SIAM J. APPL. MATH.
Vol. 47, No. 1, February 1987

011

ESTIMATING TIME AVERAGES VIA RANDOMLY-SPACED OBSERVATIONS*

BENNETT L. FOX† AND PETER W. GLYNN‡

Abstract. To estimate continuous-time averages via randomly-spaced observations of discrete-event systems, we develop a point-process framework and use it to generalize both regenerative and stationary-process oriented simulation methodologies. We give consistent estimators, central limit theorems, and an effective bias-reducing jackknife. The impact on indirect estimation of transaction (customer) averages is discussed.

Key words. simulation, point processes

AMS(MOS) subject classifications. 62M15, 65C05, 68J05, 60G55

1. Introduction. Let $0=T_0<T_1<\cdots$ be *event times* with $\lim_{k\to\infty} T_k=\infty$ a.s. Though the sequence $\tilde{X}_0, \tilde{X}_1, \tilde{X}_2, \cdots$ with $\tilde{X}_k \in R^d$ is not necessarily an imbedded Markov chain, we call $\tilde{X}_k$ the *state* at time T_k—somewhat abusing the term. To define the state $X(t)$ at an arbitrary time t, we set

$$X(t)=\sum_{k=0}^{\infty} \tilde{X}_k I_{[T_k, T_{k+1})}(t) \tag{1.1}$$

where the indicator $I_A(t)$ is 1 or 0 depending on whether or not $t\in A$. For this definition to make sense, every state change must correspond to an event time. The state does not change continuously. It jumps at discrete (possibly random) times. In other words, we have a *discrete-event system.* Let f be a real-valued function. Put

$$r(t)=\frac{1}{t}\int_0^t f(X(s))\,ds. \tag{1.2}$$

We solve the *steady-state problem*: estimate the limit (when it exists)

$$r=\lim_{t\to\infty} r(t) \quad \text{a.s.} \tag{1.3}$$

and construct confidence intervals for r.

To do this, we develop a point-process framework and use it to generalize both regenerative and stationary-process oriented simulation methodologies. Simply averaging the $\tilde{X}_k$'s generally estimates r inconsistently. The T_k's are not necessarily regeneration times. We work with generally dependent observations, in contrast to regenerative approaches.

An alternative (traditional) approach takes equally-spaced observations

$$R_i(\Delta)=\Delta^{-1}\int_{(i-1)\Delta}^{i\Delta} f(X(s))\,ds$$

($\Delta>0$), sometimes approximating $R_i(\Delta)$ by, say,

$$\tilde{R}_i(\Delta)=f(X((i-1)\Delta)).$$

* Received by the editors December 28, 1983; accepted for publication (in revised form) March 21, 1986.

† Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Québec H3C 3J7. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada.

‡ Department of Industrial Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706. The work of this author was sponsored by the United States Army under contract DAAG29-80-C-0041.

Call the average of $R_i(\Delta)$'s

$$\bar{R}_N(\Delta) = N^{-1} \sum_{i=1}^{N} R_i(\Delta).$$

If $T = N\Delta$, then $\bar{R}_N(\Delta)$ is precisely identical to the point estimate $r(T)$, so that the mean and variance are (trivially) the same, respectively. It is well known that standard methods can be applied to the time series $\{R_i(\Delta): i \geqq 1\}$ to estimate $\bar{R}_N(\Delta)$'s variance and to construct confidence intervals for r. In this paper, however, we prefer to operate on the natural (random) time scale $(T_0, T_1, \cdots)$ of the process via equations (1.29)–(1.32) below, without introducing the parameter Δ. The time scale $(T_0, T_1, \cdots)$ has the computational advantage that it runs on the time scale of events simulated, thereby simplifying data collection. In addition, we expect that the correlation structure of the process X is more suitably estimated on the intrinsic natural time scale $(T_0, T_1, \cdots)$ than some arbitrary sequence of equally spaced instants $(\Delta, 2\Delta, \cdots)$. For example, if the time between events tends to be large, then X will be highly correlated, and one would generally prefer to take Δ large. However, by using the T_i's instead, one automatically compensates for this correlation effect, without any need to deal with choice of the parameter Δ.

Another estimator for r uses the $\tilde{R}_i(\Delta)$'s:

$$\hat{R}_N(\Delta) = N^{-1} \sum_{i=1}^{N} \tilde{R}_i(\Delta).$$

However, $\hat{R}_N(\Delta)$ need not even be consistent, for example, if X is periodic.

This discussion suggests that our approach is generally better than traditional ones, but probably not universally so.

1.1. Shifts. We use the set of functions $\omega : [0, \infty) \to \mathbb{R}^d$ that are right continuous and have left limits to describe the sample space Ω; denote the probability measure on Ω by P. Define $X : \Omega \to \Omega$ via

$$X(\cdot) = X(\cdot, \omega) = \omega(\cdot) \tag{1.4}$$

for $\omega \in \Omega$. For any random variable $R : \Omega \to [0, \infty)$, define a *right shift* $\theta_R : \Omega \to \Omega$ via

$$\theta_R = \theta_{R(\omega)}(\omega) = \omega(R(\omega) + \cdot). \tag{1.5}$$

Let $A \subset \mathbb{R}^d$, put $S_0 = 0$, and set

$$S_{k+1} = \inf\{t > S_k : X(t-) \neq X(t), X(t) \in A\} \tag{1.6}$$

where $X(t-)$ is the left limit of X at t. If $A = \mathbb{R}^d$, then $S_i = T_i$. Define

$$P_n\{X \in \cdot\} = P\{X \circ \theta_{S_n} \in \cdot\} \tag{1.7}$$

and

$$P_t\{X \in \cdot\} = P\{X \circ \theta_t \in \cdot\}. \tag{1.8}$$

For most simulations where the steady-state limit r exists, limit probabilities $\hat{P}$ and $\tilde{P}$ exist such that weak convergence holds:

$$P_n \Rightarrow \hat{P} \tag{1.9}$$

and

$$P_t \Rightarrow \tilde{P} \tag{1.10}$$

where

$$\hat{P}\{X \circ \theta_{S_1} \in \cdot\} = \hat{P}\{X \in \cdot\} \tag{1.11}$$

and for $t \geqq 0$ (*shift invariance*)

$$\tilde{P}\{X \circ \theta_t \in \cdot\} = \tilde{P}\{X \in \cdot\}. \tag{1.12}$$

Standard weak-convergence theory (Billingsley (1968), for example) assumes a closed bounded interval as domain, typically [0, 1]. Stone (1963) shows that this theory extends to semi-infinite domains $[0, \infty)$.

