] \\
&= E[f(\mathbf{X})]E[g(\mathbf{X})]
\end{aligned}$$

□

Corollary If $h(x_1, \ldots, x_n)$ is a monotone function of each of its arguments, then, for a set $U_1, \ldots, U_n$ of independent random numbers,

$$\text{Cov}[h(U_1, \ldots, U_n), h(1-U_1, \ldots, 1-U_n)] \leqslant 0$$

Proof By redefining h we can assume, without loss of generality, that h is increasing in its first r arguments and decreasing in its final $n-r$. Hence, letting

$$\begin{aligned}
f(x_1, \ldots, x_n) &= h(x_1, \ldots, x_r, 1-x_{r+1}, \ldots, 1-x_n) \\
g(x_1, \ldots, x_n) &= -h(1-x_1, \ldots, 1-x_r, x_{r+1}, \ldots, x_n)
\end{aligned}$$

it follows that f and g are both increasing functions. Thus, by the preceding theorem,

$$\text{Cov}[f(U_1, \ldots, U_n), g(U_1, \ldots, U_n)] \geqslant 0$$

or, equivalently,

$$\begin{aligned}
\text{Cov}[&h(U_1, \ldots, U_r, 1-U_{r+1}, \ldots, 1-U_n), \\
&h(1-U_1, \ldots, 1-U_r, U_{r+1}, \ldots, U_n)] \leqslant 0
\end{aligned}$$

The result now follows since the random vector $h(U_1, \ldots, U_n), h(1 - U_1, \ldots, 1 - U_n)$ has the same joint distribution as does the random vector

$$h(U_1, \ldots, U_r, 1 - U_{r+1}, \ldots, 1 - U_n),$$
$$h(1 - U_1, \ldots, 1 - U_r, U_{r+1}, \ldots, U_n)$$

Exercises

1. Suppose we wanted to estimate θ, where

$$\theta = \int_0^1 e^{x^2}\, dx$$

Show that generating a random number U and then using the estimator $e^{U^2}(1 + e^{1-2U})/2$ is better than generating two random numbers U_1 and U_2 and using $[\exp(U_1^2) + \exp(U_2^2)]/2$.

2. Explain how antithetic variables can be used in obtaining a simulation estimate of the quantity

$$\theta = \int_0^1 \int_0^1 e^{(x+y)^2}\, dy\, dx$$

Is it clear in this case that using antithetic variables is more efficient than generating a new pair of random numbers?

3. Let $X_i, i = 1, \ldots, 5$, be independent exponential random variables each with mean 1, and consider the quantity θ defined by

$$\theta = P\left\{\sum_{i=1}^{5} i X_i \geqslant 21.6\right\}$$

(a) Explain how we can use simulation to estimate θ.
(b) Give the antithetic variable estimator.
(c) Is the use of antithetic variables efficient in this case?

4. Show that if X and Y have the same distribution then $\text{Var}[(X + Y)/2] \leqslant \text{Var}(X)$, and conclude that the use of antithetic variables can never increase variance (although it need not be as efficient as generating an independent set of random numbers).

5. (a) If Z is a standard normal random variable, design a study using antithetic variables to estimate $\theta = E[Z^3 e^Z]$.

(b) Using the above, do the simulation to obtain an interval of length no greater than 0.1 that you can assert, with 95 percent confidence, contains the value of θ.

6. Suppose that X is an exponential random variable with mean 1. Give another random variable that is negatively correlated with X and that is also exponential with mean 1.

7. Verify Equation (9.1).

8. Verify Equation (9.2).

9. Let $U_n, n \geqslant 1$, be a sequence of independent uniform (0, 1) random variables. Define

$$S = \min(n : U_1 + \cdots + U_n > 1)$$

It can be shown that S has the same distribution as does N in Example 9e, and so $E[S] = e$. In addition, if we let

$$T = \min(n : 1 - U_1 + \cdots + 1 - U_n > 1)$$

then it can be shown that $S + T$ has the same distribution as does $N + M$ in Example 9e. This suggests the use of $(S + T + N + M)/4$ to estimate e. Use simulation to estimate $\text{Var}(N + M + S + T)/4$.

