S.G. Henderson and B.L. Nelson (Eds.), *Handbook in OR & MS, Vol. 13*

DOI: 10.1016/S0927-0507(06)13008-X

Chapter 8
Statistical Estimation in Computer Simulation

Christos Alexopoulos
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA
E-mail: christos@isye.gatech.edu

Abstract

A fundamental goal of a discrete-event simulation experiment is the computation of point and confidence interval estimators for various performance measures, such as means and quantiles. Data generated during a single simulation run are sample paths from serially correlated stochastic processes. This chapter presents various forms of point estimators for means and quantiles, and shows how the user can account for the dependencies in within-run data towards the computation of valid confidence intervals. Finally, this chapter reviews methods for estimating a density function using a random sample.

1 Introduction

The ultimate goal of discrete-event simulations is the computation of point estimators and confidence intervals (CIs) for a variety of system performance measures. Towards this goal simulations use the generated data during event times to compute finite-sample time averages or sample averages as approximations to unknown means. There are two types of simulations, *finite-horizon* ones and *steady-state* ones. The first type is concerned with performance measures evaluated over a specific time window, such as the mean cost incurred by an inventory policy during a month. On the other hand, a steady-state simulation studies the long-term behavior of the system of interest. A performance measure of a system is called a steady-state parameter if it is a characteristic of the equilibrium distribution of an output process. An example is the simulation of a continuously operating communication system where the objective is the computation of the mean delay of a data packet. Regardless of their type, simulations produce within-run data that are correlated, making the application of "standard" statistical techniques nontrivial.

In finite-horizon simulations systems start at a pre-determined state, and the primary issue is the number of independent replications (or runs) that need to

be made to compute estimators with a pre-specified precision. This issue can be addressed by standard statistical techniques, as we will illustrate in Section 5. On the other hand, steady-state simulations often require substantial statistical expertise. The most notable issues related to steady-state simulations are:

- Choosing the initial state for the system; the effect of this choice lessens as the run length increases.
- Choosing a *warm-up* interval that elapses before data collection can begin. The objective is to reduce the impact of the initial conditions on the finite-sample averages.
- Choosing a run length (time interval length or number of observations to be collected) to yield estimators that meet certain precision criteria.
- Choosing between a long run and several short runs – not a simple issue by any means.

Virtually all commercial software packages contain efficient routines for the recursive computation of finite-sample averages. For instance, they compute estimates for mean queue lengths, mean entity delays in queues, and server utilizations. However, they typically provide limited guidance for performing "sound" statistical analysis. The main reason is the difficulty in automating these analyses and the (often justifiable) desire of analysts to look at the final results without "wasting time" with the intricacies of the underlying statistical methodologies. This collective attitude ends up misguiding analysts and managers with regard to the usefulness of the simulation's estimates. The aforementioned issues are discussed in a variety of texts, e.g., Fishman (2001) and Law and Kelton (2000). This chapter formalizes them within a single theme and serves as a guide for several forthcoming chapters that address them in more detail.

In addition to point and interval estimates, users are often interested in estimating the density functions (or cumulative distribution functions) of random variables generated by a computer simulation. Several simulation packages can generate histograms, but such plots are often "poor" estimates of the unknown density function because their shape depends heavily on the chosen origin and the bin width. Although the statistical literature contains many state-of-the-art density estimation techniques, such as those based on kernel functions, the simulation literature (in particular texts) barely mentions such techniques, and only within the context of independent input data. This chapter aims to close the gap between the statistical and simulation literatures by reviewing univariate kernel density estimators based on independent samples and sample paths of stationary dependent processes.

Section 2 lists notation and reviews basic concepts from probability theory and statistics. Section 3 uses a simple queueing system to illustrate the properties of estimators based on sample averages and time averages. Section 4 reviews issues related to dependencies in stochastic processes and sets the stage for the methods described in Chapter 15. Section 5 applies concepts from

Section 2 to single measurements from independent replications. Section 6 discusses modern kernel density estimators.

2 Background

This section reviews tools needed to establish asymptotic (as the sample size increases) properties of estimators and to obtain confidence intervals. Let $X, X_1, X_2, \ldots$ be random variables (RVs) from a common probability space $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ is a σ-field of events and P is the probability measure. Recall that an RV on $(\Omega, \mathcal{F}, P)$ is a real-valued function $X = X(\omega)$ such that for each $x \in \mathbb{R}$, the set $\{\omega\colon X(\omega) \leqslant x\}$ is an event. The cumulative distribution function (CDF) of X is $F_X(x) = P(X \leqslant x)$, $x \in \mathbb{R}$. L^p is the space of RVs X with $\mathrm{E}(|X|^p) < \infty$.

We say that

$$X_n \to X \begin{cases} \text{almost surely (w.p.1)} & \text{if } \Pr\bigl(X_n(\omega) \to X(\omega) \\ & \qquad \text{as } n \to \infty\bigr) = 1, \\ \text{in probability} & \text{if } \lim_{n\to\infty} \Pr\bigl(|X_n - X| \leqslant \varepsilon\bigr) = 1 \\ & \quad \forall \varepsilon > 0, \\ \text{in distribution} & \text{if } \lim_{n\to\infty} F_{X_n}(x) = F_X(x) \\ & \quad \text{at all continuity points } x \text{ of } F_X, \\ \text{in } L^p,\ p \geqslant 1, & \text{if } \lim_{n\to\infty} \mathrm{E}\bigl[|X_n - X|^p\bigr] = 0. \end{cases}$$

The respective notations for these modes are $\xrightarrow{\text{a.s.}}$, $\xrightarrow{\text{p}}$, $\xrightarrow{\text{d}}$ and $\xrightarrow{L^p}$. Convergence in L^1 (L^2) is also called convergence in mean (quadratic mean). Among the first three modes, almost sure convergence is the strongest while convergence in distribution is the weakest (and easiest to establish). For additional details and related results, see Chapter 5 of Karr (1993).

Now suppose that the RVs $X_1, X_2, \ldots$ are from some distribution with an unknown parameter θ and the objective is to estimate a quantity δ that is a function of θ, say $\delta := g(\theta)$. For fixed n, let $\delta_n = \delta_n(X_1, \ldots, X_n)$ be an estimator of δ. The bias of δ_n is defined as $\mathrm{Bias}(\delta_n) := \mathrm{E}(\delta_n) - \delta$, and the mean squared error of δ_n is $\mathrm{MSE}(\delta_n) := \mathrm{E}[(\delta_n - \delta)^2] = \mathrm{Bias}^2(\delta_n) + \mathrm{Var}(\delta_n)$. If $\mathrm{E}(\delta_n) = \delta$, then δ_n is said to be *unbiased*. Furthermore, δ_n is said to be a *consistent* (respectively, *strongly consistent*) estimator of δ if $\delta_n \xrightarrow{\text{p}} \delta$ (respectively, $\delta_n \xrightarrow{\text{a.s.}} \delta$). If δ_n is unbiased for each n and $\mathrm{Var}(\delta_n) \to 0$ as $n \to \infty$, then δ_n is also consistent. This is a direct implication of Chebyshev's inequality (Karr, 1993, p. 122): $\Pr(|\delta_n - \delta| \geqslant \varepsilon) \leqslant \mathrm{E}[(\delta_n - \delta)^2]/\varepsilon^2 = \mathrm{Var}(\delta_n)/\varepsilon^2$.

The remainder of this section illustrates the aforementioned concepts with a few classical results. Suppose that $X_1, X_2, \ldots$ are independent and identically distributed (IID) RVs with finite mean μ. The sample mean $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator of μ because $\mathrm{E}(\overline{X}_n) = \mu$. $\overline{X}_n$ is also a strongly consis-

tent estimator of μ by the *strong law of large numbers* (Karr, 1993, pp. 188–189),

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu \quad \text{as } n \to \infty.$$

If $\sigma_X^2 := \text{Var}(X_1) \in (0, \infty)$, the *central limit theorem* (CLT) (see Karr, 1993, p. 174) states that

$$\frac{\overline{X}_n - \mu}{\sigma_X/\sqrt{n}} \xrightarrow{\text{d}} \text{N}(0, 1) \quad \text{as } n \to \infty,$$

where N(0, 1) is a standard normal RV. The CLT remains valid when the typically unknown variance σ_X^2 is replaced by its unbiased and consistent estimator $S_n^2(X) := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, the sample variance of the X_i. Therefore, for sufficiently large n,

$$p_{n,1-\alpha} := \Pr\left(\frac{|\overline{X}_n - \mu|}{S_n(X)/\sqrt{n}} \leqslant z_{1-\alpha/2}\right) \approx 1 - \alpha, \tag{1}$$

where z_β denotes the β-quantile of the standard normal distribution. Solving the inequality in the middle of (1) for μ, one has the well-known approximate (two-sided) $1 - \alpha$ CI for μ,

$$\overline{X}_n \pm z_{1-\alpha/2}\frac{S_n(X)}{\sqrt{n}}. \tag{2}$$

We call $p_{n,1-\alpha}$ the "coverage probability" of the CI (2). One interprets this CI as follows: Suppose that a large number of independent trials are performed; in each trial, n observations are collected and a CI for μ is computed using (2). As the number of trials grows, the proportion of CIs that contain μ approaches $1 - \alpha$.

The number of observations n required for $p_{n,1-\alpha} \approx 1 - \alpha$ depends on the symmetry of the distribution of X_i. The more skewed (asymmetric) the density/probability function of X_i, the larger n required. To reduce potential undercoverage problems ($p_{n,1-\alpha} < 1 - \alpha$) for small n, one may replace the normal quantile $z_{1-\alpha/2}$ by the larger quantile $t_{n-1,1-\alpha/2}$ of Student's t_{n-1} distribution with $n - 1$ degrees of freedom. This choice for degrees of freedom is due to the fact that for IID normally distributed X_i,

$$\frac{\overline{X}_n - \mu}{S_n(X)/\sqrt{n}} \sim t_{n-1},$$

where the notation $X \sim Y$ is used to indicate that the RVs X and Y have the same distribution.