Example 1.1. Let X be a delayed (resp., nondelayed) regenerative process under P (resp., $\hat{P}$). Assume that X regenerates when it hits A. Then $P_n = \hat{P}$ for $n \geqq 1$ though generally $P_0 \neq \hat{P}$.

Example 1.2. Let X be as in Example 1.1. If the regenerative process is positive recurrent and the regeneration-spacing distribution (or the n-fold convolution of it for some n) has a nontrivial Lebesgue density component, then (1.10) and (1.12) hold (see Miller (1972)). Also, $\tilde{P}$ and $\hat{P}$ are related by

$$\tilde{P}\{X \in \cdot\} = \frac{1}{\hat{E}S_1} \int_0^\infty \hat{P}\{X \circ \theta_t \in \cdot\,; S_1 > t\}\, dt, \tag{1.13}$$

where $\hat{E}$ denotes expectation under $\hat{P}$. In § 3 (Theorem 6 (iii)) we show that (1.13) holds more generally. See the remarks following (1.41).

A process X satisfying

$$P\{X \circ \theta_{S_1} \in \cdot\} = P\{X \in \cdot\} \tag{1.14}$$

is *synchronous* with respect to the *imbedded point process* sequence $\{S_n\}$.

Example 1.3. Suppose that $\{\tilde{X}_n\}$ is a stationary sequence with $S_n = T_n = n$. Then $X(t)$ is synchronous with respect to $\{S_n\}$.

From Examples 1.1 and 1.3 we see that synchronous processes generalize both nondelayed regenerative processes and stationary sequences. Assume that the simulator (somehow) chooses the time origin so that everything representing the initial "transient" phase is to its left. This is trivial for regenerative processes but not in general. This deletion assumption translates mathematically as: (1.14) holds. Without the deletion assumption, we would have only (1.11) and then our results (1.23)-(1.28) under P would hold generally only under (the typically unknown) $\hat{P}$.

Let

$$\alpha_n = S_{n+1} - S_n \tag{1.15}$$

and

$$X_n(t) = \begin{cases} X(S_n + t), & 0 \leqq t < \alpha_n, \\ \infty, & t \geqq \alpha_n. \end{cases} \tag{1.16}$$

We assume that

(1.17) $\{X_n\}$ is a ϕ-mixing sequence of processes (Billingsley (1968, pp. 166–168)) with ϕ-mixing coefficients satisfying $\sum_{k=1}^\infty \phi_k^{1/2} < \infty$,

and

$$E[Y_n(|f|)^2 + \alpha_n^2] < \infty \tag{1.18}$$

where

$$Y_n(f) = \int_0^{\alpha_n} f(X_n(t))\, dt = \int_{S_n}^{S_{n+1}} f(X(t))\, dt. \tag{1.19}$$

Several times we use the fact that (1.17) implies that sequences $\{Y_n(f)\}, \{\alpha_n\}$, and $\{Y_n(f)\alpha_n\}$ are also ϕ-mixing with (different) mixing coefficients satisfying $\sum \phi_k^{1/2} < \infty$. The *cycle* sequence $\{X_n\}$ is not necessarily i.i.d. If it were, we could simply apply regenerative methodology. Three more definitions set the stage:

$$\bar{Y}_n(f) = \sum_{k=0}^{n-1} \frac{Y_k(f)}{n}, \tag{1.20}$$

$$\bar{\alpha}_n = \frac{S_n}{n}, \tag{1.21}$$

$$r_n = \frac{\bar{Y}_n(f)}{\bar{\alpha}_n}. \tag{1.22}$$

1.2. Preview of results. In § 2 we assume that (1.14), (1.17), and (1.18) hold under P, usually without further explicit mention. We show that under P,

$$r_n \to r \quad \text{a.s.}, \tag{1.23}$$

$$r(t) \to r \quad \text{a.s.}, \tag{1.24}$$

$$r = \frac{EY_0(f)}{E\alpha_0}, \tag{1.25}$$

and generalize the "inspection paradox"

$$\frac{1}{t}\int_0^t I_{[x,\infty)}(\alpha_{N(s)})\, ds \to \frac{1}{E\alpha_0}\int_x^\infty sP\{\alpha_0 \in ds\}, \tag{1.26}$$

where $N(s) = \max\{n\colon S_n \leqq s\}$ for all $s \geqq 0$. The left side of (1.26) is the proportion of time over $[0, t]$ that the cycle in progress had length at least x. Its limit, the right side of (1.26), is precisely that for the regenerative case; e.g., see Bratley, Fox and Schrage (1983, problem 3.7.4) or Heyman and Sobel (1982, § 5.5). Thus, for synchronous processes the cycle in progress at time t tends to be longer than average, thereby biasing $r(t)$.

Next, § 2 proves a central limit theorem (CLT) under P:

$$n^{1/2}(r_n - r) \Rightarrow \frac{\sigma N(0,1)}{E\alpha_0}, \tag{1.27}$$

$$t^{1/2}(r(t) - r) \Rightarrow \frac{\sigma N(0,1)}{(E\alpha_0)^{1/2}} \tag{1.28}$$

where $N(0, 1)$ is a zero-mean, unit-variance, normal random variable,

$$\sigma^2 = EY_0(\hat{f})^2 + 2\sum_{k=1}^{\infty} EY_0(\hat{f})Y_k(\hat{f}), \tag{1.29}$$

$$\hat{f}(x) = f(x) - r. \tag{1.30}$$

In other words, $Y_k(\hat{f}) = Y_k(f) - \alpha_k r$. If the synchronous process X is either regenerative or stationary with $S_n = n$, the CLT simplifies. For the former, the second term in (1.29)—corresponding to covariances—vanishes. For the latter, $\alpha_n \equiv 1$. It is perhaps surprising that in (1.27) we divide by $E\alpha_0$ whereas in (1.28) we divide by $(E\alpha_0)^{1/2}$. To check that this is reasonable, consider the i.i.d. case with $T_n = S_n = 2n$. Then r_n is (still) the average of n terms but $r(n)$ is the average of $n/2$ terms modulo rounding. So

