Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Letter to the Editor—Stopping Rules for Queuing Simulations

Irwin W. Kabak,

To cite this article:
Irwin W. Kabak, (1968) Letter to the Editor—Stopping Rules for Queuing Simulations. Operations Research 16(2):431-437. https://doi.org/10.1287/opre.16.2.431

Solutions to the Three-Dimensional Assignment Problem," Tech. Memo. No. 71, Operations Research Group, Case Institute of Technology, Cleveland, Ohio, October, 1966.

11. SCHELL, E., "Distribution of a Product by Several Properties," *Proc. Second Symp. in Linear Programming*, Washington, D. C., January 27–29, Vol. 1–2, pp. 615–642, 1955.

STOPPING RULES FOR QUEUING SIMULATIONS

Irwin W. Kabak

New York University, New York, New York

(Received March 8, 1967)

In queuing simulations when service times and/or inter-arrival times are exponentially distributed it is possible to obtain independent estimates of the quantities of interest, such as the probability of a request being served immediately or the proportion of requests that are delayed more than (say) t time units. It is shown that a weighted average of estimates of the probability of being served immediately is asymptotically unbiased for two simple queuing systems; it is also shown that an unweighted average is biased for one of these systems. Because the estimates are independent, the calculation of the variance of the weighted average is simplified. Expressions are presented for the calculation of the mean and variance of the estimate of interest. This paper presents only the nucleus of an idea and indicates several areas where research should prove useful.

AN EVER-PRESENT question in all simulations is 'Have enough trials been processed by the simulator?' A usual stopping rule is: stop the simulation when the (theoretical) variance of the statistic that estimates the quantity of interest is within given limits. However, there is an operational difficulty in the administration of this measure, i.e., because of the correlation of successive trials one cannot easily calculate the required variance. (This is discussed in some detail by HAUSER, ET AL.[3] If one could avoid the entire problem of correlated estimates and obtain independent estimates, the calculation of the required variance could be done more readily.

INDEPENDENT ESTIMATES FOR QUEUING SIMULATIONS

IT IS indeed fortunate for simulations of queuing systems that, under certain circumstances, independent estimates of the quantity of interest are available. COX AND SMITH[1] have introduced the concept of tours that we now present.

The *state* of the system is defined as the number of calls in the system either being served or waiting to be served. When an event (as arrival or a departure) causes the system to be in state j, we begin a *tour*. A departure from state j and a subsequent return, of the same type (arrival or departure) that began the tour completes it. Referring to Fig. 1, all tours illustrated for state C are [1, 3), [3, 7), [7, 10), [2, 4), and [4, 8). Note that [1, 2) is not a tour. In some queuing simulations cer-

tain tours are independent of other tours and therefore statistics from these tours are independent.

We will restrict ourselves to cases where interarrival times and service times are *independent* samples from given distributions.

The exponential distribution is the only continuous distribution that possesses no memory, i.e., if X is a positive random variable with continuous c.d.f. then

$$P(X<t)=P(X<t+t^*|X>t^*) \tag{1}$$

if and only if

$$P(X<t)=1-\exp(-at). \qquad (a>0) \tag{2}$$

For a proof, *see* PARZEN.[4] It is this property of regeneration by which the system under study gives rise to independent tours.

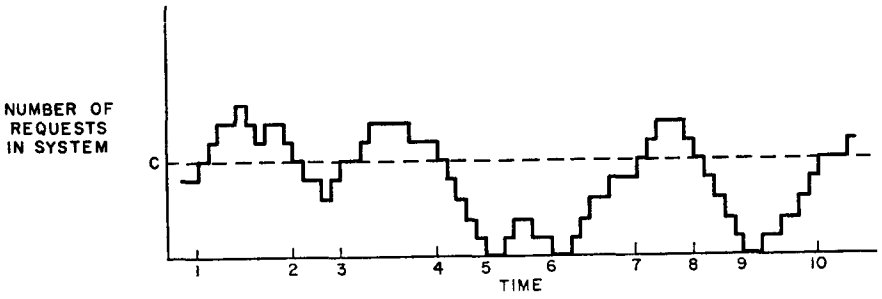

Fig. 1. Number of calls in system vs. time for a delay simulator with C servers.

In queuing simulations two factors control the state of the system: the manner in which requests arrive and the manner in which requests leave. Most queuing simulations concern themselves with waiting for the next event (an arrival or a departure) to occur.