3 Sample averages and time averages

Simulation output data are realizations (or *sample paths*) of *stochastic processes*. A stochastic process is a probabilistic model of a system that evolves

randomly. Let $X := \{X_t, t \in T\}$ be a stochastic process with state space S and time set T. Each X_t is a random element in S: a measurable mapping from an underlying probability space $(\Omega, \mathcal{F}, P)$ to the measurable space $(S, \mathcal{E})$ (all functions herein will be measurable). Typically, the state space S is a countable set, the real line $\mathbb{R}$, or a subset of a Euclidean space or a metric space; and $T = \mathbb{R}$ or $\mathbb{R}_+$ for continuous-time processes and $T = \mathbb{Z}$ or $\mathbb{Z}_+$ for discrete-time processes. We use the following simple system to illustrate the main concepts associated with data collection and analysis.

The notation $GI/G/c$ denotes a queueing system with c parallel servers and a single queue with unlimited buffer. The times between successive job arrivals are IID RVs, say $\{A_i, i \geqslant 1\}$ with finite mean $\mathrm{E}(A) := 1/\lambda$ and variance. The service times are also IID RVs, say $\{S_i, i \geqslant 1\}$ with finite mean $\mathrm{E}(S) := 1/\omega$ ($\lambda < c\omega$ for stability) and variance, and are independent of the arrival process. Let X_t denote the number of units in the system at time t, let B_t denote the number of busy servers at time t, and let W_i denote the time the ith job spends in the system. The system is formally described by the stochastic process $Y_t = (X_t, R_t^1, \ldots, R_t^c)$, where R_t^j is the remaining service time for the unit served at server j at time t. Suppose that the rule for assigning customers to servers is such that the process $\{Y_t, t \geqslant 0\}$ is *regenerative* over a sequence of epochs $\{\tau_n\}$ (see Chapter 16). A typical rule is to send an arrival to an arbitrary idle server or to the lowest numbered idle server. If the system empties out with a reasonable frequency, it is natural to let τ_n be the nth time an arrival finds the system empty. Below we consider two data collection mechanisms often encountered in practice.

Scheme I. Suppose data collection starts at some time $s \geqslant 0$ and ends at time $s + \tau$ ($\tau > 0$). This approach is convenient for continuous-time sample paths. The *time-average*

$$\overline{X}(s, s+\tau) = \frac{1}{\tau} \int_s^{s+\tau} X_t \, \mathrm{d}t \tag{3}$$

estimates the mean number of units in the system during the time interval $[s, s+\tau]$, while the time-average

$$\overline{B}(s, s+\tau) = \frac{1}{\tau} \int_s^{s+\tau} B_t \, \mathrm{d}t \tag{4}$$

estimates the (mean) number of busy servers during $[s, s+\tau]$.

The processes $\{X_t\}$ and $\{B_t\}$ have limiting distributions with finite means, denoted by L and $\nu := \lambda/\omega$ (see Serfozo, 1999). We study the properties of the estimators (3) and (4) by first fixing τ. One can show that as $s \to \infty$,

$$\overline{X}(s, s+\tau) - L \xrightarrow{\mathrm{d}} \xi(\tau) \quad \text{and} \quad \overline{B}(s, s+\tau) - \nu \xrightarrow{\mathrm{d}} \zeta(\tau),$$

where $\xi(\tau)$ and $\zeta(\tau)$ are RVs with zero mean and distribution that is independent of the system state at time s. Furthermore, as $\tau \to \infty$, $\xi(\tau)$ and $\zeta(\tau)$

converge in distribution (and hence in probability) to zero so that

$$\overline{X}(s, s+\tau) \xrightarrow{\mathrm{d}} L \quad \text{and} \quad \overline{B}(s, s+\tau) \xrightarrow{\mathrm{d}} \nu \quad \text{as } s, \tau \to \infty. \tag{5}$$

Now, under some additional mild assumptions (e.g., $\mathrm{E}(A^4) < \infty$ and $\mathrm{E}(S^4) < \infty$), one can show that for fixed s there exist finite constants $\sigma^2(X) > 0$ and $\sigma^2(B) > 0$ so that, as $\tau \to \infty$,

$$\sqrt{\tau}\big[\overline{X}(s, s+\tau) - L\big] \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, \sigma^2(X)\big) \quad \text{and}$$
$$\sqrt{\tau}\big[\overline{B}(s, s+\tau) - \nu\big] \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, \sigma^2(B)\big).$$

The last two properties allow the user to compute CIs for the respective steady-state means L and ν.

The estimation of measures other than means is also possible. For instance, the time average

$$\frac{1}{\tau}\int_s^{s+\tau} \mathbf{1}\{B_t \geqslant r\}\,\mathrm{d}t$$

is an estimator for the probability that at least r servers are busy during the time window $[s, s+\tau]$.

Scheme II. Data collection starts when unit m departs and ends when unit $m+n$ ($n \geqslant 1$) departs. Denote the respective departure times by T_m and T_{m+n}. Then the mean time a job spends in the system during the interval $[T_m, T_{m+n}]$ can be estimated by the sample average

$$\overline{W}(T_m, T_{m+n}) = \frac{1}{n}\sum_{j=m+1}^{m+n} W_j. \tag{6}$$

This estimator has properties analogous to the properties of the estimator $\overline{X}(s, s+\tau)$. Specifically, as m is held constant and $n \to \infty$, $\overline{W}(T_m, T_{m+n}) \xrightarrow{\text{a.s.}} W$, where $W = L/\lambda$ by Little's law; see Stidham (1974) and Serfozo (1999, Chapter 5). As with Scheme I, usually there exists a finite constant $\sigma^2(W) > 0$ so that, as $n \to \infty$,

$$\sqrt{n}\big[\overline{W}(T_m, T_{m+n}) - W\big] \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, \sigma^2(W)\big).$$

Remark 1. Sample averages can be collected under Scheme I, but have random denominators. For instance, let N_t be the numbers of departures by time t. The sample average

$$\overline{W}(s, s+\tau) = \frac{1}{N_{s+\tau} - N_s}\sum_{j=N_s+1}^{N_{s+\tau}} W_j \tag{7}$$

estimates the mean time a job that completed service in $[s, s+\tau]$ spent in the system. Since the denominator of Equation (7) is an RV, the derivation of the limiting properties of the estimator $\overline{W}(s, s+\tau)$ requires extra care, but they are similar to the limiting properties of its counterpart in Equation (6).

Finally, parallel properties hold for the estimator of the mean number of jobs in the system

$$\overline{X}(T_m, T_{m+n}) = \frac{1}{T_{m+n} - T_m} \int_{T_m}^{T_{m+n}} X_t \, \mathrm{d}t$$

computed under Scheme II.

4 Stationary processes

The scope of many simulation studies is the estimation of limiting measures associated with stochastic processes. This section starts with some common structural assumptions that allow us to study the impact of the dependence within a stochastic process and proceeds with limit theorems. This section contains a synopsis of material from Alexopoulos et al. (2006) and Durrett (2005).

One way to describe a stochastic process is to specify the joint distribution of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ for each set of times $t_1 < t_2 < \cdots < t_n$ and each n. This approach is typically too complicated to be attempted in practice. A simpler approach relies on the specification of the first and second moment functions of the process. These functions are the *mean* function $\mu_t := \mathrm{E}[X_t]$ and the *autocovariance* function

$$C_{t_1,t_2} := \mathrm{Cov}[X_{t_1}, X_{t_2}], \quad t_1 \leqslant t_2.$$

Notice that $C_{t_1,t_2} = C_{t_2,t_1}$. The *variance* function is then given by $\sigma_t^2 := \mathrm{Var}(X_t) = C_{t,t}$ and the *autocorrelation* function is defined as

$$\rho_{t_1,t_2} := \frac{C_{t_1,t_2}}{\sigma_{t_1}\sigma_{t_2}}, \quad t_1 \leqslant t_2.$$

(Strict) Stationarity. The stochastic process X is *stationary* if, for any $t_1, \ldots, t_k$, $t \in T$,

$$(X_{t_1+t}, \ldots, X_{t_k+t}) \stackrel{\mathrm{d}}{=} (X_{t_1}, \ldots, X_{t_k}),$$

where "$\stackrel{\mathrm{d}}{=}$" denotes equality in distribution. This says that the distribution of the process with the time origin shifted to t is equal to that of X. In other words, $\theta^t X \stackrel{\mathrm{d}}{=} X$ for any $t \in T$, where $\{\theta^t, t \in T\}$ is the family of time-shift transformations defined by $\theta^t x = (x_{s+t}, s \in T)$. An immediate result is that the joint distribution of $X_{t_1}, X_{t_2}, \ldots, X_{t_k}$ depends only on the distances between $t_1, t_2, \ldots, t_k$.

Weak stationarity. The process X is said to be *weakly stationary* if its mean and variance functions are constant and its autocovariance function satisfies

$$\mathrm{Cov}[X_s, X_{s+t}] = C_t, \quad s \geqslant 0, t \geqslant 0,$$

that is, it depends only on the *lag* t. If the process X is Gaussian whence all finite-dimensional distributions are multivariate normal, then weak stationarity implies stationarity.

Ergodicity. A stationary process X is *ergodic* if $P\{X \in A\} = 0$ or 1, for each $A \in \mathcal{E}$ that is *time-shift invariant*: $\{X \in A\} = \{\theta^t X \in A\}$, $t \in T$. This property is important for steady-state simulations because it allows the estimation of limiting performance measures based on a single realization (sample path) of the process X. Examples are stationary Markov chains and stationary functionals of a stationary, ergodic process.

Example 1 (The $M/M/1$ queueing system). This is a special case of the $GI/G/c$ system with one server and exponentially distributed interarrival and service times. The ratio $\nu = \lambda/\omega$ is the traffic intensity or (long-run) server utilization. Suppose that the service discipline is first-come, first-served. Let D_i be the delay time in queue of the ith customer and assume that the system starts empty. Since $D_1 = 0$, we have $\mathrm{E}(D_1) = 0$. However, the first ($i = 1$) of the following recursive equations (Lindley, 1952)

$$D_{i+1} = \max\{D_i + S_i - A_{i+1}, 0\}, \quad i \geqslant 1, \tag{8}$$

implies $\Pr(D_2 > 0) = \Pr(S_1 > A_2) = \lambda/(\lambda + \omega) > 0$; hence $\mathrm{E}(D_2) > 0$. Therefore the delay process $\{D_i, i \geqslant 1\}$ is not stationary. Using queueing theory (Gross and Harris, 1998, Chapter 2), one has

$$\lim_{i\to\infty} \Pr(D_i \leqslant x) = 1 - \nu + \nu\bigl(1 - \mathrm{e}^{-(\omega-\lambda)x}\bigr), \quad x \geqslant 0, \tag{9}$$

$$\mu = \lim_{i\to\infty} \mathrm{E}(D_i) = \frac{\nu}{(1-\nu)\omega} \quad \text{and} \quad \lim_{i\to\infty} \mathrm{Var}(D_i) = \frac{\nu(2-\nu)}{\omega^2(1-\nu)^2}.$$

Equation (9) suggests that the delay process becomes asymptotically stationary. Indeed, if D_1 has the distribution on the right-hand side of (9), Equation (8) implies (after some work) that all D_i have the same distribution and the delay process is stationary.