$\sqrt{n}\,(r_n - r) \Rightarrow (\sigma/2)N(0,1)$ in agreement with (1.27) and $(\sqrt{n/2})\,[r(n)-r] \Rightarrow (\sigma/2)N(0,1)$, or equivalently $\sqrt{n}\,[r(n)-r] \Rightarrow (\sigma/\sqrt{2})N(0,1)$, in agreement with (1.28). Also when $A = \mathbb{R}^d$,

$$\sigma^2 = E\hat{f}^2(\tilde{X}_0)\tau_0^2 + 2\sum_{k=1}^{\infty} E\hat{f}(\tilde{X}_0)\hat{f}(\tilde{X}_k)\tau_0\tau_k, \tag{1.31}$$

where $\tau_k = T_{k+1} - T_k$.

To estimate σ^2 in the regenerative case, one uses the knowledge that the covariance terms in (1.29) vanish to construct the estimator σ_n^2:

$$\sigma_n^2 = \frac{1}{n-1}\sum_{k=0}^{n-1} (Y_k(f) - r_n\alpha_k)^2;$$

the estimator σ_n^2 is easily shown to be strongly consistent for σ^2.

In the general case, estimation of σ^2 is more complicated. The parameter σ^2 can be expressed in the form

$$\begin{aligned}\sigma^2 = &\left[\operatorname{var} Y_0(f) + 2\sum_{k=1}^{\infty} \operatorname{cov}(Y_0(f), Y_k(f))\right]\\ &-r\left[\operatorname{cov}(Y_0(f), \alpha_0) + 2\sum_{k=1}^{\infty}\operatorname{cov}(Y_0(f), \alpha_k)\right.\\ &\left.+\operatorname{cov}(\alpha_0, Y_0(f)) + 2\sum_{k=1}^{\infty}\operatorname{cov}(\alpha_0, Y_k(f))\right]\\ &+r^2\left[\operatorname{var}\alpha_0 + 2\sum_{k=1}^{\infty}\operatorname{cov}(\alpha_0, \alpha_k)\right].\end{aligned} \tag{1.32}$$

Still assuming (1.14) and hence no transient-phase contamination, the three bracketed terms appearing in (1.32) can be consistently estimated by c_{1n}, c_{2n}, c_{3n} (say); this can be accomplished by standard techniques (e.g. batch means, spectral methods, autoregressive procedures). Such methods include parameters (e.g. batch size, spectral window, autoregressive order) which must be keyed to sample size. In any case, the estimator σ_n^2 is then given by

$$\sigma_n^2 = c_{1n} - r_n c_{2n} + r_n^2 c_{3n}.$$

When using $r(t)$, replace n by

$$N(t) = \#S_i\text{'s observed in } (0, t]. \tag{1.33}$$

Glynn and Iglehart (1981), Jow (1982, pp. 54-56), and Streller (1980, Thm. 3.2) prove essentially (1.27) but under significantly different hypotheses.

From (1.27) and (1.28) we get the respective confidence intervals

$$[r_n - z_\delta\sigma_n/\bar{\alpha}_n\sqrt{n},\ r_n + z_\delta\sigma_n/\bar{\alpha}_n\sqrt{n}]$$

and

$$[r(t) - z_\delta\sigma_{N(t)}/(t\bar{\alpha}_{N(t)})^{1/2},\ r(t) + z_\delta\sigma_{N(t)}/(t\bar{\alpha}_{N(t)})^{1/2}],$$

where the percentile z_δ is the unique solution of $P[N(0,1) \leqq z_\delta] = 1 - \delta/2$. If σ_n is a strongly consistent estimator for σ, then these are asymptotically exact $100(1-\delta)\%$ confidence intervals for r.

Under an additional assumption

$$\text{(1.34)}\qquad \begin{array}{l}\text{there exists } K>0 \text{ such that } P\{\alpha_0 \leqq K\}=1 \\ \text{and } \sup\{|f(x)|: x\in\mathbb{R}^d\}\leqq K,\end{array}$$

§ 2 proves that

$$\text{(1.35)}\qquad Er_n = r - \frac{\beta}{(E\alpha_0)^2 n} + o\left(\frac{1}{n}\right)$$

and finds an expression (2.12) for β. The form of (1.35) motivates a jackknife:

$$\text{(1.36)}\qquad \tilde{r}_{2n} = 2r_{2n} - (r(0, n-1) + r(n, 2n-1))/2,$$

where

$$\text{(1.37)}\qquad r(a,b) = \frac{(\sum_{j=a}^{b} Y_j(f))}{(\sum_{j=a}^{b} \alpha_j)}.$$

It reduces bias by an order of magnitude:

$$\text{(1.38)}\qquad E\tilde{r}_{2n} = r + o\left(\frac{1}{n}\right).$$

Sometimes there is no bias. If $S_n = T_n = n$, then $Er_n = r$.

For another case, let $A = \mathbb{R}^d$ and suppose that the T_k's see time averages, in the sense (1.39) that the long-run average of the $f(\tilde{X}_k)$'s equals the continuous-time average r a.s. As an example, suppose that

(i) $N(t)$ is a stationary Poisson process;

(ii) for each $t \geqq 0$, $\{X(s-): 0 \leqq s \leqq t\}$ is independent of $\{N(t+u) - N(t): u \geqq 0\}$.

Wolff (1982) shows that, if $r(t) \to r$ a.s., then

$$\text{(1.39)}\qquad \hat{r}_n = \frac{1}{n}\sum_{k=0}^{n-1} f(\tilde{X}_k) \to r \quad \text{a.s.}$$

Since the summands are identically distributed, $E\hat{r}_n = r$, so $\hat{r}_n$ is unbiased.

Section 2 concludes by proving that

$$\text{(1.40)}\qquad \sqrt{2n}\,(\tilde{r}_{2n} - r) \Rightarrow \frac{\sigma N(0,1)}{E\alpha_0}.$$

Combining (1.27) and (1.40), we see that $\tilde{r}_{2n}$ and r_{2n} have the same asymptotic variance. So our jackknife reduces bias without increasing asymptotic variance.

