RATIO ESTIMATES IN MONTE CARLO SIMULATIONS

Irwin Kabak and Bertram Price

New York University

1. Introduction

Simulation is a method of analysis that has found widespread application. Almost all simulations involve some probabilistic aspects and are known as Monte Carlo simulations. Because of the probabilistic aspects of these types of simulations the outputs are themselves probabilistic in nature. Actually, the outputs are random variables and as such, possess distribution functions which must be estimated. An example follows.

Consider a warehousing system with a given set of initial conditions and a proposed operating procedure that is to be evaluated via the simulation. The output of the simulation is to be in the following scenario:

a. the proportion of orders received that are shipped on the day received.

b. the proportion of orders that are shipped within one day, two days, one week and two weeks.

Another example might be the proportion of airline flights that take off within five minutes of schedule and another the distribution of profit from operations in a manufacturing enterprise.

In all of the above examples, estimates of proportions are required. Although the entire distribution function (or equivalent) might be desirable, it is impractical to estimate more than just a few points. An analogy exists in comparing acceptance sampling plans; in such comparisons only the manufacturer's risk (also known as the producer's risk), the 50% point and the consumer's risk are often used.

The everpresent question of length of run is with us. In order to ascertain whether or not the estimated proportions (the answers) are within the desired limits of accuracy, an estimate of variance is required. Herein we shall address ourselves to sampling methods for estimating a single proportion emphasizing statistical properties of the estimates.

2. Sampling methods

When estimating proportions such as those described above or when estimating the value of any parameter from simulation output, two basic approaches to data collection may be considered. Both approaches involve the summarization of data over an interval of time with each selected interval giving rise to one data point or observation. The two approaches differ with respect to the definition of the time interval or what shall be referred to as the sampling period. In one approach, the sampling period is defined as a fixed length of system time. With the other approach the sampling period is defined by the state of the system. For example, the sampling period may be a tour or it may be defined in terms of a regeneration point. (3,5) One simple example of a sampling period that depends on the state of the system is the time it takes a fixed number of elements to traverse the system. An argument for associating the sampling period with the state of the system is that through some clever definition of the interval a sample of statistically independent observations may be drawn. As is well known, observations on simulation output taken sequentially usually possess a strong dependence structure. In some situations a detailed analysis of dependence is required. (2) More often the parameters of the dependence structure (the auto-covariance function) are viewed as nuisance parameters. In this latter case, any sampling method that yields an estimate of the parameter of interest and at the same time has minimal involvement with auto-covariance is desirable.

Herein, nonzero covariance is viewed as a nuisance parameter. However, estimates based on sampling periods of fixed length will be developed. This definition for the sampling period has been chosen because it is apparent that the state dependent definition has at least one major deficiency, namely, that the system time required to obtain an observation is unknown in advance. In the case of certain types of regeneration points, the sampling period may be unreasonably long. If a reasonable sampling period is to be defined in terms of the state of the system, enough must be known about the system in advance to anticipate the length of the

sampling period, as well as to determine which type of sampling periods could yield independent observations. However, if simulation has been chosen as the method for analyzing the system, it is very unlikely that enough would be known of the characteristics of the system in advance to define effective state dependent sampling periods.

In summary, an estimation procedure is desired that is consistent with two major objectives. First the sampling plan should be easy to implement and should not depend on extensive a priori knowledge about the simulation output. Second, the estimators should retain their good properties (e.g., unbiased, efficient, etc.) for a wide class of simulation structures. It is proposed that the ratio type estimator discussed in the remainder of this paper is an estimator that meets these two objectives. The estimator is first analyzed under the assumption that statistically independent observations are available. Then it is discussed within a more realistic framework characterized by Markov-type dependence.

3. Independent Observations

Let X_i be the number of arrivals in the i^{th} sampling period and Y_i be the number of arrivals in the i^{th} sampling period with a given attribute. The sample consists of n statistically independent pairs, (X_i, Y_i). We shall interpret Y_i as the number of successes in X_i independent trials so that Y_i is conditionally distributed as Binomial $Bi(X_i;\theta)$. θ is the probability of success on a given trial or equivalently the proportion of arrivals with the given attribute. θ is the parameter to be estimated. Since we are sampling over intervals of fixed length of time, the $\{X_i\}$ are independent and identically distributed according to some frequency function $g(x)$. Note that it is possible that there may be no arrivals in a sampling period, i.e., $X_i = 0$. We choose to disregard intervals with $X_i = 0$. Therefore, $g(x)$ will be of the form $g(x) = f(x)/(1 - f(0))$, $x = 1,2,\ldots$, where $f(x)$ is a frequency function over the non-negative integers. $g(x)$ is then a conditional frequency function conditioned on the event $X_1 > 0$.

We consider two logical candidates as estimators for θ. The first is in the form of a standard ratio estimator given as $\hat{\theta}_1 = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} X_i$.

The second is the average of the ratios for each observation, $\hat{\theta}_2 = n^{-1} \sum_{i=1}^{n} Y_i/X_i$. Both estimators are unbiased which follows easily from the facts that given $X_1,\ldots,X_n$

(i) Y_i is distributed as $Bi(X_i;\theta)$

and

(ii) ΣY_i is distributed as $Bi(\Sigma X_i;\theta)$.