The autocorrelation function of the delay process $\{D_i\}$ is given by (Blomqvist, 1967)

$$\rho_j = \frac{(1-\nu)^3(1+\nu)}{(2-\nu)\nu^3} \times \sum_{k=j+3}^{\infty} \left[\frac{\nu}{(\nu+2)^2}\right]^k \frac{(2k-3)!}{k!(k-2)!}(k-j-1)(k-j-2)$$

for $j = 0, 1, \ldots$. This function is monotone decreasing with a very long tail that increases as the server utilization ν increases (for instance, $\rho_{200} \approx 0.30$ when $\nu = 0.9$). This makes the $M/M/1$ system a good test bed for evaluating simulation methodologies.

Example 2 (Moving average process). A well-studied stationary sequence is the moving average process of order q (often abbreviated to MA(q))

$$X_i = \beta_0 Z_i + \beta_1 Z_{i-1} + \cdots + \beta_q Z_{i-q}, \quad i \geqslant 0,$$

where the coefficients β_i are constants and $\{Z_i, i \in \mathbb{Z}\}$ is an IID random sequence with mean zero and finite variance σ_Z^2. MA processes have applications in several areas, particularly econometrics (Chatfield, 1989).

Clearly,

$$\mathrm{E}(X_i) = 0, \qquad \mathrm{Var}(X_i) = \sigma_Z^2 \sum_{i=0}^{q} \beta_i^2$$

while some algebra yields the autocovariace function

$$C_j = \begin{cases} \sigma_Z^2 \sum_{i=0}^{q-j} \beta_i \beta_{i+j} & \text{for } j = 0, 1, \ldots, q, \\ 0 & \text{for } j > q, \end{cases}$$

which "cuts off" at lag q.

4.1 Impact of dependence

Although this topic will be the subject of Chapter 15, we will give a brief introduction of the main issues. Suppose one observes the portion $X_1, \ldots, X_n$ of a discrete-time stationary process for the purpose of estimating the mean $\mu := \mathrm{E}(X_1)$. Clearly, $\overline{X}_n$ is an unbiased estimator of μ while some algebra yields

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma_X^2}{n}\left[1 + 2\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\rho_j\right] := \frac{\sigma_X^2}{n}(1 + \gamma_n). \tag{10}$$

In order for $\overline{X}_n$ to be a consistent estimator of μ, we require that $\lim_{n\to\infty} \mathrm{Var}(\overline{X}_n) = 0$. The last condition holds if $\lim_{n\to\infty} n\,\mathrm{Var}(\overline{X}_n) < \infty$ or, equivalently,

$$\lim_{n\to\infty} \gamma_n < \infty. \tag{11}$$

The condition $\lim_{j\to\infty} C_j = 0$ is necessary for (11) but not sufficient. A necessary and sufficient condition is

$$\sum_{j=-\infty}^{\infty} |C_j| < \infty. \tag{12}$$

In simple terms, the covariance between X_i and X_{i+j} must dissipate sufficiently fast for the summation in (12) to remain bounded. If condition (12) holds, we call

$$\sigma^2 := \sum_{j=-\infty}^{\infty} C_j \tag{13}$$

the *variance parameter* of the process X.

Under the assumption that $\overline{X}_n$ is approximately normally distributed (which is reasonable for sufficiently large n), the construction of a CI for μ requires the derivation of an estimator for $\mathrm{Var}(\overline{X}_n)$. Based on the practice for IID data, one might be inclined to estimate $\mathrm{Var}(\overline{X}_n)$ by $S_n^2(X)/n$. Is this wise? After some algebra one can establish the following formula:

$$\begin{aligned} \mathrm{E}\left[\frac{S_n^2(X)}{n}\right] &= \frac{\sigma_X^2}{n}\left[1 - \frac{2}{n-1}\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\rho_j\right] \\ &= \frac{n/(1+\gamma_n) - 1}{n-1}\,\mathrm{Var}(\overline{X}_n). \end{aligned} \tag{14}$$

For processes that are positively correlated (i.e., $\rho_i > 0\ \forall i$), Equation (14) implies that $\mathrm{E}[S_n^2(X)/n] \ll \mathrm{Var}(\overline{X}_n)$. Hence the CI in (2) can exhibit severe undercoverage.

4.2 Ergodic and central limit theorems for stationary sequences

The strong law of large numbers and the CLT for sums of IID random variables have analogues for partial sums of stationary sequences. Our discussion of these will involve the notion of ergodicity. The following ergodic theorem is in Durrett (2005).

Theorem 1 (Ergodic theorem: von Neumann, Birkhoff). *Suppose that* $\{X_i\}$ *is a stationary, ergodic sequence of random variables with* $\mathrm{E}(|X_1|) < \infty$. *Then*

$$\overline{X}_n \xrightarrow{a.s.} \mu \quad \text{as } n \to \infty. \tag{15}$$

Moreover, if $\mathrm{E}(|X_1|^p) < \infty$ *for* $p \geqslant 1$, *then*

$$\overline{X}_n \xrightarrow{L^p} \mu \quad \text{as } n \to \infty. \tag{16}$$

These w.p.1 and L^p limit properties also hold for stationary processes that are not ergodic; the $\mathrm{E}(X_1)$ is just replaced by the conditional expectation of X_1 with respect to the σ-field of invariant events. The ergodic theorem also holds for averages of continuous-time processes, in which case

$$\frac{1}{t}\int_0^t X_s\,\mathrm{d}s \xrightarrow{\text{a.s.}} \mathrm{E}(X_0) \quad \text{as } t \to \infty.$$

There are several functional central limit theorems for stationary processes under various assumptions. Two prominent ones are as follows (Durrett, 2005). Suppose that $\{X_i, i \in \mathbb{Z}\}$ is a stationary, ergodic sequence of random variables with mean μ and finite variance. The assumption that the process has an infinite past is natural and is justified since a one-sided process with time set $\mathbb{Z}_+$ can be extended by (by Kolmogorov's extension theorem) to a two-sided process.

Define the partial sums $S_n = \sum_{i=1}^{n}(X_i - \mu)$, and consider the related stochastic process

$$n^{-1/2}S_{(nt)} = n^{-1/2}\sum_{i=1}^{\lfloor nt \rfloor}(X_i - \mu), \quad t \in \mathbb{R}_+.$$

Here $\{n^{-1/2}S_{(n\cdot)}, n \geqslant 1\}$ is a sequence of random elements in the Skorohod space $D(\mathbb{R}_+)$ of real-valued functions on $\mathbb{R}_+$ that are right-continuous with left-hand limits. We use "$\Longrightarrow$" to denote weak convergence of random elements in this space.

Theorem 2. *In addition to the assumptions above, suppose that*

$$\sum_{n=1}^{\infty}\left[\mathrm{E}\left(\xi_n^2\right)\right]^{1/2} < \infty, \tag{17}$$

where $\xi_n = \mathrm{E}[X_0 - \mu | X_k - \mu\colon k \leqslant -n]$. *Then the series* (13) *converges absolutely and*

$$n^{1/2}S_{(n\cdot)} \Longrightarrow \sigma\mathcal{W},$$

where $\{\mathcal{W}(t), t \geqslant 0\}$ *denotes a standard Brownian motion process.*

The following is another functional central limit theorem with a stronger assumption. This uses the notion that the sequence X is *ϕ-mixing* if there are $\varphi_j \downarrow 0$ such that, for each $j \geqslant 1$,

$$\left|\Pr(A \cap B) - \Pr(A)\Pr(B)\right| \leqslant \phi_j \Pr(A), \quad A \in \mathcal{F}_{-\infty}^{i}, B \in \mathcal{F}_{i+j}^{\infty},$$

where $\mathcal{F}_i^j$, $i \leqslant j$, denotes the σ-field generated by $X_i, X_{i+1}, \ldots, X_j$.

Theorem 3. *The assertion of Theorem 2 is true if assumption* (17) *is replaced by the assumption that X is ϕ-mixing and* $\sum_{n=1}^{\infty}\phi_n^{1/2} < \infty$.

Assuming $\sigma > 0$, Theorems 2 and 3 imply

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1) \quad \text{as } n \to \infty. \tag{18}$$

Equation (18) can be used to obtain an asymptotically (as $n \to \infty$) CI for μ. The estimation of σ^2 is the main topic of Chapter 15.

Before we proceed, we present an even stronger property to be revisited in Section 6.7. The process X is *strongly* α-mixing if there are $\alpha_j \downarrow 0$ such that, for each $j \geqslant 1$,

$$\left|\Pr(A \cap B) - \Pr(A)\Pr(B)\right| \leqslant \alpha_j, \quad A \in \mathcal{F}_{-\infty}^{i}, B \in \mathcal{F}_{i+j}^{\infty}.$$

Since $\alpha_j \leqslant \phi_j$, a ϕ-mixing process is also strongly mixing. Informally, mixing means that events in the distant future are approximately independent of those in the past. For details see Billingsley (1968) and Rosenblatt (1956).

Remark 2. Contrary to popular belief, many stochastic processes encountered in simulation output analysis are *not* ϕ-mixing. Examples are autoregressive processes, regenerative processes (see Chapter 16) with regenerations not occurring uniformly fast over the state space, and virtually all open queueing networks. However, positive recurrent regenerative processes are strongly mixing. For details, see Glynn and Iglehart (1985).

Example 3 (Gaussian first-order autoregressive process). Another well-known stationary process is the autoregressive process of order one, denoted by AR(1), and often called the Markov process in the time series literature,

$$X_i = \mu + \rho(X_{i-1} - \mu) + Z_i, \quad i \geqslant 1,$$

where $|\rho| < 1$, $X_0 \sim \mathrm{N}(\mu, 1)$, and the Z_i are IID $\mathrm{N}(0, 1 - \rho^2)$.