If the distribution of service times is exponential then whenever an arrival causes the number of calls in the system to become equal to N (an arbitrary positive integer) we can begin an independent tour. This is so because all calls being served at the instant of this arrival are at regeneration points and are equivalently beginning their service anew, and the time until the next arrival is characterized by the distribution is exponential (i.e., Poisson arrivals) then we have regeneration points whenever a departure causes the system to be in state N. This is so because both the interarrival times and the service times are characterized by the exponential distribution, and therefore the remaining N service times and the interarrival time have regeneration points and equivalently begin anew at each departure epoch.

When the service times are nonexponential and the arrivals are Poisson then any departure that leaves the system empty causes an independent tour to begin at the departure epoch.

When the service times are nonexponential and the arrivals are Poisson then any departure that leaves the system empty causes an independent tour to begin at

the departure epoch. When neither the interarrival times nor the service times are exponentially distributed an arrival to an empty system begins a tour.

BERNOULLI TYPE ESTIMATES AND QUEUING SIMULATIONS

ONE IS often interested in the probability of blocking (not being served immediately) that attempts experience. In such cases one simply divides the blocked attempts by the total number of attempts and obtains an estimate of the probability directly. The question of how to handle the delay distribution is discussed. It is suggested that the estimate of the cumulative distribution function of the delays be obtained in a Bernoulli manner at a finite number of points. For example, we are interested in four points of the delay distribution: we want to know what proportion of the delayed requests are served within 0.1, 0.5, 1.0, and 2.0 average service times. From each tour we will obtain four estimates, $\hat{q}_j$, $j=1(1)4$, which characterize the conditional delay distribution. Each of these estimates are to be treated separately to obtain four variances. When all of the variances are within limits we may stop the simulation. Although it is recognized that the $\hat{q}_j$ are not independent and hence the respective variances are also not independent this is asserted to be a reasonable operational stopping rule.

CALCULATION OF THE VARIANCE

WE WILL now indicate specifically how we may calculate the variance of the estimate that interests us. We will restrict ourselves to statistics of a Bernoulli nature.

Let

N = the random number of arrivals during each independent tour.
N_i = the random number of arrivals during tour i.
n_i = the sample number of arrivals during tour i.

We will designate as N', N_i', and n_i' the arrivals of interest for the aforementioned tours; they may represent the number of attempts blocked, or perhaps the number of attempts served within 100 sec. The statistic of interest for tour i is

$$\hat{P}_i = (N_i')/N_i, \qquad [i=1(1)M] \quad (3)$$

where M equals the fixed number of tours under consideration.

It suggested that the weighted average over all tours be used for the grand estimate $\hat{P}$, namely

$$P = (\sum_{i=1}^{i=M} N_i')/(\sum_{i=1}^{i=M} N_i). \quad (4)$$

It is demonstrated in the section, "A Cleared System," that, for a particular queuing simulation, an unweighted estimate yields a biased result while the estimate in (4) is unbiased. We are interested in the variance of $\hat{P}$, viz.

$$V(\hat{P}) = V[(\sum_{i=1}^{i=M} N_i')/(\sum_{i=1} N_i)]. \quad (5)$$

An approximate expression for the variance of a ratio can be obtained by expanding the ratio in a Taylor series about the mean of the numerator and the mean of the denominator and then taking expected values. The result, given in reference 4 is (for $B>0$)

$$V(A/B)\approx E^2(A)V(B)/E^4(B)+V(A)/E^2(B)-2E(A)\mathrm{Cov}(A, B)/E^3(B). \quad (6)$$

For our specific case we have

$$A=\sum_{i=1}^{i=M} N_i', \quad (7)$$

$$B=\sum_{i=1}^{i=M} N_i. \quad (8)$$

The moments of (7) and (8) are in general not known, so sample estimates will be used on the right-hand side of (6).

Because the tours are independent we have

$$E(A)=E(N')\cdot M, \quad (9)$$

$$E(B)=E(N)\cdot M, \quad (10)$$

$$V(A)=V(N')\cdot M, \quad (11)$$

$$V(B)=V(N)\cdot M. \quad (12)$$

When sample values are used we obtain

$$E(A)\approx\sum_{j=1}^{j=M} n_j', \quad (13)$$

$$E(B)\approx\sum_{j=1}^{j=M} n_j, \quad (14)$$

$$V(A)\approx M[\sum_{j=1}^{j=M} n_j'^2-(\sum_{j=1}^{j=M} n_j')^2/M]/(M-1), \quad (15)$$

and

$$V(B)\approx M[\sum_{j=1}^{j=M} n_j^2-(\sum_{j=1}^{j=M} n_j)^2/M]/(M-1). \quad (16)$$

