

## EMPIRICAL EVALUATION OF DATA-BASED DENSITY ESTIMATION

E. Jack Chen

BASF Corporation  
333 Mount Hope Avenue  
Rockaway, New Jersey 07866, U.S.A.

W. David Kelton

Department of Quantitative Analysis and  
Operations Management  
University of Cincinnati  
Cincinnati, Ohio 45221, U.S.A.

### ABSTRACT

This paper discusses implementation of a sequential procedure to estimate the steady-state density of a stochastic process. The procedure computes sample densities at certain points and uses Lagrange interpolation to estimate the density  $f(x)$ . Even though the proposed sequential procedure is a heuristic, it does have strong basis. Our empirical results show that the procedure gives density estimates that satisfy a pre-specified precision requirement. An experimental performance evaluation demonstrates the validity of using the procedure to estimate densities.

### 1 INTRODUCTION

Simulation studies have been used to investigate the characteristics of the system under study, for example the mean and the variance of certain system performance. The probability density function  $f$  gives a natural description of the distribution of the output random variable  $X$  produced by a simulation. The density function associated with  $X$  satisfies

$$P(a < X < b) = \int_a^b f(x)dx \text{ for all } a < b.$$

We investigate the performance of using the technique of Chen and Kelton (2006) to estimate the density of a simulation output random variable.

*Density estimation* is the construction of an estimate of the density function from observed data. Silverman (1986, p. 5) points out that “density estimates are ideal for presentation of data to provide explanation and illustration of conclusions, since they are fairly easily comprehensible to non-mathematicians.” One approach to density estimation is *parametric*, assuming that the data are drawn from a known parametric family of distributions. Another approach is *nonparametric*, where less rigid assumptions are made about the distribution of the observed data. We consider the nonparametric approach. Furthermore, the procedure is a

data-based algorithm, i.e., the procedure can be embodied in a software package whose input is the data  $(X_1, \dots, X_n)$  and whose output is the density estimate. Several different approaches have received extensive treatment; see Scott and Factor (1981) and the references in the paper.

The most widely used density estimator is the *histogram*, a graphical estimate of the underlying probability density function and reveals all the essential distributional features of an output random variable analyzed by simulation, such as skewness and multimodality. Hence, a histogram is often used in the informal investigation of the properties of a given set of data. Given an origin  $g_0$  and a bin width  $w$ , the bins of the histogram are the intervals  $[g_0 + mw, g_0 + (m+1)w]$  for positive and negative integers  $m$ . Suppose that we have any division of the real line into bins; then the histogram density estimator is

$$\hat{f}_h(x) = \frac{1}{n} \times \frac{(\text{no. of } X_i \text{ in same bin as } x)}{(\text{width of bin containing } x)}.$$

Hence, to construct the histogram, we need to choose both an origin and a bin width. It is the bin width that, primarily, controls the amount of smoothing inherent in the procedure. Scott and Factor (1981) investigate the optimal bin width given a sample size. They point out that the optimal smoothing parameter can be computed if the true underlying sampling density  $f$  is known.

A histogram can be constructed with a properly selected set of quantiles. It is known that for both independent and identically distributed (i.i.d.) and  $\phi$ -mixing sequences sample quantiles will be asymptotically unbiased if certain conditions are satisfied; see Sen (1972). Intuitively, a stochastic process is  $\phi$ -mixing if its distant future is essentially independent of its present and past (Billingsley 1999).

In Section 2 we discuss some theoretical bases of density estimation in the context of simulation output analysis. In Section 3 we present our methodologies and the proposed procedure for density estimation. In Section 4 we show

our empirical-experimental results of density estimation. In Section 5 we give concluding remarks.