The autocorrelation function of this process $\rho_j = \rho^j$, $j \geqslant 0$, is monotone decreasing if $\rho > 0$ with a tail that becomes longer as ρ increases, and exhibits damped harmonic behavior around the zero axis if $\rho < 0$. Applying Equation (10) one has

$$n \operatorname{Var}(\overline{X}_n) = 1 + 2\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\rho^j \longrightarrow \frac{1+\rho}{1-\rho} \quad \text{as } n \to \infty.$$

Hence $\overline{X}_n$ is a consistent estimator of the mean $\mu = \mathrm{E}(X_i)$ and $\sigma^2 = (1+\rho)/(1-\rho)$. This process does not possess the ϕ-mixing property, but the CLT (18) holds by Theorem 2.

5 Analyzing data from independent replications

Suppose that a simulation run yields n output observations, say $X_1, \ldots, X_n$. Let $Y = f(X_1, \ldots, X_n)$ be a measurable function, and assume that the objective of the study is the estimation of $\mu := \mathrm{E}(Y)$. For instance, if X_i is the time spent in system by customer i, then $f(X_1, \ldots, X_n)$ could be the average time

in system experienced by the first n customers or the maximum time in system. By definition, Y is an unbiased estimator for μ.

To bypass the impact of the autocorrelations, one can run k independent replications of the system simulation. Each replication starts in the same state and uses a portion of the random number stream that is different from the portions used to run the other replications. Assume that replication i produces the output data $X_{i,1}, X_{i,2}, \ldots, X_{i,n}$ and $Y_i = f(X_{i,1}, X_{i,2}, \ldots, X_{i,n})$. Since the RVs Y_i are IID, their sample mean $\overline{Y}_k$ is also an unbiased estimator of μ, and their sample variance $S_k^2(Y)$ is an unbiased estimator of $\mathrm{Var}(Y)$. If, in addition, k is sufficiently large, an approximate $1-\alpha$ CI for μ is

$$\overline{Y}_k \pm t_{k-1,1-\alpha/2}\frac{S_k(Y)}{\sqrt{k}}. \tag{19}$$

Denote the half-width of the interval (19) by $\delta(k,\alpha) = t_{k-1,1-\alpha/2}S_k(Y)/\sqrt{k}$.

5.1 Sequential estimation

Suppose that one wishes to estimate μ within a tolerance $\pm d$, where d is user-specified. More formally, one would like to make k runs so that

$$\Pr(\overline{Y}_k - d \leqslant \mu \leqslant \overline{Y}_k + d) \geqslant 1-\alpha, \tag{20}$$

where $\alpha \in (0,1)$. The sequential procedure of Chow and Robbins (1965) (see also Nadas, 1969) is to run one replication at a time and stop at run k^* such that

$$k^* = \min\left[k\colon\ k \geqslant 2,\ \delta(k,\alpha) \leqslant \sqrt{\frac{k}{k-1}d^2 - \frac{t_{k-1,\alpha/2}^2}{k(k-1)}}\right]. \tag{21}$$

The stopping rule (21) is based on the limiting result

$$\lim_{d\to 0}\Pr(\overline{Y}_{k^*} - d \leqslant \mu \leqslant \overline{Y}_{k^*} + d) = 1-\alpha. \tag{22}$$

Equation (22) indicates that as the tolerance d decreases, the probability that the interval $\overline{Y}_{k^*} \pm d$ contains μ converges to $1-\alpha$. Notice that as k increases, the right-hand side of the last inequality in (21) approaches d.

Now suppose that $Y_1, Y_2, \ldots$ are normally distributed. Starr (1966) showed that the choice

$$k^* = \min\bigl[k\colon\ k \geqslant 3,\ k \text{ odd},\ \delta(k,\alpha) \leqslant d\bigr]$$

yields

$$\Pr(\overline{Y}_{k^*} - d \leqslant \mu \leqslant \overline{Y}_{k^*} + d) \geqslant \begin{cases} 0.928 & \text{if } \alpha = 0.05, \\ 0.985 & \text{if } \alpha = 0.01. \end{cases}$$

The last inequalities indicate little loss in the confidence level for arbitrary d. Based on Starr's result and Equation (22), we recommend the simpler and

more intuitive stopping rule that starts with at least 5 runs,

$$k^* = \min\bigl[k\colon\ k \geqslant 5,\ \delta(k,\alpha) \leqslant d\bigr].$$

Often, the t quantile is replaced by the normal quantile $z_{1-\alpha/2}$.

An alternative *two-stage* approach for computing a CI for μ with half-width at most d works as follows: The first stage uses k_0 replications to compute a variance estimate $S^2_{k_0}(Y)$ and a CI with half-width $\delta(k_0,\alpha)$. Assume that the estimate $S^2_{k_0}(Y)$ does not change significantly as k_0 increases. If $\delta(k_0,\alpha) \leqslant d$, the procedure terminates. Otherwise, an estimate of the total number of replications required to obtain a half-width of at most d is computed from

$$\begin{aligned}\hat{k} &= \min\left[k\colon\ k \geqslant k_0,\ \frac{t_{k_0-1,1-\alpha/2} S_{k_0}(Y)}{\sqrt{k}} \leqslant d\right] \\ &= \max\left\{k_0,\ \left\lceil \frac{t^2_{k_0-1,1-\alpha/2} S_{k_0}(Y)^2}{d^2} \right\rceil\right\},\end{aligned}$$

where $\lceil\cdot\rceil$ is the ceiling function. The efficacy of this method depends on the closeness of $S^2_{k_0}(Y)$ to the unknown $\mathrm{Var}(Y)$. If $S^2_{k_0}(Y)$ underestimates $\mathrm{Var}(Y)$, then $\hat{k}$ will be smaller than actually needed. Conversely, if $S^2_{k_0}(Y)$ overestimates $\mathrm{Var}(Y)$, then unnecessary replications will have to be made.

Alexopoulos and Seila (1998) studied the performance of the sequential and two-stage methods by a set of experiments concerning the throughput of an $M/M/1$ system. Based on their experiments, the sequential procedure with an initial sample of at least 5 replications appears to outperform the two-stage procedure as: (a) the resulting CI half-width is always less than or equal to the target value; (b) the variation in the final sample sizes and CI half-widths is substantially smaller; and (c) the CIs had estimated coverage close to the nominal probability.

An alternative problem is the computation of an estimate for μ with relative error $|\overline{Y}_k - \mu|/|\mu| \leqslant c$, where c is a positive constant. Formally, one requests

$$\Pr\left(\frac{|\overline{Y}_k - \mu|}{|\mu|} \leqslant c\right) \geqslant 1-\alpha.$$

Let $c' = c/(1+c)$ and assume that we perform one replication at a time until the relative half-width $\delta(k,\alpha)/|\overline{Y}_k| \leqslant c'$. Since $\delta(k,\alpha)/|\overline{Y}_k|$ estimates the actual relative error, we have

$$\begin{aligned}1-\alpha &\approx \Pr\left(\frac{|\overline{Y}_k - \mu|}{|\overline{Y}_k|} \leqslant \frac{\delta(k,\alpha)}{|\overline{Y}_k|}\right) \\ &\leqslant \Pr\bigl(|\overline{Y}_k - \mu| \leqslant c'|\overline{Y}_k|\bigr) \\ &= \Pr\bigl(|\overline{Y}_k - \mu| \leqslant c'|\overline{Y}_k - \mu + \mu|\bigr) \\ &\leqslant \Pr\bigl(|\overline{Y}_k - \mu| \leqslant c'|\overline{Y}_k - \mu| + c'|\mu|\bigr)\end{aligned}$$

$$= \Pr\bigl[(1-c')|\overline{Y}_k - \mu| \leqslant c'|\mu|\bigr]$$
$$= \Pr\left(\frac{|\overline{Y}_k - \mu|}{|\mu|} \leqslant \frac{c'}{1-c'}\right)$$
$$= \Pr\left(\frac{|\overline{Y}_k - \mu|}{|\mu|} \leqslant c\right).$$

Based on these observations, one can use the following stopping rule:

$$k^* = \min\left[k\colon\ k \geqslant k_0,\ \frac{\delta(k,\alpha)}{|\overline{Y}_k|} \leqslant c'\right]. \tag{23}$$

Law et al. (1981) showed that when c is close to 0, the coverage of the CI $\overline{Y}_k \pm \delta(k,\alpha)$ can be arbitrarily close to $1-\alpha$. They recommend that rule (23) be used with $c \leqslant 0.15$ and $k_0 \geqslant 10$.

5.2 *Quantile estimation*

The method of replications can also be used to implement nonparametric methods for estimating performance measures other than means. For example, let Y be the total cost incurred by an inventory system during a month. Suppose we wish to estimate the p-quantile $y_p = F_Y^{-1}(p) := \inf\{y \in \mathbb{R}\colon\ F_Y(y) \geqslant p\}$, $0 < p < 1$, of Y. If the cumulative distribution function (CDF) $F_Y(\cdot)$ is continuous and strictly monotone, then y_p is the unique solution to the equation $F_Y(y) = p$.

Let $Y_1, \ldots, Y_k$ be a random sample from a continuous CDF F_Y obtained by performing k independent replications, and let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(k)}$ be the order statistics corresponding to the Y_i's (without loss of generality we assume that the Y_i are distinct). Then a point estimator for y_p is

$$\hat{y}_p = \begin{cases} Y_{(kp)}, & \text{if } kp \text{ is integer,} \\ Y_{(\lfloor kp+1 \rfloor)}, & \text{otherwise,} \end{cases}$$

where $\lfloor\cdot\rfloor$ is the floor function.

To compute a $1-\alpha$ CI for y_p, one identifies indices $i < j$ such that $\Pr(Y_{(i)} < y_p < Y_{(j)}) \geqslant 1-\alpha$. Then $(Y_{(i)}, Y_{(j)})$ is the required interval. The event $Y_{(i)} < y_p < Y_{(j)}$ has the binomial probability

$$\Pr(Y_{(i)} < y_p < Y_{(j)}) = \sum_{\ell=i}^{j-1} \binom{k}{\ell} p^\ell (1-p)^{k-\ell}$$
$$\approx \Phi\left(\frac{j-1-kp}{\sqrt{kp(1-p)}}\right) - \Phi\left(\frac{i-1-kp}{\sqrt{kp(1-p)}}\right),$$

where the normal approximation is recommended for $kp \geqslant 5$ (Hogg and Craig, 1995). Since several index pairs can satisfy the last inequality, one would

choose a symmetric range of indices. In this case, the indices would be

$$i = \lfloor kp + 0.5 - z_{1-\alpha/2}\sqrt{kp(1-p)} \rfloor \quad \text{and}$$
$$j = \lceil kp - 0.5 + z_{1-\alpha/2}\sqrt{kp(1-p)} \rceil.$$

(The last expressions involve continuity corrections.) It should be noted that quantile estimation is usually more difficult than estimation of the mean because point estimates for quantiles are biased and substantially larger sample sizes are required to obtain reasonably tight CIs. These problems are much more severe for more extreme quantiles, i.e., for p close to 0 or 1. Exceptions are heavy-tailed distributions (Gross et al., 2002). An introduction to nonparametric interval estimation methods is given in Hogg and Craig (1995).