The covariance term is given by

$$\mathrm{Cov}(a, B)=E[\sum_{i=1}^{i=M} N_i'\sum_{j=1}^{j=M} N_j]-E[\sum_{i=1}^{i=M} N_i']E[\sum_{j=1}^{j=M} N_j], \quad (17)$$

or when using sample variances

$$\mathrm{Cov}(A, B)\approx\sum_{i=1}^{i=M} n_i'n_i-M^{-1}\sum_{i=1}^{i=M} n_i'\sum_{i=1}^{i=M} n_i. \quad (18)$$

There are instances where some of the values on the right side of (6) can be calculated theoretically, and this should be done whenever possible. For example, when one has a Poisson input of known intensity a, then

$$E(B)=aT,$$

and

$$V(B)=aT,$$

where T is the average length of tour and is known theoretically under certain circumstances.

Note that in a simple binomial case, with parameter p, with the total number of independent trials per tour fixed equal to N_0, A is the random number of trials that result in a success. Our estimate $\hat{p}$ that is, A/B, and the formula equation (6) is exact, yielding $p(1-p)/N_0$.

TWO EXAMPLES

We will show that the estimate in equation (4) is an asymptotically unbiased estimate of the probability of blocking in two simple examples.

A DELAY SYSTEM

Consider the simulation of a c-server delay system with Poisson input, exponential service times with unit mean, and any order of service. The probability of blocking

is independent of the order of service and equals $C(c, \lambda)$* where λ is the arrival rate and also the load in erlangs since the average service time is unity.

By the integral stationary theorm (*see* LOEVE[5]) we can express the expected value of the probability of blocking $E(\hat{p})$, from equation (4), as

$$E(\hat{p}) \approx E(N')/E(N) \quad \text{for} \quad M \text{ large}; \tag{19}$$

for tours using c, and for Poisson input and exponential service times we have (since the busy and nonbusy† periods are independently distributed)

$$E(\hat{p}) \approx (\lambda M_b)/(\lambda M_b + \lambda M_{nb}), \tag{20}$$

where

M_b = the mean length of time of the busy period, and

M_{nb} = the mean length of time of the nonbusy period.

This system, for multiple servers, has been analyzed by SEGAL[7] who obtained recurrence relations for the Laplace transforms, first and second moments, as well as explicit formulas for these moments of the distribution of time of a tour. Thus, we have

$$M_b = (c-\lambda)^{-1} \tag{21}$$

and

$$M_{nb} = (c-1)!\lambda^{-c}\sum_{i=0}^{c-1}(\lambda^i/i!). \tag{22}$$

Upon substituting one finds that

$$E(\hat{p}) \approx C(c, \lambda).$$

Thus we have shown that equation (4) yields an asymptotically unbiased estimate (as $M \to \infty$) for the probability of blocking when tours utilizing the number of servers are used, the input is Poisson, and the service time is exponential.

A CLEARED SYSTEM

CONSIDER a single-server system with Poisson input and exponential service times with unit mean in which attempts that are not serviced immediately are dismissed (cleared) from the sytem. The probability of blocking for this case is equal to $\lambda/(\lambda+1)$‡, where λ is the arrival rate and also the offered load in erlangs. As in the preceding section, we may write for the expected value of our estimate of the probability of blocking

$$E(\hat{p}) \approx E(N')/E(N) \quad \text{for} \quad M \text{ large}.$$

The average number of arrivals during a busy period is simply λ since the mean holding time is taken as unity. The total number of arrivals during a tour for this system is $\lambda+1$, since the first arrival after the nonbusy (idle) period begins, termi-

* This system is known as Erlang's delay system when the order of service is first-come-first-served. The probability of blocking was first obtained by the Danish mathematician A. K. ERLANG early in this century. It is given by

$$C(c, \lambda) = [1 + (c-1)!(c-\lambda)\lambda^{-c}\sum_{i=0}^{c-1}\lambda^i/i!]^{-1}$$

A derivation is given in DESCLOUX.[2]

† The nonbusy period is taken to be when not all servers are occupied. In the case of a single-server system the nonbusy period is known as the idle period.

‡ *See* COX AND SMITH,[1] section 24, iv.

nates the tour. Thus

$$E(\hat{p}) \approx \lambda/(\lambda+1), \tag{23}$$

and is asymptotically unbiased for this system.