## 2 THEORETICAL BASIS

From the definition of a probability density, if the random variable  $X$  has density  $f$ , then

$$f(x) = \lim_{h \rightarrow 0} \frac{1}{2h} P(x-h < X < x+h).$$

### 2.1 The Histogram Density Estimator

A natural estimator by histogram  $\hat{f}_h$  of the density is given by choosing a small number  $h$  ( $h = w/2$ ) and setting

$$\begin{aligned} & \hat{f}_h(x) \\ = & \frac{1}{2nh} [\text{no. of } X_1, \dots, X_n \text{ falling in } (x-h, x+h)]. \end{aligned}$$

Let  $p = P(|x - X| < h)$  and

$$I_i = \begin{cases} 1 & \text{if } |x - X_i| < h, \\ 0 & \text{otherwise.} \end{cases}$$

The estimator  $\hat{f}_h(x)$  is based on a transformation of the output sequence  $\{X_i\}$  to the sequence  $\{I_i\}$ ,  $i = 1, 2, \dots, n$ :

$$\hat{f}_h(x) = \frac{1}{2nh} \sum_{i=1}^n I_i.$$

For data that are i.i.d., the following properties of  $I_i$  are well known (Hogg and Craig 1995, pp. 116-117):  $E(I_i) = p$  and  $\text{Var}(I_i) = p(1-p)$ . Chen (2001) has developed a procedure to estimate proportion of simulation output sequences based on these properties. Since  $\hat{f}_h(x)$  is based on the mean of the random variable  $I_i$ , we can use any method developed for estimating the variance of the mean to estimate  $\text{Var}(\hat{f}_h(x))$ . Let  $I(\cdot)$  denote the indicator function for the interval  $(x-h, x+h)$ . It can be shown that

$$E(\hat{f}_h(x)) = \frac{1}{2h} \int I(y) f(y) dy,$$

and  $\text{Var}(\hat{f}_h(x)) = p(1-p)/(4nh^2)$ .

Note that  $\hat{f}_h(x)$  has a binomial distribution. It follows from the definition that  $\hat{f}_h$  is not a continuous function, but has jumps at the points  $X_i \pm h$  and has zero derivative everywhere else. This gives the estimate a somewhat ragged character.

### 2.2 The Kernel Density Estimator

To overcome the difficulties of ragged character, one can use a *kernel function*  $K$ , which satisfies the condition

$$\int_{-\infty}^{\infty} K(x) dx = 1.$$

For simplicity, the kernel  $K$  usually is a symmetric function satisfying  $\int K(x) dx = 1$ ,  $\int xK(x) dx = 0$ , and  $\int x^2 K(x) dx = k_2 \neq 0$ ; an example is the normal density. The kernel estimator with kernel  $K$  is defined by

$$\hat{f}_k(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

Silverman (1986, pp. 15-17) points out that “ $\hat{f}_k$  will inherit all the continuity and differentiability properties of the kernel  $K$ , so that if  $K$  is the normal density function, then  $\hat{f}_k$  will be a smooth curve having derivatives of all orders.” However, the kernel method often underestimates the density at the boundary when the domain of the density being estimated is not the whole real line but an interval bounded on one or both sides; see Silverman (1986, p. 29).

It can be shown that

$$E(\hat{f}_k(x)) = \frac{1}{h} \int K\left(\frac{x-y}{h}\right) f(y) dy; \quad (1)$$

$$\begin{aligned} & \text{Var}(\hat{f}_k(x)) \\ = & \frac{1}{n} \left[ \frac{1}{h^2} \int K\left(\frac{x-y}{h}\right)^2 f(y) dy - E(\hat{f}_k(x))^2 \right] \\ \approx & \frac{1}{nh} f(x) \int K(y)^2 dy, \end{aligned}$$

and

$$\begin{aligned} \text{bias}_h(x) &= E(\hat{f}_k(x)) - f(x) \\ &= \frac{1}{2} h^2 f''(x) k_2 + \text{higher-order terms in } h. \end{aligned}$$

See Silverman (1986, pp. 37-40) for details.