Section 3 begins by proving that

$$\text{(1.41)}\qquad \frac{1}{t}\int_0^t \hat{P}\{X\circ\theta_s \in \cdot\}\, ds \to \tilde{P}\{X \in \cdot\},$$

where $\tilde{P}$ satisfies (1.13), assuming that (1.11), (1.17), and (1.18) hold under $\hat{P}$ but not necessarily under P. The point is that (1.13) remains valid when X is (merely) synchronous. Franken et al. (1982, Chap. 1) show that it holds under even weaker conditions.

When X is regenerative, $\hat{P}\{X \in \cdot\} = \tilde{P}\{X\circ\theta_{S_1} \in \cdot\}$, "inverting" (1.13). This simple inversion generally fails, assuming only (1.9)-(1.12). Recall that under $\tilde{P}$, the cycle trapping a fixed time tends to be longer than a typical cycle. In the regenerative case, only the actual trapping cycle is affected, due to independence of cycles. In general, however, neighboring cycles are also affected due to cycle correlation. Franken et al. (1982, p. 23) show that the general inverse to (1.13) is

$$\hat{P}\{X \in \cdot\} = \frac{1}{\tilde{E}N(1)} \sum_{k=1}^{\infty} \tilde{P}\{X \circ \theta_{S_k} \in \cdot\,; S_k \leqq 1\}, \tag{1.42}$$

where $\tilde{E}$ denotes expectation under $\tilde{P}$. Comparing (1.11) and (1.14), we see that X is synchronous under $\hat{P}$, the *Palm distribution* of $\tilde{P}$. Here X is *stationary* under $\tilde{P}$.

Chapter 1 of Franken et al. (1982) thoroughly discusses the relationships between $\hat{P}$ and $\tilde{P}$ and proves an intuitive alternative to (1.42):

$$\hat{P}\{X \in \cdot\} = \lim_{h \downarrow 0} \tilde{P}\{X \circ \theta_{S_1} \in \cdot \mid S_1 \leqq h\}. \tag{1.43}$$

This shows that the Palm distribution $\hat{P}$ is the stationary distribution $\tilde{P}$ conditioned on hitting A at time 0.

We want (1.12) to hold with P replacing $\tilde{P}$. This means that the simulator (somehow) deletes the entire transient phase, choosing the time origin so that this phase is to its left. Typically this is impossible to do exactly in practice, but to proceed mathematically we assume it has been done exactly. This translates as

$$P\{X \circ \theta_t \in \cdot\} = P\{X \in \cdot\} \tag{1.44}$$

holds. This is stronger than (1.14) because $Er(t) = r$ under (1.44) but not generally under (1.14). In fact (1.44) usually holds literally only if the initial state is generated by the (generally unknown) stationary distribution $\tilde{P}$. Even for regenerative processes, where synchronization via (1.14) is trivial, making (1.44) hold even to a first approximation is generally hard. Nevertheless, in the rest of this section and in § 3, usually without further explicit mention, we assume that (1.44) holds and that, under $\hat{P}$ but not necessarily under P, (1.11), (1.17), and (1.18) hold. By contrast, § 2 assumes that (1.14), (1.17), and (1.18) hold under P but does not assume (1.44). Thus neither section subsumes the other.

Several counterparts to results in § 2 are proved in § 3. There we show that, under $\tilde{P}$, $r(t)$ is (still) strongly consistent and

$$\sqrt{t}\,(r(t) - r) \Rightarrow \frac{\sigma N(0,1)}{(\hat{E}\alpha_0)^{1/2}} \tag{1.45}$$

where

$$\sigma^2 = \hat{E}Y_0(\hat{f})^2 + 2 \sum_{k=1}^{\infty} \hat{E}Y_0(\hat{f})\,Y_k(\hat{f}). \tag{1.46}$$

This CLT corresponds to (1.28) with $\hat{E}$ (expectation under $\hat{P}$) replacing E. Estimate σ^2 and construct confidence intervals just as before.

The stationarity assumptions (1.14) and (1.44) can be significantly relaxed. Our CLT's can be extended to include certain nonstationary processes by appealing to Billingsley (1968, Thm. 20.2). Furthermore, our discussion carries over to the steady-state estimation problem for

$$r = \lim_{t \to \infty} \frac{1}{t} \sum_{k=0}^{\infty} f(X(U_k))\, I_{[U_k, \infty]}(t)$$

(when the limit exists), where the U_k's are an increasing sequence of random times. Such limits are of interest, for example, in queues where lump-sum rewards are paid out to the server at customer departure epochs. Our arguments go through provided that one modifies the definition of $Y_n(f)$ to

$$Y_n(f) = \sum_{j=0}^{\infty} f(X(U_j)) I_{[S_n, S_{n+1})}(U_j).$$

1.3. Transactions. To essentially every time average r there corresponds a transaction (customer) average s and conversely. Heyman and Stidham (1980) and Heyman and Sobel (1982, § 11.3) establish this correspondence explicitly. Thus, every estimator of r yields an *indirect estimator* of s and conversely. In our framework, § 4 indicates that on balance estimating s indirectly is often better than estimating it directly. This conclusion appears contrary to some folklore.

Readers wishing to skip to § 4 can do so without loss of continuity.

2. Results and proofs. I. Throughout this section we assume that (1.14), (1.17) and (1.18) hold under P.

THEOREM 1. *Formulas* (1.23), (1.24), *and* (1.25) *hold.*

Proof. First, observe that $(Y_n(f), \alpha_n)$ is trivially a functional of $X_n(\cdot)$ and thus $\{(Y_n(f), \alpha_n): n \geqq 0\}$ is ϕ-mixing with the same mixing coefficients as the X_n's. Since any ϕ-mixing sequence is ergodic (Lamperti (1977, pp. 95–96)), apply Birkhoff's ergodic theorem (Heyman and Sobel (1982, p. 366)) to conclude that

$$\bar{Y}_n(f) \to EY_0(f) \quad \text{a.s.,} \tag{2.1}$$

$$\bar{\alpha}_n \to E\alpha_0 \quad \text{a.s.} \tag{2.2}$$

proving (1.23) with r given by (1.25). With $N(t)$ defined by (1.33),

$$\frac{S_{N(t)}}{N(t)} \leqq \frac{t}{N(t)} \leqq \frac{S_{N(t)+1}}{N(t)} \tag{2.3}$$

from the definitions. From $N(t) \uparrow \infty$ a.s. and (2.2), the extreme terms of (2.3) converge to $E\alpha_0$ a.s.; as $t/N(t)$ gets squeezed, it converges to $E\alpha_0$ a.s. For nonnegative f,

$$\frac{N(t)+1}{t}\left(\frac{N(t)}{N(t)+1}\right)\bar{Y}_{N(t)}(f) \leqq r(t) \leqq \frac{N(t)}{t}\left(\frac{N(t)+1}{N(t)}\right)\bar{Y}_{N(t)+1}(f). \tag{2.4}$$