For this loss system we can also show that an unweighted average of the estimate of the probability of blocking from each tour, viz., $\hat{p}_i$ yields a biased result. Since the tours are independent the estimate $\hat{p}_i$ is identically distributed for all i. This being the case we may take the expected value of $\hat{p}_i$ for just one tour. We will begin a tour immediately after an arrival epoch that finds the system empty; the tour will end at the next arrival epoch that finds the system empty. In this single-server system the probability of an arrival before a departure is $\lambda/(\lambda+1)$ and thus the number of arrivals before the departure* is distributed geometrically. The expected value of $\hat{p}_i$ is therefore given by

$$E(\hat{p}_i) = \sum_{j=0}^{j=\infty} j(j+1)^{-1}\lambda^j(\lambda+1)^{-j},^{-1}, \tag{24}$$

and after some simplification

$$E(\hat{p}_i) = 1-\lambda^{-1}\ln(\lambda+1), \tag{25}$$

so that

$$E(\hat{p}_i) \neq \lambda/(\lambda+1). \tag{26}$$

Thus, an unweighted average of the $\hat{p}_i$ yields a biased result.

THE NEED FOR FURTHER RESEARCH

WHEN SERVICE times are exponentially distributed one has the opportunity to choose the tours in many ways as compared to the case when only the interarrival distribution is exponential. In the latter one can choose tours only when a departure leaves the system empty. In the former case it is not clear if the tours should be chosen to, e.g., maximize the number of tours for a given length of run. If L is the average number of servers that are busy then it has been conjectured that the shortest tours occur at either $[L]$ or $[L]+1$. This conjecture has been proven by the writer for delay systems with one- and two-servers and Poisson input and exponential service time by using some results of Segal.[7] When the service times are *not* exponential the departure rate when a given number of servers are busy *might* approximate the situation when the service times *are* exponential. Under such circumstances the said number of servers should be able to be used for tours that are *essentially* independent. The approximate variance formula obtained from the propagation of errors could be evaluated thoroughly considering the myriad of distributions that arise in simulation.

In some simulations it is possible that the tours are so very long that length of run might become excessive. Allowing for such a contingency could wisely collect the simulation data to possibly use a subgrouping approach (for a complete discussion *see* HAUSER ET AL.[3]). It is also possible that the initial estimates $\hat{p}_i$ have a very small spurious variance that would result in stopping the simulation too soon. It is therefore desirable that some lower bound be placed on the number of tours

* This is equivalent to the number of arrivals before the server is idle.

completed before one considers stopping the simulation. Investigations into the determination of this lower bound should prove challenging.

REFERENCES

1. D. R. Cox and W. C. Smith, *Queues*, p. 136, Wiley, New York, 1961.
2. A. Descloux, *Delay Tables for Finite and Infinite Source Systems*, McGraw-Hill, New York, 1962.
3. N. Hauser, N. N. Barish, and S. Ehrenfeld, "Design Problems in a Process Control Simulation," *J. Indust. Eng.* **17** (1966).
4. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. I, p. 232, Griffin, London, 1958.
5. M. Loeve, *Probability Theory*, p. 419, Van Nostrand, New York, 1955.
6. E. Parzen, *Modern Probability Theory and its Applications*, p. 262, Wiley, New York, 1960.
7. M. Segal, "Some First Passage Problems in Queues," Unpublished Memorandum, Bell Telephone Laboratories, Inc., Holmdel, N.J., 1963.

QUASILINEARIZATION AND INVERSE PROBLEMS FOR LANCHESTER EQUATIONS OF CONFLICT

J. D. Buell, H. H. Kagiwada, and R. E. Kalaba

The Rand Corporation, Santa Monica, California

(Received October 24, 1966)

This letter applies quasilinearization and digital computation to the solution of certain problems concerning the Lanchester equations of combat. The study shows how the enemy's strength and replenishment rate may be estimated on the basis of information on friendly units. The problem is phrased as a nonlinear, multipoint, boundary-value problem, and quasilinearization is used for the numerical solution. The results of some numerical experiments performed with the aid of an IBM 7044 computer are presented, and the problem of errors in the observations is examined.

IN A recent survey article devoted to the Lanchester equations of combat, Dolanský[1] called attention to a number of unresolved problems in the theory:

1. Development of outcome-predicting relations that use information pertaining only to friendly units.

2. Determination of the parameters of enemy units on the basis of the numbers of friendly units and higher derivatives of these numbers.

3. Additional verification studies to establish the validity of Lanchester equations more firmly in more sophisticated situations.

The purpose of this letter is to show how a mathematical technique, quasilinearization,[2] together with digital computers, can shed light on these and related topics in military operations research. These techniques have been applied to