5.3 Using independent replications to estimate steady-state measures

The method of independent replications can also be used for estimating the steady-state mean μ of an output process. Unfortunately, most simulations start at an initial state that is either fixed (e.g., the empty and idle state in a queueing system) or is randomly chosen from a distribution other than the steady-state one. As a result, the process X goes through an initial phase before reaching steady state. How can one deal with the initial transient phase? The simulation folklore suggests that, to bypass trouble arising from this portion of a run, the experimenter should simply allow the simulation to "warm up" before retaining data. Then the question is, how much truncation (deletion) on every replication is enough to produce unbiased, low-variance point estimators and valid CIs?

Suppose that the output data form a discrete-time process. Several heuristic methods try to identify an appropriate index ℓ and then truncate the first ℓ observations. For instance, Welch (1983) gives a graphical truncation method that has met some success; see Alexopoulos and Seila (1998) or Law and Kelton (2000) for a description and illustrative examples. A host of other heuristic methods are somewhat naive and ill-suited for use in practical problems, and so will not be discussed here. References for various statistics-based tests are listed in Alexopoulos and Seila (1998) and Law and Kelton (2000); they too are not completely trustworthy in some situations, and so we will dispense with any additional details for now. For these and other reasons, Alexopoulos and Seila (1998) state that it is difficult and problematic to select a proper truncation index, especially for congested queueing systems with slowly decreasing autocorrelation functions.

Since the initial conditions induce a systematic error that has a more-pronounced impact when multiple independent replications start at the same state, it is of interest to see what happens if we indeed incorporate a truncation strategy into the CI for the steady-state mean μ. Assuming that one still uses k independent replications with n observations per replication (after truncation) to compute a CI for μ, the replicate means are $Y_i(\ell) = n^{-1} \sum_{j=\ell+1}^{\ell+n} X_{i,j}$,

$i = 1, 2, \ldots, k$, the point estimator for μ is

$$\hat{\mu}_{\mathrm{R}} = \overline{Y}_k(\ell) = \frac{1}{k}\sum_{i=1}^{k} Y_i(\ell),$$

and the estimator for the variance parameter σ^2 is

$$\widehat{\sigma^2_{\mathrm{R}}} = nS_k^2\big(Y(\ell)\big) = \frac{n}{k-1}\sum_{i=1}^{k}\big[Y_i(\ell) - \overline{Y}_k(\ell)\big]^2.$$

Glynn and Heidelberger (1991) study truncation schemes for processes whose bias satisfies the condition

$$\sum_{j=1}^{\infty}\big|\mathrm{E}(X_j) - \mu\big| < \infty. \tag{24}$$

They propose rules for choosing the truncation index ℓ so that the replicate truncated mean $\overline{Y}_k(\ell)$ satisfies a central limit theorem. Two cases that satisfy condition (24) are $\mathrm{E}(X_j) - \mu = \mathrm{O}(1/j^r)$, for some constant $r > 1$, and $\mathrm{E}(X_j) - \mu = \mathrm{O}(\beta^j)$, for some constant $\beta \in [0, 1)$. The "big-oh" notation $\mathrm{O}(h(u))$ denotes a function $g(u)$ for which there exist constants c and u_0 such that $|g(u)| \leqslant c|h(u)|$ for all $u \geqslant u_0$.

Fishman (2001) considers the common practice of fixing the replication length $\ell + n$ (after a truncation index has been identified) and increasing the number of replications. Theorem 4 describes conditions for assuring proper coverage for the resulting CI. Assumptions (F.1)–(F.3), which feed into Theorem 4 as sufficient conditions, are satisfied by a variety of processes encountered in simulation output analysis, e.g., the AR(1) process, irreducible aperiodic finite-state-space Markov chains, and irreducible finite-state-space Markov processes. S_0 denotes the (possibly random) initial state of the underlying system.

(F.1) There exists a constant $\beta \in [0, 1)$ such that for any initial state s_0,

$$\mu_j(s_0) := \mathrm{E}(X_j|S_0 = s_0) = \mu + \mathrm{O}\big(\beta^j\big) \quad \text{as } j \to \infty.$$

This implies that $\mathrm{E}(Y_1(\ell)|S_0 = s_0) = \mu + \mathrm{O}(\beta^\ell/n)$ as $\ell \to \infty$ and $n \to \infty$.

(F.2) For each s_0 and n, there is a constant $\sigma_n^2 \in (0, \infty)$ such that

$$\begin{aligned}\sigma_n^2(\ell; s_0) &:= n\,\mathrm{Var}\big(Y_1(\ell)|S_0 = s_0\big)\\ &= \sigma_n^2 + \mathrm{O}\left(\frac{\beta^\ell}{n}\right) \quad \text{as } \ell \to \infty \text{ and } n \to \infty.\end{aligned} \tag{25}$$

Equation (25) implies $\sigma_n^2(\ell; s_0) \to \sigma_n^2$ as $\ell \to \infty$, independently of s_0.

(F.3) There is a constant $\sigma^2 \in (0, \infty)$ such that

$$\sigma_n^2 = \sigma^2 + \mathrm{O}\left(\frac{1}{n}\right) \quad \text{as } n \to \infty.$$

Hence for all s_0 and ℓ, $\sigma_n^2(\ell; s_0) \to \sigma^2$, as $n \to \infty$.

Theorem 4 (Fishman, 2001, Section 6.4). *If assumptions* (F.1)–(F.3) *hold*, *n is fixed*, *and* $\ell/\ln(k) \to \infty$ *as* $k \to \infty$ *and* $\ell \to \infty$, *then*

$$\frac{\overline{Y}_k(\ell) - \mu}{S_k(Y(\ell))/\sqrt{k}} \xrightarrow{d} \mathrm{N}(0, 1).$$

We see that the CI for μ will be of the form

$$\overline{Y}_k(\ell) \pm t_{1-\alpha/2,k-1} \frac{S_k(Y(\ell))}{\sqrt{k}}.$$

Further, if one fixes the number of observations n retained during a replication and attempts to compute a narrower CI by making more runs k, the theorem says that we can achieve the nominal coverage $1 - \alpha$ by forcing the truncation index ℓ to grow faster than the logarithm of k. This requirement is hard to implement in practice as the user has to consider three parameters (n, k and ℓ). This "systematic error" due to the initial transient is not as pertinent to the methods presented in Chapter 15 (e.g., the batch means method and the standardized time series method) and Chapter 16 (the regenerative method).

We finish this discussion by pointing out that Fishman (2001) and Glynn and Heidelberger (1991) give several additional sufficient conditions among n, k and ℓ that yield a central limit theorem for the grand replicate mean $\overline{Y}_k(\ell)$. For additional results and experiments related to this issue, see Alexopoulos and Goldsman (2004).

6 Density estimation

The probability density function is a fundamental concept in probability and statistics. Suppose that we possess a finite univariate sample, say $X_1, \ldots, X_n$ from an unknown density function f. *Density estimation* refers to the construction of an estimate of f using the observed data. In simulation experiments this need arises in input data analysis, in analysis of data from independent replications, and in the estimation of the marginal density function of a stationary process. We start with estimators for IID data; the case that has received most attention.

If one assumes that the data come from a parametric distribution, then the parameters can be estimated by "standard" methods, such as the method of maximum likelihood, and the goodness-of-fit can be assessed by a battery of tests (e.g., the omnibus chi-square test, the Kolmogorov–Smirnov test and

the Anderson–Darling test). Unfortunately, this approach is quite restrictive because the commonly-used parametric families have unimodal density functions. Below we give a brief introduction to *nonparametric* methods for estimating the unknown density. In this case, we will allow the data themselves to guide the estimation of f. For a thorough treatment of this subject, we refer the eager reader to Härdle (1991), Silverman (1986), and many references therein. Multivariate density estimation is treated thoroughly by Scott (1992).

6.1 Histograms

The oldest and most commonly-used density estimator is the histogram. Given an origin x_0 and a bin width h, the bins are the intervals $[x_0 + \ell h, x_0 + (\ell + 1)h)$, for $\ell = 0, \pm 1, \pm 2, \ldots$. Notice that the leftmost interval can be open and the rightmost interval can be closed. The histogram is then defined by

$$\hat{f}_{\mathrm{H}}(x) = \frac{1}{nh}(\text{number of } X_i \text{ in the same bin as } x), \quad x \in \mathbb{R}.$$

The choice of the bin width h controls the smoothness of the histogram. Unfortunately, the choice of the origin and the bin width can have a severe impact on the histogram. Clearly, more robust methods are required. We start with the naive estimator, and proceed with various state-of-the-art estimators.

6.2 The naive estimator

By definition, the density function of an absolutely continuous random variable X obeys

$$f(x) = \lim_{h \to 0} \frac{1}{2h} \Pr(x - h < X < x + h), \quad x \in \mathbb{R}.$$

For given h, we can approximate the $\Pr(x - h < X < x + h)$ by the fraction of the sample that falls in the interval $(x - h, x + h)$. Then the *naive* estimator of f is

$$\hat{f}_{\mathrm{N}}(x) = \frac{1}{2nh}\big[\text{number of } X_i \text{ in } (x - h, x + h)\big], \quad x \in \mathbb{R}. \tag{26}$$

Notice that the naive estimator is neither continuous nor differentiable – it has jumps at the points $X_i \pm h$ and has zero derivative everywhere else.

To motivate the estimators in the following subsections, we write (26) as

$$\hat{f}_{\mathrm{N}}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right),$$

where the weight function,

$$w(x) = \begin{cases} \frac{1}{2}, & \text{if } |x| < 1, \\ 0, & \text{otherwise}, \end{cases} \tag{27}$$

is the uniform density in the interval $(-1, 1)$.