The variances may be computed using the formula

$$\text{Var}(U) = E[\text{Var}(U|V)] + \text{Var}[E(U|V)]. \qquad (3.1)$$

For $\hat{\theta}_1$, replace U with $\hat{\theta}_1$ and V with ΣX_i. Then

$$\text{Var}(\hat{\theta}_1) = E[\theta(1 - \theta)/\Sigma X_i] + \text{Var}[\theta]$$
$$= \theta(1 - \theta)\, E(1/\Sigma X_i).$$

For $\hat{\theta}_2$, let $U = \hat{\theta}_2$ and V be the vector $(X_1,\ldots,X_n)$. Then $\text{Var}(\hat{\theta}_2) = \theta(1 - \theta)n^{-1}\, E(1/X)$ where X is a random variable with frequency function $g(x)$.

To compare $\hat{\theta}_1$ and $\hat{\theta}_2$, we must compare their variances. Let $X_1,\ldots,X_n$ be given. Then it is easy to show that in the class of linear unbiased estimators of θ, the weighted average of $\{Y_i/X_i\}$ with weights equal to $\{X_i/\sum_{i=1}^{n} X_i\}$ has minimum variance.(6) Therefore,

$$\text{Var}(\hat{\theta}_1|X_1,\ldots,X_n) \le \text{Var}(\hat{\theta}_2|X_1,\ldots,X_n)$$

and using (3.1) it follows that

$$\text{Var}(\hat{\theta}_1) \le \text{Var}(\hat{\theta}_2).$$

To substantiate the claim that $\hat{\theta}_1$ is a best estimator, we compute a lower bound for the variance of an unbiased estimator of θ. (6) With the joint frequency of (X_i,Y_i) given as

$$h(x,y) = \binom{x}{y} \theta^y (1 - \theta)^{x-y} g(x) \quad y \le x,\ x > 0,$$

it follows that the bound is

$$1/nE\left[\frac{\partial \ln h}{\partial \theta}\right]^2 = \theta(1 - \theta)/nE(X).$$

Since $E(1/\bar{X})$ approaches $1/E(X)$ as n becomes large, it follows that $\text{Var}(\hat{\theta}_1)$ approaches the bound so that $\hat{\theta}_1$ is an asymptotically efficient estimator for θ.

4. Dependent Observations

The estimation problem with dependent data will be discussed from the point of view of Markov dependence. First the general Markov structure of the data is developed. Then the stimation problem is formulated and the estimator is analyzed.

Let $Z_1,Z_2,\ldots$ be a sequence of random variables defined as follows: Z_i is equal to 1 if the i^{th} arrival has a given attribute and Z_i is equal to zero in all other cases. Clearly Z_i is a Bernoulli type random variable. As above, let θ be the probability that the i^{th} arrival has the given attribute, or equivalently

$$P(Z_i = 1) = \theta$$

Assuming that the sequence $\{Z_i\}$ has Markov structure, let

$$P(Z_i = 1|Z_{i-1} = 1) = \lambda$$

from which it follows that

$$P(Z_i = 1 | Z_{i-1} = 0) = (1 - \lambda)\,\theta/(1 - \theta)$$

and

$$\text{Cov}\,(Z_i, Z_{i+k}) = \theta(1 - \theta)[(\lambda-\theta)/1-\theta)]^k$$

(See (4).) For the types of dependence under consideration (e.g. delayed services, etc.) it is implied that Cov (Z_i, Z_{i+k}) should be positive. Therefore, we shall always have $\lambda \geq \theta$, noting that $\lambda = \theta$ corresponds to independence of the $\{Z_i\}$.

Turning to the estimation of θ, as above, let X_i be the number of arrivals in the i^{th} sampling period and let Y_i denote the number of arrivals in the i^{th} sampling period with the given characteristic. Note that Y_i is the sum of X_i variates from the sequence $\{Z_i\}$. We shall focus our attention on the better of the two estimators from section 3. Therefore, let $\hat{\theta} = \Sigma Y_i / \Sigma X_i$. As before, the estimator is easily seen to be unbiased. Utilizing formula (3.1) the variance of θ is

$$\text{Var}\,(\hat{\theta}) = E\,\{(\Sigma X_i)^{-2}\,[\Sigma\,\text{Var}(Y_i | X_1, \ldots, X_n) + 2 \underset{i<j}{\Sigma} \text{Cov}\,(Y_i, Y_j | X_1, \ldots, X_n)] \qquad (4.1)$$

where all summations are for $1 \leq i \leq n$.
For $n = 1$,

$$\text{Var}\,(\hat{\theta}) = E\,\{\frac{\theta(1 - \theta)}{X_1}\,[1 + 2\,\frac{(\lambda - \theta)}{(1 - \lambda)}\,[(X_1 - 1) - \frac{(\lambda - \theta)}{(1 - \lambda)}\,(1 - (\frac{\lambda - \theta}{1 - \theta})^{X_1 - 1})]]\} \qquad (4.2)$$

where the expectation is taken with respect to X_1. For $n > 1$, the variance of $\hat{\theta}$ is a function of covariances between Y_i and Y_{i+k}. We shall evaluate the covariances conditioned on $X_1, \ldots, X_n$. It is sufficient to consider the case $i = 1$ and arbitrary k. Then with $S_m = \sum_{i=1}^{m} X_i$, it follows that $Y_1 = \sum_{i=1}^{X_1} Z_i$ and $Y_{1+k} = \sum_{i=1}^{X_k} Z_{S_{k-1}+i}$. In order to simplify the notation but still retain the essence of the calculation, let $X_1 = X_k = X$.