10. In certain situations a random variable X, whose mean is known, is simulated so as to obtain an estimate of $P\{X \leqslant a\}$ for a given constant a. The raw simulation estimator from a single run is I, where

$$I = \begin{cases} 1 & \text{if } X \leqslant a \\ 0 & \text{if } X > a \end{cases}$$

Because I and X are clearly negative correlated, a natural attempt to reduce the variance is to use X as a control—and so use an estimator of the form $I + c(X - E[X])$.

(a) Determine the percentage of variance reduction over the raw estimator I that is possible (by using the best c) if X were uniform on (0, 1).
(b) Repeat (a) if X were exponential with mean 1.
(c) Explain why we knew that I and X were negatively correlated.

11. Show that $\text{Var}(\alpha X + (1 - \alpha)W)$ is minimized by α being equal to the value given in Equation (9.3) and determine the resulting variance.

12. (a) Explain how control variables may be used to estimate θ in Exercise 1.
(b) Do 100 simulation runs, using the control given in (a), to estimate first c^* and then the variance of the estimator.

(c) Using the same data as in (b), determine the variance of the antithetic variable estimator.
(d) Which of the two types of variance reduction techniques worked better in this example?

13. Repeat Exercise 12 for θ as given in Exercise 2.

14. Repeat Exercise 12 for θ as given in Exercise 3.

15. Show that in estimating $\theta = E[(1 - U^2)^{1/2}]$ it is better to use U^2 rather than U as the control variate. To do this, use simulation to approximate the necessary covariances.

16. Let $U_i, i \geqslant 1$, be independent uniform (0, 1) random variables and let

$$N = \min(n : U_n > .8).$$

(a) What is the distribution of N?
(b) Use Wald's equation to find $E[\sum_{i=1}^{N} U_i]$.
(c) What is $E[U_i|N = n]$ when $i < n$?
(d) What is $E[U_n|N = n]$?
(e) Verify the result of Wald's equation by conditioning on N. That is, by using

$$E[S] = \sum_{n=1}^{\infty} E[S|N = n]P(N = n)$$

where $S = \sum_{i=1}^{N} U_i$.

17. Let X and Y be independent with respective distributions F and G and with expected values μ_x and μ_y. For a given value t, we are interested in estimating $\theta = P\{X + Y \leqslant t\}$.

(a) Give the raw simulation approach to estimating θ.
(b) Use "conditioning" to obtain an improved estimator.
(c) Give a control variable that can be used to further improve upon the estimator in (b).

18. Suppose that Y is a normal random variable with mean 1 and variance 1, and suppose that, conditional on $Y = y$, X is a normal random variable with mean y and variance 4. We want to use simulation to efficiently estimate $\theta = P\{X > 1\}$.

(a) Explain the raw simulation estimator.
(b) Show how conditional expectation can be used to obtain an improved estimator.
(c) Show how the estimator of (b) can be further improved by using antithetic variables.

(d) Show how the estimator of (b) can be further improved by using a control variable.

Write a simulation program and use it to find the variances of

(e) The raw simulation estimator.
(f) The conditional expectation estimator.
(g) The estimator using conditional expectation along with antithetic variables.
(h) The estimator using conditional expectation along with a control variable.
(i) What is the exact value of θ?

[Hint: Recall that the sum of independent normal random variables is also normal.]

19. The number of casualty insurance claims that will be made to a branch office next week depends on an environmental factor U. If the value of this factor is $U = u$, then the number of claims will have a Poisson distribution with mean $\frac{15}{0.5+u}$. Assuming that U is uniformly distributed over (0, 1), let p denote the probability that there will be at least 20 claims next week.