6.3 The kernel estimator

The naive estimator in Equations (26) and (27) can be generalized by replacing the uniform weight by a *kernel* function $K(\cdot)$ obeying

$$\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1, \qquad \sup_x \big|K(x)\big| < \infty, \quad \text{and} \quad \lim_{|x|\to\infty} \big|xK(x)\big| = 0. \tag{28}$$

Most often, the kernel K is an even and symmetric density, e.g., the normal density. The general form of the kernel estimators is

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{29}$$

and h is frequently called the *smoothing parameter* or the *bandwidth*. The kernel estimator inherits the properties of the kernel related to continuity and differentiability: if the kernel K is nonnegative (i.e., a density function), then $\hat{f}_K$ is also a probability density function as it integrates to one.

Clearly, the kernel estimator is the sum of "bumps" $K\{(x - X_i)/h\}/(nh)$ placed at the observations X_i. Although the function K determines the shape of the bumps, the bandwidth h affects the smoothness of the estimator. If h is chosen too small, the estimate becomes overly spurious. A few of the most common kernels are listed in Table 1 ($\mathbf{1}(\cdot)$ is the indicator function).

6.4 Bandwidth selection

Although the bandwidth h plays an important role, there are no firm rules for selecting it. A common approach relies on the minimization of the integrated MSE of the kernel estimator $\hat{f}_K$, namely,

$$\mathrm{IMSE}(h) = \int_{-\infty}^{\infty} \mathrm{E}\big[\big(\hat{f}_K(x; h) - f(x)\big)^2\big]\,\mathrm{d}x.$$

Assuming that the kernel is a symmetric density and that the density f has continuous derivatives of all orders needed (Silverman, 1986, Section 3.3), the

Table 1.
Common kernels

Name	$K(x)$
Triangular	$(1 - \lvert x\rvert)\mathbf{1}(\lvert x\rvert < 1)$
Epanechnikov	$\frac{3}{4}(1 - x^2)\mathbf{1}(\lvert x\rvert < 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$
Quartic	$\frac{15}{16}(1 - x^2)^2\mathbf{1}(\lvert x\rvert < 1)$
Triweight	$\frac{35}{32}(1 - x^2)^3\mathbf{1}(\lvert x\rvert < 1)$
Cosine	$\frac{\pi}{4}\cos(\pi x/2)\mathbf{1}(\lvert x\rvert < 1)$

IMSE is asymptotically (as $n \to \infty$) approximated by the asymptotic mean integrated squared error (AMISE)

$$\mathrm{AMISE}(h) = \frac{R(K)}{nh} + \frac{h^4 R(f'')\kappa^4}{4}, \tag{30}$$

where $R(f'') = \int_{-\infty}^{\infty} [f''(x)^2]\,\mathrm{d}x$ is the *roughness* of f'' and $\kappa^2 = \int_{-\infty}^{\infty} x^2 \times K(x)\,\mathrm{d}x$. (If the kernel is a symmetric density, then κ^2 is the variance of K.) Expression (30) quantifies the effects of the bandwidth h: the first term (from the integrated variance) is large when h is too small, whereas the second term (from the integrated squared bias) is large when h is too large.

The minimizer of $\mathrm{AMISE}(h)$, namely

$$h_{\mathrm{AMISE}} = \left[\frac{R(K)}{nR(f'')\kappa^4}\right]^{1/5}, \tag{31}$$

depends on the unknown quantity $R(f'')$ but offers a couple of interesting suggestions. First, the optimal bandwidth will approach zero as the sample size increases, but at a very slow rate. Second, small bandwidths will be required for rapidly varying densities f. Silverman (1986), Section 6.1.5, and Jones et al. (1996) review a variety of bandwidth selection methods, in particular, the method of Sheather and Jones (1991). An older simple rule is given in Section 3.4.1 of Silverman (1986):

$$h = 1.06 \min\bigl\{S_n(X), R/1.3\bigr\} n^{-1/5}, \tag{32}$$

where R is their inter-quartile range (difference between the 75 and 25 percentiles).

Remark 3. Epanechnikov's kernel from Table 1 is optimal as it minimizes $\mathrm{AMISE}(h_{\mathrm{AMISE}})$; see Bartlett (1963) and Epanechnikov (1969).

Remark 4. On top of conditions (28), assume that $\int_{-\infty}^{\infty} uK(u)\,\mathrm{d}u = 0$. Then the mean of $\hat{f}_K$ is

$$\begin{aligned}
\int_{-\infty}^{\infty} x\hat{f}_K(x)\,\mathrm{d}x &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{1}{h} xK\left(\frac{x - X_i}{h}\right)\mathrm{d}x \\
&\quad \text{(using the transformation } u = (x - X_i)/h\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} (X_i + uh)K(u)\,\mathrm{d}u \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i \int_{-\infty}^{\infty} K(u)\,\mathrm{d}u + \frac{1}{n}\sum_{i=1}^{n} h \int_{-\infty}^{\infty} uK(u)\,\mathrm{d}u \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i,
\end{aligned}$$

the sample mean of the data. Notice that the last expression is *not* the mean $\mathrm{E}[\hat{f}_K(x)]$.

Working similarly, one can show that the second moment of $\hat{f}_K(x)$ can be written as

$$\int_{-\infty}^{\infty} x^2 \hat{f}_K(x)\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\kappa^2.$$

Hence the variance of the density $\hat{f}_K$ is

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + h^2\kappa^2,$$

which is the sample second central moment of the data inflated by the term $h^2\kappa^2$.

Example 4. Table 2 contains 100 interarrival times from two customer classes into a system. Unfortunately, the data collector failed to record the proper class type; as a result, we cannot "separate" the data into the respective sets and proceed with each set.

Figure 1 displays density estimates based on the Gaussian kernel and bandwidths h equal to 6.258 and 2, respectively, in conjunction with a histogram based on a bin width of 10. The first value of h is based on Equation (32). The estimates were computed using MATLAB. Although both kernel density curves capture the bimodality of the data, the estimate based on $h = 2$ is very "jagged".

6.5 The variable kernel estimator

A potential problem arises when the kernel estimator is applied to data from long-tailed distributions: since the bandwidth h is fixed across the range of the

Table 2.
Interarrival times from 2 customer classes

11.97	2.04	12.02	2.81	11.45	8.19	1.05	7.35	13.42	2.28
24.22	14.44	17.42	21.47	38.53	41.06	29.89	20.40	44.11	38.31
13.48	1.49	1.67	10.44	29.79	9.21	1.88	2.94	6.07	3.49
25.43	21.05	30.31	37.41	78.71	19.95	42.67	31.26	30.55	19.09
14.49	2.10	6.36	15.36	8.17	6.75	4.37	6.98	3.83	17.30
36.61	27.80	31.72	69.06	20.34	36.21	44.65	52.72	12.83	21.00
5.12	17.86	17.76	13.16	4.31	5.92	36.88	3.53	10.53	5.05
22.22	40.69	37.38	32.02	32.71	17.28	16.13	20.45	38.72	24.35
2.04	8.63	4.10	8.26	4.86	7.68	11.93	14.31	1.02	10.11
19.57	14.07	12.95	21.64	19.52	18.02	29.37	15.33	21.12	23.35

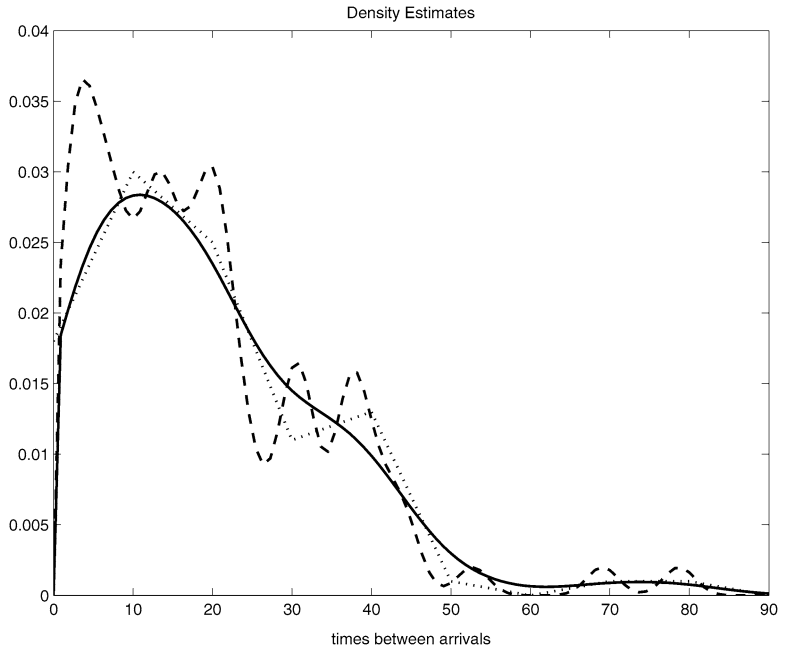

Fig. 1. Density estimates for the data set in Table 2 based on Gaussian kernels with bandwidths $h = 6.258$ (obtained from (32)) and $h = 2$. (⋯ histogram; — Gaussian kernel ($h = 6.258$); - - Gaussian kernel ($h = 2$).)

data, there is a tendency for spurious "noise" at the tails of the estimate due to the local sparsity of the data. This problem can be addressed as follows.

Let k be a positive integer, and define $d_{i,k}$ to be the distance from X_i to the kth closest observation among the set of remaining data. The *variable kernel* estimator with bandwidth h is defined by

$$\hat{f}_{\mathrm{VK}}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{hd_{i,k}} K\left(\frac{x - X_i}{hd_{i,k}}\right), \quad x \in \mathbb{R}. \tag{33}$$

Since the window width of the kernel placed at the point X_i is proportional to $d_{i,k}$, the observations in sparse regions will be associated with flatter kernels. As expected, for fixed k, the overall degree of smoothness in the estimate $\hat{f}_{\mathrm{VK}}$ will depend on the choice of h. A typical choice of k is $k \approx \sqrt{n}$.

Again, if K is a density, then $\hat{f}_{\mathrm{VK}}$ will also be a density that inherits the continuity and differentiability properties of K.

6.6 *Bounded domains*

Often, the natural support of a density is an interval bounded on one or both sides. An example, is the density of the interarrival times in Example 4.