Then the conditional covariance of Y_1 with Y_{1+k} is given as

$$\text{Cov}\,(Y_1, Y_{1+k} | X_1, \ldots, X_n) = X\,\alpha^{S_{k-1}}\,[1 + 2 \sum_{t=1}^{X-1} (1 - \frac{t}{X})\,\alpha^t] \qquad (4.3)$$

where $\alpha = (\lambda - \theta)/(1 - \theta)$. The form of this expression for arbitrary i and k can easily be obtained. Then substituting these expressions into the right hand side of equation (4.1) results in an expression for Var $(\hat{\theta})$.

Rather than displaying the full expression for Var $(\hat{\theta})$ it is more productive to analyze the role played by covariance terms such as the one derived above. Note that its contribution to the variance may be expressed as

$$E\,\{\frac{2X}{S_n^2}\,\alpha^{S_{k-1}}\,[1 + 2 \sum_{t=1}^{X-1} (1 - \frac{t}{X})\,\alpha^t]\}$$

where the expectation is taken with respect to $(X_1, \ldots, X_n)$. The right hand factor is bounded by $(1 + \alpha)/(1 - \alpha)$ and $0 < \alpha < 1$. Recall that X is an observation from a population with frequency function $g(x)$ and S_n and S_{k-1} are sums of n and k-1 independent observations respectively with that same frequency function. Therefore, X/S_n is small and approaching zero as n increases and $\alpha^{S_{k-1}}$ approaches zero as k is increased. Arguing heuristically it follows that for appropriate choices of n and k, the covariance of Y_i and Y_{i+k} is negligible and may be omitted in the computation of Var $(\hat{\theta})$.

The above analyses lead to an approach for sampling simulation output that has also been suggested by other authors. (1) For estimating θ, the sampling plan may be described as follows. Divide the total simulation running time into sample periods of equal length. Then observe X_i and Y_i for every k^{th} period and use the resulting data to construct $\hat{\theta}$. If k is sufficiently large, Y_i and Y_{i+k} will effectively have a zero covariance as shown above. It is clear that the sample size, n, the value of k and the total running time of the simulation are interrelated.

The problem of making an optimal choice of values for these three parameters is a problem in sampling design that remains to be investigated. However, with a good design the asymptotic variance of θ is determined from the equation

$$\lim_{n \to \infty} n\,E(X)\,\text{Var}(\hat{\theta}) = \theta(1-\theta)(1-2\theta+\lambda)/(1-\lambda) \qquad (4.4)$$

where E(X) is the average number of transactions observed per sampling period. If the number of observations were set in advance to be of the form $m = nE(X)$, then in the limit, m Var$(\Sigma Y_i/m)$ is identical to the right side of (4.4). Furthermore, $\Sigma Y_i/m$ is asymptotically efficient for θ. (4) Therefore $\hat{\theta}$ is asymptotically efficient.

5. Conclusions and Future Research

Analyses of ratio estimators as they apply to the estimation of proportions from simulation output have been presented. For the case of independent observations it is shown that the standard ratio estimator is a best linear unbiased estimator and that it is asymptotically efficient. For the more realistic case of dependent observations, a Markovian structure is described and the ratio-type estimator is shown to be unbiased and asymptotically efficient. It is noted that the variance expression depends on the value of a parameter that characterizes the Markov dependence.

The next step after insuring that efficient estimators are being used is to establish confidence intervals for the proportions being estimated. In future research we shall analyze the confidence intervals that are associated with the sampling plans

described in section 4 above. The analysis will be carried out under the assumption that the attribute data under observation may be approximately described by the Markov dependence structure. The problem of simultaneously estimating many proportions will also be considered.

BIBLIOGRAPHY

1. Conway, R.W., "Some Tactical Problems in Digital Simulation," in Management Science, Vol. 10, No. 1, October, 1963.

2. Fishman, George S., and Kiviat, Phillip J., "The Analysis of Simulation--Generated Time Series," in Management Science, Vol. 13, No. 7, March, 1967.

3. Kabak, Irwin W., "Stopping Rules for Queueing Simulations" in Operations Research, Vol. 16, No. 2, March-April, 1968.

4. Klotz, Jerome, "Statistical Inference in Bernoulli Trials with Dependence" in Annals of Statistics, Vol. 1, No. 2, March, 1973.

5. Saaty, Thomas L., Elements of Queueing Theory, McGraw-Hill, 1961.

6. Wilks, S.S., Mathematical Statistics, Wiley & Sons, 1962.