The approximation of bias and variance indicates one of the fundamental difficulties of density estimation. To eliminate the bias, a small value of  $h$  should be used, but then the variance will become large. On the other hand, a large value of  $h$  will reduce the variance, but will increase the bias. Nevertheless, the ideal window width  $h$  should satisfy:

$$\lim_{n \rightarrow \infty} h = 0, \quad \lim_{n \rightarrow \infty} nh = \infty. \quad (2)$$

That is,  $h$  should converge to zero as the sample size increases but at a slower rate than  $n$ . Furthermore, smaller values of  $h$  will be appropriate for more rapidly fluctuating densities. Let  $J_n = \int |\hat{f}_n - f|$ , where  $\hat{f}_n(x) = f_n(x; X_1, \dots, X_n)$  is a real-valued Borel measurable function of its arguments. Conditions (2) imply that there exists  $r \in R$  such that  $P(J_n \geq \varepsilon) \leq e^{-rne^2}$  for all  $\varepsilon \in (0, 1)$  and all  $n \geq n_f$ , where  $n_f$  depends upon  $f$  and  $\varepsilon$ ; see Devroye and Györfi (1985).

Since the shape of the true density is of most interest, a relevant criterion is the integrated mean squared error (IMSE) (Rosenblatt 1971) defined as

$$\text{IMSE} = \int E[\hat{f}(x) - f(x)]^2 dx = E \int [\hat{f}(x) - f(x)]^2 dx.$$

There exists extensive research on the selection of an optimal kernel function to minimize the IMSE. Scott and Factor (1981) point out that many symmetric uni-modal kernel functions are nearly optimal. We use the Gaussian kernel in this paper. Moreover, it can be shown that the asymptotically optimal smoothing parameter is

$$h = \alpha(K)\beta(f)n^{-1/5}$$

where

$$\alpha(K) = \left[ \int K(y)^2 dy \right]^{1/5} \left[ \int K(y)y^2 dy \right]^{-2/5}$$

and

$$\beta(f) = \left[ \int f''(x)^2 dx \right]^{-1/5}.$$

With this choice, the IMSE decreases in proportion to  $n^{-4/5}$ ; see Scott and Factor (1981).

### 3 METHODOLOGIES

This section presents the methodologies we will use for our density estimation. A flow chart of the procedure is depicted in Figure 1. An imbedded pilot run is executed to set up the grid points. On each iteration, the algorithm operates as follows. The simulation outputs are funnelled into grids. The number of observations in each grid is updated as the observation is processed. The systematic samples are obtained through lag- $l$  observations and are stored in a buffer. The initial value of  $l$  is 1. If lag- $l'$  systematic samples appear to be dependent, then the lag  $l$  is doubled every other iteration and the process is repeated until the lag- $l'$  systematic samples appear to be independent. The initial value of  $l'$  is 0 and will be updated each iteration by the following rule: “if  $l' < 3$ , then  $l' = l' + 1$ ; else  $l' = 2$ .”

#### 3.1 Determine the Window Width

Scott and Factor (1981) point out that “the great potential of nonparametric density estimators in data analysis is not being fully realized, primarily because of the practical difficulty associated with choosing the smoothing parameter given only data  $X_1, X_2, \dots, X_n$ .” Silverman (1986, p. 47) suggests that the window width of the kernel estimator be

$$h = 0.9An^{-1/5},$$

where

$$A = \min(\text{standard deviation, interquartile range}/1.34).$$

For many purposes this will be an adequate choice of window width in terms of minimizing the IMSE. For others, it will be a good starting point for subsequent fine tuning. Let  $x_p$  be the  $p$  sample quantile. In our procedure, we set

$$A = \min(\text{standard error}, (x_{0.75} - x_{0.25})/2.68).$$

Let  $x_{[1]}$  and  $x_{[n]}$ , respectively, denote the minimum and maximum of the initial  $n_0, 2n_0$ , or  $3n_0$  observations, depending on the correlation of the output sequences. If  $(x_{[n]} - x_{[1]})/(2h) < 25$ , then  $h$  will be halved. This adjustment is needed for distributions that have relatively large variance with a small range of the initial observations, so too large a window width. For example, if the estimated number of bins is 10, the procedure increases the number of bins to 20 and reduces the window width by half.