In this case, the estimated density should also be zero outside the support of the unknown density. Silverman (1986) discussed this issue and a variety of ways to address it. We discuss a singe approach for the case where X is a non-negative variable. The method starts by augmenting the original set of data by adding their reflections with respect to zero; the new data set becomes $\{\pm X_1, \pm X_2, \ldots, \pm X_n\}$. If $\tilde{f}_K$ is a kernel estimate constructed based on the augmented set of data, then a natural estimate based on the original data set is the "folded" estimate

$$\hat{f}_K(x) = \begin{cases} 2\tilde{f}_K(x) & \text{for } x \geqslant 0, \\ 0 & \text{for } x < 0. \end{cases}$$

6.7 Density estimation for stationary processes

Kernel estimation for dependent stationary processes has received increasing interest during the last two decades. As we mentioned in the Introduction, the main objective is the estimation of the marginal density of a stationary process.

The literature considers two classes of estimators. The first class includes nonrecursive estimators of the form

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_n(x - X_i),$$

where the $K_n(\cdot)$ are kernel functions and the bandwidth for each n is part of K_n. (In what follows we replace the subscript "K" by "n" to denote the sample size.) Asymptotic properties of such estimators (e.g., convergence in mean, quadratic means and normality) are studied in Rosenblatt (1970) for Markov processes, and in Masry (1983), Robinson (1983, 1986) for strongly mixing processes.

Below we focus on estimators of the form

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right), \quad x \in \mathbb{R}, \tag{34}$$

where $\{h_i\}$ is a bandwidth sequence. The estimators (34) are attractive because they can be computed recursively by

$$\hat{f}_n(x) = \frac{n-1}{n}\hat{f}_{n-1}(x) + \frac{1}{nh_n} K\left(\frac{x - X_n}{h_n}\right).$$

Further we review results from Masry (1986). The reader is also pointed to Masry (1989), Masry and Györfi (1987), Györfi and Masry (1990), and several other references therein, for additional properties and alternative estimators. Clearly, the asymptotic results for the stationary processes generalize results for IID processes.

We will assume that the bandwidth sequence satisfies $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n = \infty$. A popular choice is

$$h_n = hn^{-\lambda}, \quad 0 < \lambda < 1, \tag{35}$$

a generalization of (32).

Theorem 5 shows that the estimator (34) is asymptotically (as $n \to \infty$) unbiased at each continuity point of f and displays the finite-sample bias. The "little-oh" notation $o(h(n))$ denotes a function $g(n)$ such that $g(n)/h(n) \to 0$ as $n \to \infty$.

Theorem 5 (Masry, 1986). *Suppose that the kernel K satisfies conditions (28), and let x be a continuity point of f. Then*

$$\mathrm{E}\big[\hat{f}_n(x)\big] \to f(x) \quad \textit{as } n \to \infty.$$

Suppose in addition that f is $r+1$ times continuously differentiable at the point x, $\sup_u |f^{(r+1)}(u)| < \infty$, and

$$\int_{-\infty}^{\infty} |u|^j \big|K(u)\big|\,\mathrm{d}u < \infty \quad \textit{for } j = 1, \ldots, r+1.$$

Also assume that there are finite constants β_ℓ such that, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_i}{h_n}\right)^{\ell} \to \beta_\ell \quad \textit{for } \ell = 1, \ldots, r+1.$$

Then

$$\mathrm{E}\big[\hat{f}_n(x)\big] = f(x) + \big(1 + o(1)\big)\sum_{\ell=1}^{r} \frac{c_\ell \beta_\ell (-h_n)^\ell}{\ell!} f^{(\ell)}(x) + O\big(h_n^{(r+1)}\big), \tag{36}$$

where

$$c_\ell = \int_{-\infty}^{\infty} u^\ell K(u)\,\mathrm{d}u \quad \textit{for } \ell = 1, \ldots, r.$$

Remark 5. If K is an even function, then $c_\ell = 0$ when ℓ is odd. For $r \geqslant 2$ in Theorem 5, the dominant term in the right-hand side of Equation (36) is $c_2\beta_2 f^{(2)}(x)/2$ so that the bias of $\hat{f}_n(x)$ decreases at the rate h_n^2. Also the bandwidth sequence in Equation (35) yields

$$\beta_\ell = \frac{1}{1 - \lambda\ell} \quad \text{if } 0 < \lambda\ell < 1.$$

Theorem 6 establishes the rate of convergence of the variance of $\hat{f}_n(x)$ to zero and the rate at which the covariance between values of $\hat{f}_n$ at distinct points approaches zero.

Theorem 6 (Masry, 1986). *Suppose that*:

(a) *The joint probability density* $f(x, y; j)$ *of the RVs* X_i *and* X_{i+j} *exists and satisfies*

$$\left|f(x, y; j) - f(x)f(y)\right| \leqslant M \quad \text{for all } (x, y) \text{ and } j \geqslant 1,$$

where M *is a finite constant.*

(b) $\{X_j\}$ *is strongly* α*-mixing with*

$$\sum_{j=1}^{\infty} \alpha_j^p < \infty \quad \text{for some } p \in \left(0, \frac{1}{2}\right).$$

(c) *The kernel* K *satisfies conditions* (28).

(d) *The bandwidth sequence satisfies*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_n}{h_i}\right)^r \to \theta_r < \infty \quad \text{as } n \to \infty,$$

for $r \in [1, 2)$, *where* θ_r *are constants.*

Then at all continuity points of f,

$$\lim_{n\to\infty} nh_n \operatorname{Var}\left[\hat{f}_n(x)\right] = \theta_1 f(x) \int_{-\infty}^{\infty} K^2(u)\,\mathrm{d}u \tag{37}$$

and

$$\lim_{n\to\infty} nh_n \operatorname{Cov}\left[\hat{f}_n(x), \hat{f}_n(y)\right] = 0 \quad \text{for } x \neq y.$$

Remark 6. The asymptotic variance expression (37) coincides with the expression obtained by Parzen (1962) for independent samples. Additional terms due to the serial dependencies are listed in the proof of Theorem 3 of Masry (1986).

Remark 7. If f is three times continuously differentiable ($r = 2$ in Theorem 5) and K is an even function (e.g., a density), we have

$$\operatorname{Bias}\left[\hat{f}_n(x)\right] \approx \frac{c_2\beta_2 f^{(2)}(x)}{2} h_n^2$$

and

$$\operatorname{Var}\left[\hat{f}_n(x)\right] \approx \frac{\theta_1 f(x) \int_{-\infty}^{\infty} K^2(u)\,\mathrm{d}u}{nh_n}.$$

One can show that the bandwidth that minimizes the mean squared error $\operatorname{MSE}[\hat{f}_n(x)]$ is $h_n = hn^{-1/5}$ and the convergence rate of $\operatorname{MSE}[\hat{f}_n(x)]$ to zero is $n^{-4/5}$. Further, $\beta_2 = 5/3$ and $\theta_1 = 5/6$.

We finish our discussion with the asymptotic normality of the recursive estimators. The following theorem is similar to Theorem 8 of Masry (1986).

Theorem 7. *Assume that the conditions of Theorem 6 hold. Also assume that there exist a constant* $\gamma \in (0, 1)$ *and a sequence* $\{m_n\}$ *of integers with* $\lim_{n\to\infty} m_n = 0$ *such that* $\lim_{n\to\infty} nh_n^{3-2\gamma} = 0$, $m_n = o(n^{1/2}h_n^{3/2-\gamma})$, *and*

$$\sqrt{\frac{n}{h_n}} \sum_{j=m_n}^{\infty} \alpha_j^{1-\gamma} \to 0 \quad \text{as } n \to \infty.$$

Then at each continuity point of f *with* $f(x) > 0$,

$$\frac{\hat{f}_n(x) - \mathrm{E}[\hat{f}_n(x)]}{\sigma_f/\sqrt{nh_n}} \xrightarrow{d} \mathrm{N}(0, 1) \quad \text{as } n \to \infty, \tag{38}$$

where

$$\sigma_f^2 = \theta_1 f(x) \int_{-\infty}^{\infty} K^2(u)\,\mathrm{d}u.$$

Equation (38) is not convenient for computing confidence intervals for $f(x)$ because it does not involve the actual error $\hat{f}_n(x) - f(x)$. To overcome this problem we write

$$\frac{\hat{f}_n(x) - \mathrm{E}[\hat{f}_n(x)]}{\sigma_f/\sqrt{nh_n}} = \frac{\hat{f}_n(x) - f(x)}{\sigma_f/\sqrt{nh_n}} - \frac{\mathrm{E}[\hat{f}_n(x)] - f(x)}{\sigma_f/\sqrt{nh_n}}. \tag{39}$$

Under the conditions in Remark 7, we have $\mathrm{Bias}[\hat{f}_n(x)] = \mathrm{O}(h_n^2)$. Hence if $\{h_n\}$ is such that $\lim_{n\to\infty} nh_n^5 = 0$, the last term in Equation (39) converges to zero and Slutsky's theorem (Karr, 1993, Theorem 5.20), implies

$$\frac{\hat{f}_n(x) - f(x)}{\sigma_f/\sqrt{nh_n}} \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1) \quad \text{as } n \to \infty. \tag{40}$$

Equation (40) yields the following approximate $1 - \alpha$ CI for $f(x)$:

$$\hat{f}_n(x) \pm z_{1-\alpha_2} \frac{\hat{\sigma}_f}{\sqrt{nh_n}},$$

where

$$\hat{\sigma}_f = \left[\theta_1 \hat{f}_n(x) \int_{-\infty}^{\infty} K^2(u)\,\mathrm{d}u\right]^{1/2}$$

is an estimator for σ_f.

Notice that the condition $nh_n^5 \to 0$ is not satisfied by Silverman's rule in Equation (32) or by the bandwidth sequence in Remark 7 that minimizes

MSE$[\hat{f}_n(x)]$. Equations (38) and (40) reveal that the rate of the pointwise convergence of the recursive kernel estimator $\hat{f}_n(x)$ to the normal distribution is $(nh_n)^{-1/2}$ or $n^{-(1-\lambda)/2}$ for bandwidth sequences of the form $h_n = hn^{-\lambda}$, with $1/5 < \lambda < 1$ used to obtain Equation (40). Such rates are significantly slower that the usual rate of $n^{-1/2}$ for point estimators for means, but typical for kernel estimators.

Remark 8. For Markov chains (not necessarily discrete-time ones) one can obtain the convergence rate of $n^{-1/2}$ by means of the "look-ahead" density estimation method of Henderson and Glynn (2001).