(a) Explain how to obtain the raw simulation estimator of p.
(b) Develop an efficient simulation estimator that uses conditional expectation along with a control variable.
(c) Develop an efficient simulation estimator that uses conditional expectation and antithetic variables.
(d) Write a program to determine the variance of the estimators in parts (a), (b), and (c).

20. (The Hit–Miss Method.) Let g be a bounded function over the interval [0, 1]—for e xample, suppose $0 \leqslant g(x) \leqslant b$ whenever $0 \leqslant x \leqslant 1$—and suppose we are interested in using simulation to approximate $\theta = \int_0^1 g(x)\,dx$. The hit–miss method for accomplishing this is to generate a pair of independent random numbers U_1 and U_2. Now set $X = U_1$, $Y = bU_2$ so that the random point (X, Y) is uniformly distributed in a rectangle of length 1 and height b. Now set

$$I = \begin{cases} 1 & \text{if } Y < g(x) \\ 0 & \text{otherwise} \end{cases}$$

That is, I is equal to 1 if the random point (X, Y) falls within the shaded area of Figure 9.4.

(a) Show that $E[I] = [\int_0^1 g(x)\,dx]/b$.
(b) Show that $\text{Var}(bI) \geqslant \text{Var}(g(U))$ and so the hit–miss estimator has a larger variance than simply computing g of a random number.

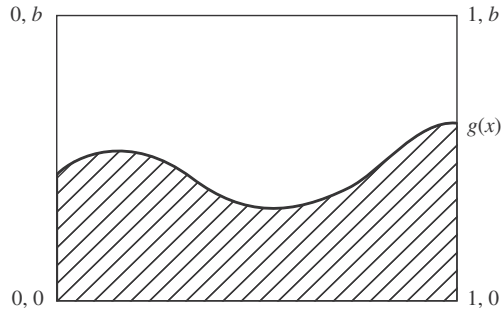

Figure 9.4. The Hit-Miss Method.

21. Let $X_1, \ldots, X_n$ be independent and identically distributed continuous random variables with distribution function F. Let $S_n = X_1 + \ldots + X_n$ and let $M_n = \max(X_1, \ldots, X_n)$. That is, let S_n and M_n be, respectively, the sum and the maximum of the n values. Suppose we want to use simulation to estimate $\theta = P(S_n > c)$.

(a) Show that $\theta = nP(S_n > c,\ M_n = X_n)$.
(b) Evaluate $P(S_n > c,\ M_n = X_n | X_1, \ldots, X_{n-1})$.
(c) Show that

$$\max_{x_1,\ldots,x_{n-1}} P(S_n > c,\ M_n = X_n | X_i = x_i, i \leqslant n-1) = nP(X_1 > c/n)$$

22. Suppose in the previous exercise that the random variables X_i are nonnegative. With $S_j = \sum_{i=1}^{j} X_i$ and $M_j = \max\{X_i,\ i = 1, \ldots, j\}$, let

$$R = \min\{n-1, \quad \min(j \geqslant 1 :\ S_j + M_j > c)\ \}$$

Let $\mathcal{E}$ be the estimator

$$\mathcal{E} = nP(S_n > c, X_n = M_n | R, X_1, \ldots, X_R)$$

Show that

$$\mathcal{E} = \begin{cases} \frac{n}{n-R}(1 - F^{n-R}(M_R)) & \text{if} \quad R < n-1 \\ P(S_n > c,\ M_n = X_n | X_1, \ldots, X_{n-1}) & \text{if} \quad R = n-1 \end{cases} \tag{9.20}$$

Explain why $\mathcal{E}$ is a better estimator of θ than is $P(S_n > c,\ M_n = X_n | X_1, \ldots, X_{n-1})$.