Apply $N(t)\uparrow\infty$ a.s., (2.1), and $N(t)/t \to 1/E\alpha_0$ a.s. to quickly see that the extreme terms of (2.4) converge to r given by (1.25), proving (1.24) for $f \geqq 0$. Split a general f into its positive and negative parts, apply (1.24) to each, and recombine. To justify the last step, use $(E|Y_1(f)|)^2 \leqq EY_1(f)^2 < \infty$ by (1.18).

THEOREM 2. *Formula* (1.26) *holds.*

Proof. Observe that

$$\frac{1}{n}\int_0^{S_n} I_{[x,\infty)}(\alpha_{N(s)})\, ds = \frac{1}{n}\sum_{k=0}^{n-1} (\alpha_k) I_{[x,\infty)}(\alpha_k) \to E\{\alpha_0; \alpha_0 \geqq x\} \quad \text{a.s.} \tag{2.5}$$

($E\{X; A\} \triangleq EXI_A$, where I_A is 1 or 0 depending on whether or not A occurs) and then mimic the proof of Theorem 1.

Actually, our proof of Theorem 2 can easily be generalized to obtain a limit theorem for the average of $I_{(-\infty,x]}(\alpha_{N(s)}) \cdot I_{(-\infty,y]}(Y_{N(s)}(f))$ as time goes to infinity. This gives the joint limit distribution of cycle length and reward received over that cycle for the cycle covering the point s. It is precisely that derived by Wilson (1983) in the regenerative case.

THEOREM 3. *Formulas* (1.27) *and* (1.28) *hold.*

Proof. The sequence $\{Y_k(\hat{f})\}$ is ϕ-mixing with zero mean. It satisfies the conditions of Billingsley (1968, Thm. 20.1), and hence

$$\sqrt{n}\ \bar{Y}_n(\hat{f}) \Rightarrow \sigma N(0,1) \tag{2.6}$$

as $n \to \infty$, for $\sigma = 0$ (see Billingsley (1968, p. 177)). Rewrite (2.6) as

$$\sqrt{n}\ \bar{\alpha}_n(r_n - r) \Rightarrow \sigma N(0,1). \tag{2.7}$$

Now (1.27) follows from (2.7), (2.2), and a routine application of the converging-together lemma (Billingsley (1968, p. 25)).

For (1.28), let $\{t_n\}$ be an arbitrary sequence converging to ∞. Apply Billingsley's (1968, p. 146) random time change theorem to the weak invariance principle version of (2.6) to get

$$\sqrt{N(t_k)}\ \bar{Y}_{N(t_k)}(\hat{f}) \Rightarrow \sigma N(0,1) \tag{2.8}$$

as $k \to \infty$. Another application of the converging-together lemma then yields

$$\sqrt{t_k}\,(r_{N(t_k)} - r) \Rightarrow \frac{\sigma N(0,1)}{\sqrt{E\alpha_0}} \tag{2.9}$$

using $N(t)/t \to E\alpha_0$ proved just after (2.3) and $\sqrt{N(t_k)} = \sqrt{t_k}\,\sqrt{N(t_k)/t_k}$. Clearly

$$\sqrt{t_k}\,|r_{N(t_k)} - r(t_k)| \leqq \sqrt{t_k}\,[Y_{N(t_k)}(|f|) + |r_{N(t_k)}|\alpha_{N(t_k)}]/S_{N(t_k)}. \tag{2.10}$$

Combining a standard Borel–Cantelli argument and (1.18), proves that

$$\alpha_k/\sqrt{k} \to 0, \qquad Y_k(|f|)/\sqrt{k} \to 0 \quad \text{a.s.} \tag{2.11}$$

But the right-hand side of (2.10) can be rewritten as

$$\begin{aligned}&(t_k/N(t_k))^{1/2} \cdot (Y_{N(t_k)}(|f|)/\sqrt{N(t_k)}) \cdot (N(t_k)/S_{N(t_k)})\\ &\quad + (t_k/N(t_k))^{1/2} \cdot |r_{N(t_k)}| \cdot (\alpha_{N(t_k)}/\sqrt{N(t_k)}) \cdot (N(t_k)/S_{N(t_k)}),\end{aligned}$$

which, by (2.11), converges to $(E\alpha_0)^{1/2} \cdot 0 \cdot (E\alpha_0)^{-1} + (E\alpha_0)^{1/2} \cdot |r| \cdot 0 \cdot (E\alpha_0)^{-1} = 0$. Apply the converging-together lemma to (2.9) and (2.10) to get (1.28) for t going to ∞ through the sequence t_k. Because that sequence is arbitrary, (1.28) holds without qualification (see Billingsley (1968, p. 16)).

Now also assuming (1.34) we have the following theorem.