We use the following strategy to determine the grid points. There are two categories of grids: main grids and auxiliary grids. Main grids are constructed based on the initial observations that “anchor” the grid of the simulation-generated histogram, while auxiliary grids are extensions of main grids to ensure that the grids cover future observations. The number of main grid points is  $G_m = \lceil (x_{[n]} - x_{[1]})/(2h) \rceil$ , and the number of auxiliary grid points is  $G_a = 2\lceil \delta G_m \rceil$ , where  $0 < \delta < 1$ . The total number of grid points is thus  $G = G_m + G_a + 1$ . Let the beginning indices of the main grid point (i.e., the origin) be  $b = \lceil \delta G_m \rceil + 1$ . The procedure sets  $g_{b+i} = x_{[1]} + 2ih$ , for  $i = 0, 1, \dots, G_m + G_a/2 - 1$ , and  $g_{b-i} = x_{[1]} - 2ih$ , for  $i = 1, 2, \dots, G_a/2 - 1$ . Grid point  $g_1$  is set to  $-\infty$  and  $g_G$  is set to  $\infty$ .

It is straightforward to compute the histogram density estimator. The array  $n_i, i = 2, 3, \dots, G$  stores the number of observations between grid points  $g_{i-1}$  and  $g_i$ , so the density of  $x_i = (g_{i-1} + g_i)/2$  can be estimated by  $\hat{f}(x_i) = n_i/(n(g_i - g_{i-1}))$ , where  $n = \sum_{i=2}^G n_i$  is the total number of observations. To obtain the kernel estimator, the procedure needs to read through the output sequence again.

In some sense, a simulation is just a *function*, which may be vector-valued or stochastic. The explicit form of this