Example 5. Consider an $M/G/1$ queueing system where units arrive according to a Poisson process with a mean interarrival time $\mathrm{E}(A) = 10$ and experience service times from the distribution with density function

$$g(x) = 0.75\left(\frac{1}{4}x\mathrm{e}^{-x/2}\right) + 0.75\left(\frac{1}{19!}x^{19}\mathrm{e}^{-x}\right).$$

This model is applicable when the units are classified into two types, 1 and 2, with respective probabilities 0.75 and 0.25; type 1 units require service times from the gamma distribution with shape parameter 2 and scale parameter 2; and type 2 units require service times from the gamma distribution with shape parameter 20 and scale parameter 1. In other words, the service time distribution is a mixture of two gamma distributions with modes of 2 and 19. Since the mean service time is $\mathrm{E}(S) = 0.75(4) + 0.25(20) = 8$, the traffic intensity is $\nu = 8/10 < 1$ and the system is stable.

The objective of this experiment is the estimation of the density of the time a unit spends in the system (flow time) in steady state. The substantial distance between the modes of the gamma distributions associated with the service times causes the steady-state density of the flow times to be bimodal. We made a single run starting with the empty and idle state and collected 7000 observations. To reduce the effect of the transient phase, we discarded the first 5000 observations and computed the density estimates in Figure 2 using the remaining 2000 observations. The data and the estimates were generated using MATLAB. The kernel estimate is based on the Gaussian kernel with bandwidth $10n^{-1/5}$ whereas the histogram is based on a bin width of 5. Both estimates clearly indicate the bimodal nature of the steady-state density of the flow times.

7 Summary

This chapter served as a "precursor" to several forthcoming chapters. Its scope was the introduction of the reader to the statistical issues involved in simulation experiments. We started with a review of probabilistic and statistical

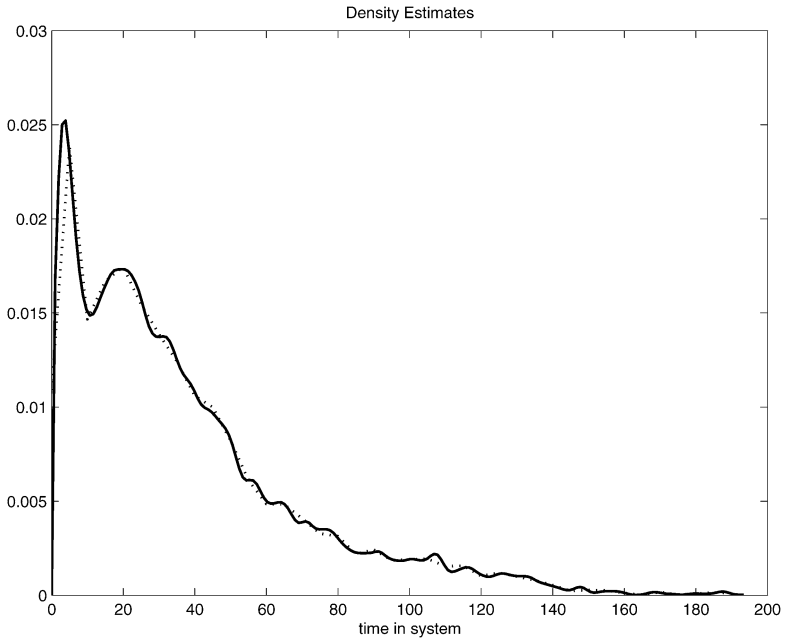

Fig. 2. Kernel estimate for the limiting distribution of the system flow time in an $M/G/1$ system. (⋯ histogram; — Gaussian kernel.)

concepts that suffice for analyzing data from finite-horizon simulations. Then we laid out the issues related to steady-state simulations. Since these issues are multi-faceted and challenging, we reviewed the basic concepts from stochastic processes and a few limit theorems that are required to carry out the necessary statistical analyses.

We proceeded with a review of a research area that has received little attention in the simulation domain, density estimation. Within this subject, we started with a review of techniques for IID data and concluded with a recent results related to data from stationary sequences. Since this subject will not receive further consideration in this volume, we encourage the eager reader to study the literature.

We finish with a couple of observations. First, kernel density estimators converge to the respective densities with rates that are significantly lower than the usual rate of $n^{-1/2}$ for point estimators for means (an exception is the estimator in Henderson and Glynn, 2001). Second, a problem that has received little or no attention in the literature is the generation of realizations from kernel density estimators (or the respective CDFs). This is an important issue as the efficient generation of random variates is a paramount issue in simulation experiments.

Acknowledgements

I would like to thank Jim Wilson for many fruitful discussions and Melike Meterelliyoz for her careful reading.

References

Alexopoulos, C., Goldsman, D. (2004). To batch or not to batch. *ACM Transactions on Modeling and Computer Simulation* 14 (1), 76–114.

Alexopoulos, C., Seila, A.F. (1998). Output data analysis. In: Banks, J. (Ed.), *Handbook of Simulation*. Wiley, New York, pp. 225–272.

Alexopoulos, C., Goldsman, D., Serfozo, R.F. (2006). Stationary processes: Statistical estimation. In: Balakrishnan, N., Read, C., Vidakovic, B. (Eds.), *Encyclopedia of Statistical Sciences*, 2nd edition. Hoboken, NJ, pp. 7991–8006.

Bartlett, M.S. (1963). Statistical analysis of density functions. *Sankhyā Series A* 25, 245–254.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Blomqvist, N. (1967). The covariance function of the $M/G/1$ queueing system. *Skandinavisk Aktuarietidskrift* 50, 157–174.

Castellana, J.V., Leadbetter, M.R. (1986). On smoothed probability density estimation for stationary processes. *Stochastic Processes and Their Applications* 21, 179–193.

Chatfield, C. (1989). *The Analysis of Time Series: An Introduction*, 4th edition. Chapman and Hall, New York.

Chow, Y.S., Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Annals of Mathematical Statistics* 36, 457–462.

Durrett, R. (2005). *Probability: Theory and Examples*, 3rd edition. Brooks/Cole–Thomson Learning, Belmont, CA.

Epanechnikov, V.A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory of Probability and Its Applications* 14, 153–158.

Fishman, G.S. (2001). *Discrete-Event Simulation: Modeling, Programming, and Analysis*. Springer-Verlag, New York.

Glynn, P.W., Heidelberger, P. (1991). Analysis of initial transient deletion for replicated steady-state simulations. *Operations Research Letters* 10, 437–443.

Glynn, P.W., Iglehart, D.L. (1985). Large-sample theory for standardized time series: An overview. In: Gantz, D., Blais, G., Solomon, S. (Eds.), *Proceedings of the 1985 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 129–134.

Gross, D., Harris, C.M. (1998). *Fundamentals of Queueing Theory*, 3rd edition. Wiley, New York.

Gross, D., Shortle, J.F., Fisher, M.J., Masi, D.M.B. (2002). Difficulties in simulating queues with Pareto services. In: Yucesan, E., Chen, C.H., Snowdon, J.L., Charnes, J.M. (Eds.), *Proceedings of the 2002 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 407–415.

Györfi, L., Masry, E. (1990). The L_1 and L_2 strong consistency of recursive density estimation from dependent samples. *IEEE Transactions on Information Theory* 36 (3), 531–539.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S*. Springer-Verlag, New York.

Henderson, S.G., Glynn, P.W. (2001). Computing densities for Markov chains via simulation. *Mathematics of Operations Research* 26 (2), 375–400.

Hogg, R.V., Craig, A.T. (1995). *Introduction to Mathematical Statistics*, 5th edition. Prentice Hall, Upper Saddle River, NJ.

Jones, M.C., Marron, J.S., Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the Americal Statistical Association* 91 (433), 402–407.

Karr, A.F. (1993). *Probability*. Springer-Verlag, New York.

Law, A.M., Kelton, W.D. (2000). *Simulation Modeling and Analysis*, 3rd edition. McGraw-Hill, New York.

Law, A.M., Kelton, W.D., Koenig, L.W. (1981). Relative width sequential confidence intervals for the mean. *Communications in Statistics B* 10, 29–39.

Lindley, D.V. (1952). The theory of queues with a single server. *Proceedings of the Cambridge Philosophical Society* 48, 277–289.

Masry, E. (1983). Probability density estimation from sampled data. *IEEE Transactions on Information Theory* 29, 696–709.

Masry, E. (1986). Recursive probability density estimation for weakly dependent stationary processes. *IEEE Transactions on Information Theory* 32 (2), 254–267.

Masry, E. (1989). Nonparametric estimation of conditional probability densities and expectations of stationary processes: Strong consistency and rates. *Stochastic Processes and Their Applications* 31, 109–127.

Masry, E., Györfi, L. (1987). Strong consistency and rates for recursive probability density estimators of stationary processes. *Journal of Multivariate Analysis* 22, 79–93.

Nadas, A. (1969). An extension of the theorem of Chow and Robbins on sequential confidence intervals for the mean. *Annals of Mathematical Statistics* 40, 667–671.

Parzen, E. (1962). On estimation of probability density function and mode. *Annals of Mathematical Statistics* 33, 1065–1076.

Robinson, P.M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis* 4, 185–207.

Robinson, P.M. (1986). On the consistency and finite sample properties of nonparametric kernel time series regression, autoregression and density estimators. *Annals of the Institute of Mathematical Statistics* 38, 539–549.

Rosenblatt, M. (1956). A central limit theorem and strong mixing condition. *Proceedings of the National Academy of Sciences* 42, 43–47.

Rosenblatt, M. (1970). Density estimates and Markov sequences. In: Puri, M. (Ed.), *Nonparametric Techniques in Statistical Inference*. Cambridge University Press, London, pp. 199–210.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley-Interscience, New York.

Serfozo, R. (1999). *Introduction to Stochastic Networks*. Springer-Verlag, New York.

Sheather, S.J., Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B* 53, 683–690.

Silverman, B.W. (1986). *Density Estimation*. Chapman and Hall, London.

Starr, N. (1966). The performance of a statistical procedure for the fixed-width interval estimation for the mean. *Annals of Mathematical Statistics* 37 (1), 36–50.

Stidham, S. (1974). A last word on $L = \lambda W$. *Operations Research* 22, 417–421.

Welch, P.D. (1983). The statistical analysis of simulation results. In: Lavenberg, S. (Ed.), *The Computer Performance Modeling Handbook*. Academic Press, New York, pp. 268–328.