THEOREM 4. *Formula* (1.35) *holds with*

$$\beta = E\alpha_0 Y_0(\hat{f}) + \sum_{k=1}^{\infty} (E\alpha_0 Y_k(\hat{f}) + EY_0(\hat{f})\alpha_k). \tag{2.12}$$

Proof. The infinite series for β converges absolutely by a remark of Billingsley (1968, p. 177). Turning to the bias expansion itself, observe that

$$r_n - r = \bar{Y}_n(\hat{f})g(\bar{\alpha}_n), \tag{2.13}$$

where

$$g(x) = \frac{1}{x}. \tag{2.14}$$

Choosing ε small enough,

$$\sup\{g''(x): |x - E\alpha_0| \leqq \varepsilon\} = M < \infty. \tag{2.15}$$

Letting event $B_n = \{|\bar{\alpha}_n - E\alpha_0| \leqq \varepsilon\}$, use (1.34) and then Chebyshev's inequality to get

$$|E\{r_n - r;\ B_n^c\}| \leqq 2KP\{|\bar{\alpha}_n - E\alpha_0| > \varepsilon\} \leqq \frac{2KE(\bar{\alpha}_n - E\alpha_0)^4}{\varepsilon^4}. \tag{2.16}$$

By Billingsley (1968, Lemma 4, p. 172),

$$E(\bar{\alpha}_n - E\alpha_0)^4 = O(n^{-2}). \tag{2.17}$$

Next

$$E\{r_n - r\} = E\{r_n - r;\ B_n\} + E\{r_n - r;\ B_n^c\}.$$

Combining (2.16) and (2.17) gives

$$E\{r_n - r\} = E\{r_n - r;\ B_n\} + O(n^{-2}). \tag{2.18}$$

On the event B_n, expand $g(\bar{\alpha}_n)$ in a Taylor series around $E\alpha_0$ to get

$$E\{r_n - r;\ B_n\} = S + T \tag{2.19}$$

where

$$S = E\left\{\bar{Y}_n(\hat{f})\left[\frac{1}{E\alpha_0} - \frac{1}{(E\alpha_0)^2}(\bar{\alpha}_n - E\alpha_0)\right];\ B_n\right\}, \tag{2.20}$$

$$T = E\{\bar{Y}_n(\hat{f})g''(\xi_n)(\bar{\alpha}_n - E\alpha_0)^2/2;\ B_n\}, \tag{2.21}$$

and

$$|\xi_n - E\alpha_0| < \varepsilon. \tag{2.22}$$

Apply Cauchy-Schwartz and then Billingsley (1968, Lemmas 3 and 4, p. 172) with (1.17) to get

$$T \leqq ME(\bar{Y}_n(\hat{f})^2)^{1/2}(E(\bar{\alpha}_n - E\alpha_0)^4)^{1/2} = O(n^{-3/2}). \tag{2.23}$$

An argument similar to that justifying (2.18) gives

$$S = -E(\bar{Y}_n(\hat{f})\bar{\alpha}_n)/(E\alpha_0)^2 + O(n^{-2}). \tag{2.24}$$

A calculation similar to that in the proof of Lemma 3 in Billingsley (1968, p. 172) shows that

$$E(\bar{Y}_n(\hat{f})\bar{\alpha}_n) = \frac{\beta}{n} + o\left(\frac{1}{n}\right) \tag{2.25}$$

with β given by (2.12). Combining (2.18)-(2.25) gives (1.35), finishing the proof.

COROLLARY. *Jackknifing works*: (1.38) *holds.*

THEOREM 5. *Formula* (1.40) *holds.*

Proof. By the converging-together lemma, it suffices to prove that

$$\sqrt{n}\,(r_{2n} - \tilde{r}_{2n}) \to 0 \tag{2.26}$$

in probability. Straightforward algebra shows that the left side of (2.26) equals

$$-\sqrt{n}\,B_n(C_n - D_n)/2(C_n + D_n), \tag{2.27}$$

where

$$B_n = r(0, n-1) - r(n, 2n-1),$$

$$C_n = \alpha(0, n-1)/n, \qquad D_n = \alpha(n, 2n-1)/n,$$

$$\alpha(a, b) = \sum_{j=a}^{b} \alpha_j.$$

An argument similar to the proof of (1.27) shows that

$$\sqrt{n}\, B_n \Rightarrow \sqrt{2}\, \sigma N(0, 1)/E\alpha_0. \tag{2.28}$$

Clearly

$$C_n - D_n \to 0 \quad \text{a.s.} \tag{2.29}$$

Combining (2.27)-(2.29) verifies (2.26).

3. Results and proofs. II.

THEOREM 6. *Assume that* (1.14), (1.17), *and* (1.18) *hold under* $\hat{P}$. *Then, there exists a probability* $\tilde{P}$ *such that*:

(i) $\displaystyle \frac{1}{t}\int_0^t \hat{P}\{X \circ \theta_s \in \cdot\}\, ds \to \tilde{P}\{X \in \cdot\} \quad \textit{as } t \to \infty,$

(ii) $\tilde{P}\{X \circ \theta_u \in \cdot\} = \tilde{P}\{X \in \cdot\} \quad \textit{for } u \geqq 0,$

(iii) $\displaystyle \tilde{P}\{X \in \cdot\} = \frac{1}{\hat{E}S_1}\int_0^\infty \hat{P}\{X \circ \theta_s \in \cdot\,; S_1 > s\}\, ds.$

Proof. Let f be a bounded nonnegative function on Ω. For $u \geqq 0$, set $f_u(X) = f(X \circ \theta_u)$ and observe that

$$\frac{1}{n}\int_0^{S_n} f_u(X \circ \theta_s)\, ds = \frac{1}{n}\sum_{k=0}^{n-1} h_u(X_k, X_{k+1}, \cdots),$$

where

$$h_u(X_k, X_{k+1}, \cdots) = \int_{S_k}^{S_{k+1}} f_u(X \circ \theta_s)\, ds.$$

As in the proof of Theorem 1, Birkhoff's ergodic theorem (see Lamperti (1977)) may be applied to conclude that

$$\frac{1}{n}\int_0^{S_n} f_u(X \circ \theta_s)\, ds \to \hat{E}h_u(X_0, X_1, \cdots), \tag{3.1}$$

$\hat{P}$ a.s. as $n \to \infty$. The "squeeze" argument of (2.4), applied to (3.1), can be readily adapted to show that

$$\frac{1}{t}\int_v^{v+t} f_u(X \circ \theta_s)\, ds \to \frac{1}{\hat{E}S_1}\hat{E}h_u(X_0, X_1, \cdots),$$