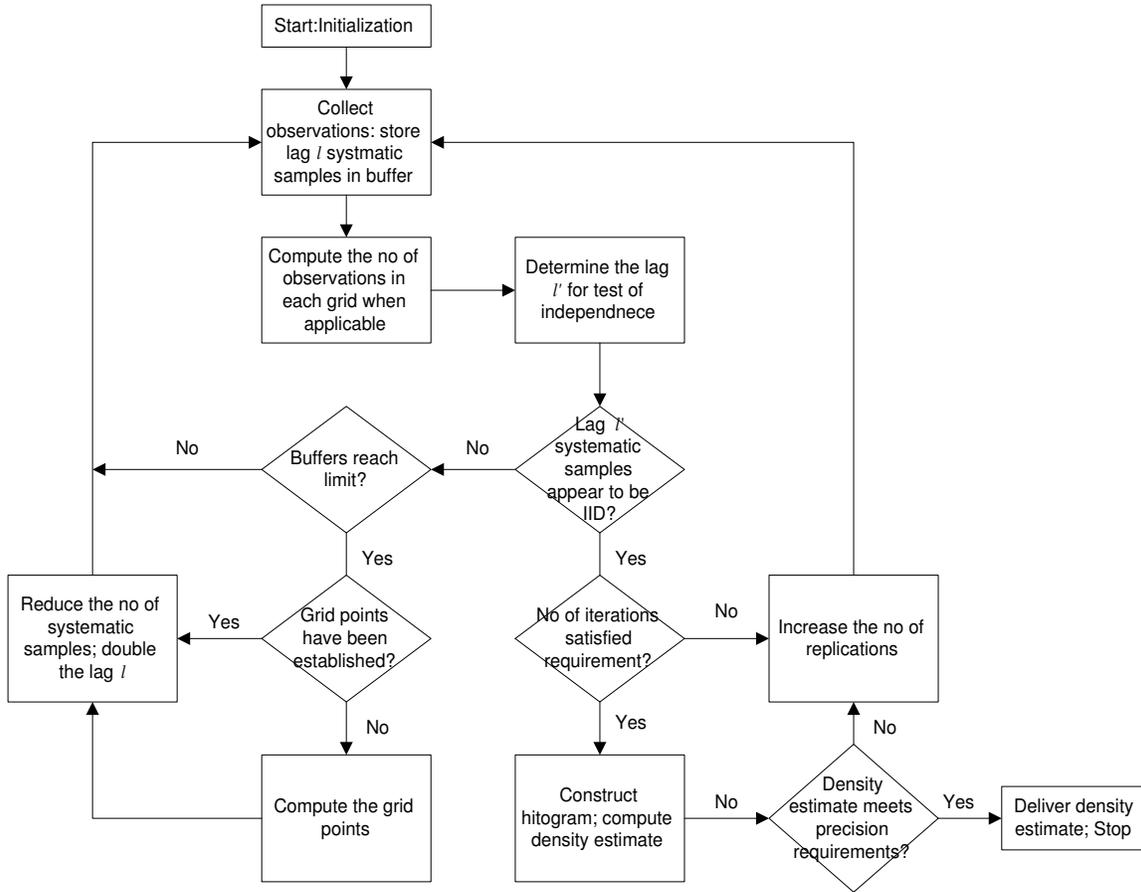


Figure 1: Flow Chart of the Procedure

function is unknown. These plots (histograms) numerically characterize the distribution of the output sequence, even though we do not have an algebraic formula with which to characterize it. The procedure then computes the point density estimator via the histogram by (four-point) Lagrange interpolation (Knuth 1998). That is, for some  $k$  such that  $x_{k-1} < x \leq x_k$ , the  $x$  density point estimator can be computed as follows. Let

$$w_j = \prod_{j'=1, j' \neq j}^4 \frac{x - x_{k+j'-3}}{x_{k+j-3} - x_{k+j'-3}}, \text{ for } j = 1, 2, 3, 4,$$

then

$$\hat{f}(x) = \sum_{j=1}^4 w_j \hat{f}(x_{k+j-3}).$$

In two extreme cases,  $x_1 < x \leq x_2$  or  $x_{G-1} < x \leq x_G$ , linear interpolation will be used.

Since the procedure uses interpolation to obtain point estimates, it eliminates the ragged character of the histogram. Hence, the density estimates for different points within the same bin can have different values.

### 3.2 Determine the Sample Size

The asymptotic validity of the density estimate is reached as the sample size or simulation run length gets large. However, in practical situations simulation experiments are restricted in time and it is not known in advance what is the required simulation run length for the estimator to become unbiased. Moreover, estimating the variance of the density estimator is needed to evaluate its precision. Therefore, a workable finite sample size must be determined dynamically for the precision required.

We use an initial sample size of  $n_0 = 600$ , which is somewhat arbitrary. If the underlying sequence is only slightly correlated and high precision is desired, a larger initial sample size should be used. For correlated sequences, the sample size  $n$  will be replaced with  $N = nl$ . Here  $l$  will be chosen sufficiently large so that systematic samples that are lag- $l$  observations apart are statistically independent; see Chen and Kelton (2003). This is possible because we assume the underlying process satisfies the property that the autocorrelation approaches zero as the lag approaches infinity. Consequently, the final sample size  $N$  increases as the auto-correlation increases. In this procedure, we

use the von Neumann (1941) test of independence instead of the runs test. We can apply the von Neumann test of independence with a smaller sample size, but it has less power. Nevertheless, it serves the purpose well.

Since we need to process the sequence again to obtain the kernel estimator, we re-compute the window width  $h$  with the final sample size  $N$  and the number of grid points with the new sample range. We only need to allocate main grids because the minimum and maximum are known. Furthermore, the sample error and the quantiles  $x_{0.25}$  and  $x_{0.75}$  will be estimated through the histogram constructed while calculating the natural estimator. That is, the variance is conservatively estimated by

$$S_H^2 = \sum_{i=2}^G \max((g_{i-1} - \bar{X}(N))^2, (g_i - \bar{X}(N))^2) P_i.$$

Note that  $N = nl = \sum_{i=2}^G n_i$ ,  $\bar{X}(N) = \sum_{j=1}^N X_j$ , and  $P_i = n_i/N$ .

To estimate the error, the IMSE is approximated by

$$\overline{IMSE} = 2 \sum_{r=1}^{\Gamma} \sum_{i=2}^{G-1} [\hat{f}(g_i) - f(g_i)]^2 h / \Gamma,$$

where  $\Gamma$  is the number of density estimates. The density of  $g_1$  and  $g_G$  is not included in the calculation because they could be  $-\infty$  and  $\infty$ , respectively. Furthermore, if the true minimum ( $\omega$ ) or the true maximum ( $\Omega$ ) are known, the values  $g_i < \omega$  or  $\Omega < g_i$  will not be included in the calculation.