23. Suppose that customers arrive at a single-server queueing station in accordance with a Poisson process with rate λ. Upon arrival they either enter

service if the server is free or join the queue. Upon a service completion the customer first in queue, if there are any customers in queue, enters service. All service times are independent random variables with distribution G. Suppose that the server is scheduled to take a break either at time T if the system is empty at that time or at the first moment past T that the system becomes empty. Let X denote the amount of time past T that the server goes on break, and suppose that we want to use simulation to estimate $E[X]$. Explain how to utilize conditional expectation to obtain an efficient estimator of $E[X]$.

[Hint: Consider the simulation at time T regarding the remaining service time of the customer presently in service and the number waiting in queue. (This problem requires some knowledge of the theory of the $M/G/1$ busy period.)]

24. Consider a single serve queue where customers arrive according to a Poisson process with rate 2 per minute and the service times are exponentially distributed with mean 1 minute. Let T_i denote the amount of time that customer i spends in the system. We are interested in using simulation to estimate $\theta = E[T_1 + \cdots + T_{10}]$.

(a) Do a simulation to estimate the variance of the raw simulation estimator. That is, estimate $\text{Var}(T_1 + \cdots + T_{10})$.
(b) Do a simulation to determine the improvement over the raw estimator obtained by using antithetic variables.
(c) Do a simulation to determine the improvement over the raw estimator obtained by using $\sum_{i=1}^{10} S_i$ as a control variate, where S_i is the ith service time.
(d) Do a simulation to determine the improvement over the raw estimator obtained by using $\sum_{i=1}^{10} S_i - \sum_{i=1}^{9} I_i$ as a control variate, where I_i is the time between the ith and $(i + 1)$st arrival.
(e) Do a simulation to determine the improvement over the raw estimator obtained by using the estimator $\sum_{i=1}^{10} E[T_i|N_i]$, where N_i is the number in the system when customer i arrives (and so $N_1 = 0$).

25. Repeat Exercise 10 of Chapter 5, this time using a variance reduction technique as in Example 9m. Estimate the variance of the new estimator as well as that of the estimator that does not use variance reduction.

26. In Example 9r, compute $E[X|i]$ for $i = 0, 1, 2$.

27. Estimate the variance of the raw simulation estimator of the expected payoff in the video poker model described in Example 9r. Then estimate the variance using the variance reduction suggested in that e xample. What is your estimate of the expected payoff? (If it is less than 1, then the game is unfair to the player.)

28. In a certain game, the contestant can quit playing at any time and receive a final reward equal to their score at that time. A contestant who does not quit plays a game. If that game is lost, then the contestant must depart with a final reward of 0; if the game is won, the contestant's score increases by a positive amount having distribution function F. Each game played is won with probability p. A new contestant's strategy is to continue to play until her score exceeds a specified value c, at which point she will quit. Let R be her final reward.

(a) If we want to use simulation to estimate $E[R]$ by sequentially generating random variables I_i, X_i, $i = 1, \ldots,$ where $P(I_i = 1) = p = 1 - P(I_i = 0)$ and X_i has distribution F, when would a run end? and what would be the estimator from a single run?

(b) Show how to improve the estimator in part (a) by giving a second estimator that in each run generates only $X_1, \ldots, X_N$ where $N = \min(n : X_1 + \ldots + X_n > c)$.

29. A knockout tournament involving n competitors, numbered 1 through n, starts by randomly choosing two of the competitors to play a game, with the loser of the game departing the tournament and the winner getting to play another game against a randomly chosen remaining competitor. This continues through $n - 1$ games, and the player who wins the final game is declared the winner of the tournament. Whenever players i and j play against each other, suppose that i wins with probability $P_{i,j}$, where $P_{i,j}, i \neq j$, are specified probabilities such that $P_{i,j} + P_{j,i} = 1$. Let W_i denote the probability that i is the winner of the tournament. A simulation study has been developed to estimate the probabilities $W_1, \ldots, W_n$. Each simulation run begins by generating a random permutation of $1, \ldots, n$. If the random permutation is $I_1, \ldots, I_n$, then contestants I_1 and I_2 play the first game, with the winner being I_1 if a generated random number is less than P_{I_1,I_2}, and being I_2 otherwise. The winner of the first game then plays I_3, with the winner of that game decided by the value of another random number, and so on. If J is the winner in a simulation run, then the estimates of W_i from that run are 0 for all $i \neq J$, and 1 for $i = J$.