$\hat{P}$ a.s. as $t \to \infty$ for $v \geqq 0$, so that bounded convergence then yields

$$\frac{1}{t}\int_v^{v+t} \hat{E}f_u(X \circ \theta_s)\, ds \to \frac{1}{\hat{E}S_1}\hat{E}h_u(X_0, X_1, \cdots) \tag{3.2}$$

as $t \to \infty$. But clearly

$$\frac{1}{t}\int_u^{u+t} f_0(X \circ \theta_s)\, ds = \frac{1}{t}\int_0^t f_u(X \circ \theta_s)\, ds,$$

from which it is evident from (3.2) that

$$\hat{E}h_u(X_0, X_1, \cdots) = \hat{E}h_0(X_0, X_1, \cdots) \tag{3.3}$$

for $u \geqq 0$. Specializing f to indicator functions in (3.2) and (3.3), we conclude that

$$\frac{1}{t}\int_0^t \hat{P}\{X\circ\theta_s \in \cdot\}\, ds \to \frac{1}{\hat{E}S_1}\hat{E}h_0(X_0, X_1, \cdots)$$

as $t\to\infty$, and that

$$\hat{E}h_u(X_0, X_1, \cdots) = \hat{E}\left(\int_0^{S_1} f_u(X\circ\theta_s)\, ds\right)$$
$$= \int_0^\infty \hat{P}\{X\circ\theta_s\circ\theta_u \in \cdot;\, S_1 > s\}\, ds$$

is independent of u. These results prove the theorem.

Throughout the remainder of this section, we assume that (1.14), (1.17), and (1.18) are in force for $\hat{P}$.

THEOREM 7.

$$\tilde{P}\{r(t)\to r\} = 1. \tag{3.4}$$

Proof. Use definitions and then the fact that (1.24) holds under $\hat{P}$ to get

$$\hat{P}\{r(t)\circ\theta_s \to r;\, S_1 > s\} = \hat{P}\left\{\int_s^{s+t} f(X(u))\, du/t \to r;\, S_1 > s\right\} = \hat{P}\{S_1 > s\}. \tag{3.5}$$

Now integrate formula (3.5) with respect to s from 0 to ∞. On the right we get $\hat{E}S_1$, positive by our assumptions. On the left, use (1.13) to get $\tilde{P}[r(t)\to r]\hat{E}S_1$. Cancelling $\hat{E}S_1$ from both sides gives (3.4).

THEOREM 8. *Under* $\tilde{P}$ *formulas* (1.45) *and* (1.46) *hold.*

Proof. For any x,

$$\begin{aligned} &\hat{P}\{\sqrt{t}\,(r(t)-r)\circ\theta_s \leqq x;\, S_1 > s\} \\ &= \hat{P}\left\{(1/\sqrt{t})\int_s^{s+t} \hat{f}(X(u))\, du \leqq x \,\middle|\, S_1 > s\right\}\hat{P}\{S_1 > s\}. \end{aligned} \tag{3.6}$$

By Billingsley (1968, Thm. 20.2) the conditional probability above converges to the appropriate normal probability as $t\to\infty$, for every x and s. Integrate both sides of (3.6) with respect to s from 0 to ∞, use (1.13) to see that the left side equals $\tilde{P}\{\sqrt{t}\,(r(t)-r)\leqq x\}\hat{E}S_1$, use dominated convergence on the right to take the limit with respect to t inside the outer integral, and recall that any nonnegative random variable Z has expectation $\int_0^\infty P\{Z > t\}\, dt$. Cancel $\hat{E}S_1$ from both sides of (3.5) integrated as above to get that $\tilde{P}\{\sqrt{t}\,(r(t)-r)\leqq x\}$ converges to the appropriate normal probability. A routine calculation verifies (1.46).

Under (1.44), we can replace $\tilde{P}$ by P everywhere in this section.

4. Indirect estimation of transaction averages. Suppose a performance measure s aggregates costs for transactions moving through the system. Let transaction i arrive at time A_i and leave at time B_i. Put $D_i = B_i - A_i$. Assume that the cost associated with transaction i is $g_i(D_i)$ and that the average cost

$$s = \lim_{n\to\infty} [g_1(D_1) + \cdots + g_n(D_n)]/n \tag{4.1}$$

exists a.s.; clearly

$$s_n = [g_1(D_1) + \cdots + g_n(D_n)]/n \tag{4.2}$$

consistently estimates s. Our discussion carries over to more general transaction averages. We use s for concreteness.

In the notation of § 1, state $\tilde{X}_k$ includes the value of A_i for every transaction i in the system at time T_k and

$$f(X(t)) = \sum g_i'(t - A_i), \tag{4.3}$$

where the sum is over those transactions i for which $A_i \leqq t \leqq B_i$. Here g_i' is the right derivative of g_i and by definition $g_i'(d) = 0$ for $d < 0$ and for $d \geqq D_i$. This setup allows rather general g_i's, though not indicators.

Now call the arrival rate λ. Assuming all limits exist, $r = \lambda s$ by Heyman and Stidham (1980). Usually λ is simply the reciprocal of the expected arrival spacing. In more general cases such as batch arrivals with batch size and spacing dependent, the theory developed earlier in this paper shows how to estimate λ efficiently.

Thus, if we want to estimate r, we can do so directly via discretized observations $Y_n(f)$ as detailed earlier or indirectly via transaction observations $g_i(D_i)$. Likewise, if we want to estimate s we can do so directly or indirectly. When λ is known, we can use $\hat{\lambda} - \lambda$ as a control variate in conjunction with any of the above four possibilities. Here $\hat{\lambda}$ is a consistent estimator of λ. Based on the theory developed in this paper and in Glynn and Whitt (1985), we recommend below particular choices among the above possibilities. Let $\hat{r}$ and $\hat{s}$ be consistent estimators for r and s, respectively.

Setting 1: λ *unknown*. Glynn and Whitt (1985, §§ 8 and 11) show in a precise sense that

$$\hat{r} \approx \hat{\lambda} \hat{s} \tag{4.4}$$

for large samples. To estimate r, we can use either $\hat{r}$ or $\hat{\lambda}\hat{s}$. To estimate s, we can use either $\hat{s}$ or $\hat{r}/\hat{\lambda}$; jackknifing the latter decreases bias. In each case, our choice is based on four criteria:

1. *Asymptotic efficiency*. Abbreviate mean square error as mse. From (4.4) we get

$$\operatorname{mse}(\hat{r}) \approx \operatorname{mse}(\hat{\lambda}\hat{s}), \tag{4.5}$$

$$\operatorname{mse}(\hat{s}) \approx \operatorname{mse}\left(\frac{\hat{r}}{\hat{\lambda}}\right). \tag{4.6}$$

2. *Ease of data collection*. Formula (4.3) indicates that gathering discretized observations at state-change epochs and using our methodology is more work, but probably not much more because the evolution of the system has to be simulated in any case. Gathering discretized observations $f(X(t))$ at (reasonably short) equally-spaced intervals that do not necessarily correspond to state-change epochs may be *much* more work.

3. *Variance estimation*. Folklore has it that the sequence of transaction observations $\{g_i(D_i)\}$ is covariance stationary if and only if the sequence of discretized observations $\{Y_n(\hat{f})\}$ is covariance stationary. Thus, loosely speaking, any variance-estimation technique (e.g., via batch means, spectral analysis, autoregressive representations, functional limit theorems) applies to both or to neither. In practice, however, one has to ask at what sample sizes asymptotic results *reasonably* apply: the faster covariance falls off with increasing lag, the better. Generally, for discretized observations, covariance does drop fairly quickly. By contrast, especially in systems that are not first-in, first-out (FIFO), a series of transaction observations scrambles past, present, and future and cuts connections between time spacing and index spacing. Thus, covariance between widely-spaced observations may well be significant. This messy covariance structure is hard to handle, at least with practical sample sizes. In § 1.2 we indicated that Var $\hat{r}$ can be estimated from $\{Y_n(f), \alpha_n\}$ by standard methods. To

estimate Var $\hat{r}/\hat{\lambda}$ from $\{Y_n(f), \alpha_n\}$ and arrival data, use a straightforward analogue of methods discussed in this paper. This appears more reliable than estimating Var $\hat{s}$ from transaction data, as argued above.

4. *Bias.* The scrambling effect mentioned in point 3 above may make it harder to detect (and hence attenuate) initialization bias for transaction observations. Incompletely processed transactions cause termination bias for transaction observations, exacerbated by the "inspection paradox" discussed earlier.

Between basing estimators on discretized observations and transaction data, criterion 1 is neutral. Criterion 2 slightly favors the latter and, at least for non-FIFO systems, criteria 3 and 4 favor the former. So, on balance, we recommend $\hat{r}$ to estimate r and $\hat{r}/\hat{\lambda}$ to estimate s.

Setting 2: λ *known.* Glynn and Whitt (1985) consider two estimators of r that use optimally weighted control variates to minimize mse asymptotically. For large samples, they show that these estimators are related by

$$\hat{r}+\hat{\alpha}(\hat{\lambda}-\lambda)\approx\lambda\hat{s}+\hat{\beta}(\hat{\lambda}-\lambda). \tag{4.7}$$

Without control variates, Carson and Law (1980) show that in some cases the efficiencies of $\hat{r}$ and $\lambda\hat{s}$ differ. However, from (4.7) we get

$$\text{mse}\,[\hat{r}+\hat{\alpha}(\hat{\lambda}-\lambda)]\approx\text{mse}\,[\lambda\hat{s}+\hat{\beta}(\hat{\lambda}-\lambda)], \tag{4.8}$$

$$\text{mse}\,[\hat{s}+\lambda^{-1}\hat{\beta}(\hat{\lambda}-\lambda)]\approx\text{mse}\,[\lambda^{-1}\hat{r}+\lambda^{-1}\hat{\alpha}(\hat{\lambda}-\lambda)], \tag{4.9}$$

counterparts to (4.5) and (4.6), respectively. The rest of the discussion in setting 1 applies without change. So, on balance, we recommend $\hat{r}+\hat{\alpha}(\hat{\lambda}-\lambda)$ to estimate r and $\lambda^{-1}\hat{r}+\lambda^{-1}\hat{\alpha}(\hat{\lambda}-\lambda)$ to estimate s. It turns out that $\hat{\alpha}$ is a ratio estimator which can be constructed using randomly-spaced observations as an easy modification of the theory developed here shows.

The simplified data collection and intuitively superior variance estimation (relative to equally-spaced observations) which our framework allows significantly strengthen the case for estimators based on discretized observations.

Acknowledgment. We thank the referee for many perceptive comments on the original version of this paper.

REFERENCES

[1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
[2] P. BRATLEY, B. L. FOX AND L. E. SCHRAGE, *A Guide to Simulation*, Springer-Verlag, New York, 1983.
[3] J. S. CARSON AND A. M. LAW, *Conservation equations and variance reduction in queueing simulations*, Oper. Res., 28 (1980), pp. 535-546.
[4] P. FRANKEN, D. KÖNIG, U. ARNDT AND V. SCHMIDT, *Queues and Point Processes*, John Wiley, New York, 1982.
[5] P. W. GLYNN AND D. L. IGLEHART, *Simulation output analysis for general state space Markov chains*, Proc. of the ORSA-TIMS Special Interest Meeting on Applied Probability-Computer Science, The Interface, 1981.
[6] P. W. GLYNN AND W. WHITT, *Indirect estimation via $L=\lambda W$*, Technical Report, AT&T Bell Laboratories, Holmdel, NJ, 1985.
[7] D. P. HEYMAN AND M. J. SOBEL, *Stochastic Models in Operations Research*, Vol. 1, McGraw-Hill, New York, 1982.
[8] D. P. HEYMAN AND S. STIDHAM, JR., *The relation between customer and time averages in queues*, Oper. Res., 28 (1980), pp. 983-994.
[9] L. JOW, *An autoregressive method for simulation output analysis*, Tech. Rep. 21, Dept. of Operations Research, Stanford University, Stanford, California, 1982.

[10] J. LAMPERTI, *Stochastic Processes*, Springer-Verlag, New York, 1977.
[11] D. R. MILLER, *Existence of limits in regenerative processes*, Ann. Math. Stat., 43 (1972), pp. 1275-1282.
[12] C. STONE, *Weak convergence of stochastic processes defined on semi-infinite time intervals*, Proc. Amer. Math. Soc., 14 (1963), pp. 694-696.
[13] A. STRELLER, *A generalization of cumulative processes*, Elektron. Informationsverarb. Kybernet., 16 (1980), pp. 449-460.
[14] J. R. WILSON, *The inspection paradox in renewal-reward processes*, Oper. Res. Lett., 2 (1983), 27-30.
[15] R. WOLFF, *Poisson arrivals see time averages*, Oper. Res., 30 (1982), pp. 223-231.