(a) Explain how conditional expectation can be used to improve the estimator of W_i. *Hint:* Condition on the permutation and whatever other information is needed to be able to determine the conditional probability that i is the winner of the tournament.

(b) Explain how post-stratification, relating to the random permutation, can be employed to further improve the estimator of W_i.

30. We proved that stratifying on Y always results in at least as great a reduction in variance as would be obtained by using Y as a control. Does that imply that in a simulation based on n runs it is always better to estimate $E[h(U)]$

by stratifying on I, where $I = i$ if $\frac{i-1}{n} < U < \frac{i}{n}$, rather than using U as a control variable?

31. For the compound random vector estimator $\mathcal{E}$ of Section 9.5.3, show that

$$\text{Var}(\mathcal{E}) \leqslant \text{Var}(g_N(X_1, \ldots, X_N))$$

Hint: Show that $\mathcal{E}$ is a conditional expectation estimator.

32. Suppose we want to use simulation to determine $\theta = E[h(Z_1, \ldots, Z_n)]$ where $Z_1, \ldots, Z_n$ are independent standard normal random variables, and where h is an increasing function of each of its coordinates. Let $W = \sum_{i=1}^{n} a_i Z_i$, where all the a_i are nonnegative. Using the following lemma, explain how we can use stratified sampling, stratifying on W, to approximate θ. Assume that the inverse transform method will be used to simulate W.

Lemma. If the standard normal random variable Z is independent of X, a normal random variable with mean μ and variance σ^2, then the conditional distribution of Z given that $Z + X = t$ is normal with mean $\frac{t-\mu}{1+\sigma^2}$ and variance $\frac{\sigma^2}{1+\sigma^2}$.

33. Explain how the approach of the preceding problem can be used when $h(x_1, \ldots, x_n)$ is an increasing function of some of its variables, and a decreasing function of the others.

34. Let $X_1, \ldots, X_k$ be independent Bernoulli random variables with parameters $p_1, \ldots, p_k$. Show how you can use the recursion formula given in Section 9.5.4 to generate $X_1, \ldots, X_k$ conditional on $\sum_{i=1}^{k} X_i = r$.

35. If X is such that $P\{0 \leqslant X \leqslant a\} = 1$, show that

(a) $E[X^2] \leqslant aE[X]$.
(b) $\text{Var}(X) \leqslant E[X](a - E[X])$.
(c) $\text{Var}(X) \leqslant a^2/4$.

[Hint: Recall that $\max_{0 \leqslant p \leqslant 1} p(1-p) = \frac{1}{4}$.]

36. Suppose we have a "black box" which on command can generate the value of a gamma random variable with parameters $\frac{3}{2}$ and 1. Explain how we can use this black box to approximate $E[e^X/(X+1)^2]$, where X is an exponential random variable with mean 1.

37. Suppose in Exercise 13 of Chapter 6 that we are interested in using simulation to estimate p, the probability that the system fails by some fixed time t. If p is very small, explain how we could use importance sampling to obtain a more efficient estimator than the raw simulation one. Choose some values for α, C, and t that make p small, and do a simulation to estimate the variance

of an importance sampling estimator as well as the raw simulation estimator of p.

38. In Example 9y, X_i are independent exponentials with rates $i/(i+2)$, $i = 1, 2, 3, 4$. With $S_j = \sum_{i=1}^{j} X_i$, that example was concerned with estimating

$$E[S_4|S_4 > 62] = \frac{E[S_4 I\{S_4 > 62\}]}{E[I\{S_4 > 62\}]}$$

(a) Determine $E[S_4 I\{S_4 > 62\}|S_3 = x]$.
(b) Determine $E[I\{S_4 > 62\}|S_3 = x]$.
(c) Explain how you can use the preceding to estimate $E[S_4|S_4 > 62]$.
(d) Using the preceding, show that $E[S_4|S_4 > 62] > 68$.

39. Consider two different approaches for manufacturing a product. The profit from these approaches depends on the value of a parameter α, and let $v_i(\alpha)$ denote the profit of approach i as a function of α. Suppose that approach 1 works best for small values of α in that $v_1(\alpha)$ is a decreasing function of α, whereas approach 2 works best for large values of α in that $v_2(\alpha)$ is an increasing function of α. If the daily value of α is a random variable coming from the distribution F, then in comparing the average profit of these two approaches, should we generate a single value of α and compute the profits for this α, or should we generate α_1 and α_2 and then compute $v_i(\alpha_i)$, $i = 1, 2$?

40. Consider a list of n names, where n is very large, and suppose that a given name may appear many times on the list. Let $N(i)$ denote the number of times the name in position i appears on the list, $i = 1, \ldots, n$, and let θ denote the number of distinct names on the list. We are interested in using simulation to estimate θ.

(a) Argue that $\theta = \sum_{i=1}^{n} \frac{1}{N(i)}$.
Let X be equally likely to be $1, \ldots, n$. Determine the name in position X and go through the list starting from the beginning, stopping when you reach that name. Let $Y = 1$ if the name is first reached at position X and let $Y = 0$ otherwise. (That is, $Y = 1$ if the first appearance of the name is at position X.)
(b) Argue that $E[Y|N(X)] = \frac{1}{N(x)}$.
(c) Argue that $E[nY] = \theta$.
(d) Now, let $W = 1$ if position X is the last time that the name in that position appears on the list, and let it be 0 otherwise. (That is, $W = 1$ if going from the back to the front of the list, the name is first reached at position X.) Argue that $n(W + Y)/2$ is an unbiased estimator of θ.
(e) Argue that if every name on the list appears at least twice, then the estimator in (d) is a better estimator of θ than is $(nY_1 + nY_2)/2$ where Y_1 and Y_2 are independent and distributed as is Y.

(f) Argue that $n/(N(X))$ has smaller variance than the estimator in (e), although the estimator in (e) may still be more efficient when replication is very high because its search process is quicker.

41. Let $\Phi^{-1}(x)$ be the inverse function of the standard normal distribution function $\Phi(x)$. Assuming that you can efficiently compute both $\Phi(x)$ and $\Phi^{-1}(x)$, show that you can generate a standard normal random variable X that is conditioned to exceed c by generating a random number U, letting $Y = U + (1 - U)\Phi(c)$, and setting

$$X = \Phi^{-1}(Y)$$

Explain how you could generate a standard normal random variable X that is conditioned to lie between a and b.

Bibliography

Hammersley, J. M., and D. C. Handscomb, *Monte Carlo Methods*. Wiley, New York, 1964.

Hammersley, J. M., and K. W. Morton, "A New Monte Carlo Technique: Antithetic Variables," *Proc. Cambridge Phil. Soc.*, **52**, 449–474, 1956.

Lavenberg, S. S., and P. D. Welch, "A Perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo Simulations," *Management Sci.*, **27**, 322–335, 1981.

Morgan, B. J. T., *Elements of Simulation*. Chapman and Hall, London, 1983.

Ripley, B., *Stochastic Simulation*. Wiley, New York, 1986.

Ross, S. M., and K. Lin, "Applying Variance Reduction Ideas in Queuing Simulations," *Probability Eng. Informational Sci.*, **15**, 481–494, 2001.

Rubenstein, R. Y., *Simulation and the Monte Carlo Method*. Wiley, New York, 1981.

Siegmund, D., "Importance Sampling in the Monte Carlo Study of Sequential Tests," *Ann. Statistics*, **4**, 673–684, 1976.