### 3.3 Density Confidence Interval

An approximate pointwise confidence interval (c.i.) for the density  $f(x)$  can be obtained using the binomial distribution from the histogram density estimate. The usual unbiased estimator of the variance of  $\hat{f}_h(x)$  is  $S_b^2 = \text{Var}(\hat{f}_h(x)) = p(1-p)/(4h^2N)$ . This would then lead to the  $100(1-\alpha)\%$  c.i., for  $f(x)$ ,

$$\hat{f}_h(x) \pm z_{1-\alpha/2} S_b,$$

where  $z_{1-\alpha/2}$  is the  $1-\alpha/2$  quantile for the standard normal distribution. On the other hand, the distribution of  $\hat{f}_k(x)$  is unknown, hence, a c.i. cannot be constructed through one replication of  $\hat{f}_k(x)$ .

Let  $\hat{f}_r(x)$  denote the (histogram or kernel) estimator of  $f(x)$  in the  $r^{\text{th}}$  replication. We use

$$\bar{f}(x) = \frac{1}{R} \sum_{r=1}^R \hat{f}_r(x)$$

as a point estimator of  $f(x)$ . Assuming  $\bar{f}(x)$  has a limiting normal distribution, by the central limit theorem a c.i. for  $f(x)$  using the i.i.d.  $\hat{f}_r(x)$ 's can be approximated using standard statistical procedures. That is, the ratio

$$T = \frac{\bar{f}(x) - f(x)}{S/\sqrt{R}}$$

would have an approximate  $t$  distribution with  $R-1$  d.f. (degrees of freedom), where

$$S^2 = \frac{1}{(R-1)} \sum_{r=1}^R (\hat{f}_r(x) - \bar{f}(x))^2$$

is the usual unbiased estimator of the variance of  $f(x)$ . This would then lead to the  $100(1-\alpha)\%$  c.i., for  $f(x)$ ,

$$\bar{f}(x) \pm t_{R-1, 1-\alpha/2} \frac{S}{\sqrt{R}}, \tag{3}$$

where  $t_{R-1, 1-\alpha/2}$  is the  $1-\alpha/2$  quantile for the  $t$  distribution with  $R-1$  d.f. ( $R \geq 2$ ).

Let the half-width  $H$  be  $t_{R-1, 1-\alpha/2} S/\sqrt{R}$ . The final step in the procedure is to determine whether the c.i. meets the user's half-width requirement, a maximum absolute half-width  $\epsilon'$  or a maximum relative fraction  $\gamma$  of the magnitude of the final point density estimator  $\bar{f}(x)$ . If the relevant requirement  $H \leq \epsilon'$  or  $H \leq \gamma|\hat{f}(x)|$  for the precision of the confidence interval is satisfied, then the procedure terminates, returns the point density estimator  $\hat{f}(x)$ , and the c.i. with half-width  $H$ . If the precision requirement is not satisfied with  $R$  replications, then the procedure will increase the number of replications by one. This iterates until the pre-specified half-width is achieved.

## 4 EMPIRICAL EXPERIMENTS

We tested the proposed procedure with several i.i.d. and correlated sequences. In these experiments, we used  $R = 3$  independent replications to construct c.i.'s. We constructed density c.i.'s at four points for each distribution. The confidence level  $1-\alpha$  of the density c.i.— i.e., (3)—is set to 0.90. Moreover, the confidence level of the von Neumann test of independence is set to 0.90 as well.

We tested the following independent sequences:

- Observations are i.i.d. from the tri-modal density  $f(x) =$

$$\frac{1}{3\sqrt{2\pi}} (e^{-x^2/2} + \frac{1}{2}e^{-(x-5)^2/8} + e^{-(x-10)^2/2}).$$

- Observations are i.i.d. from the exponential density

$$f(x) = \begin{cases} e^{-x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Tables 1 and 2 list the experimental results using the tri-modal and exponential distributions, respectively. Each design point was based on 1000 replications. The *avg N* row lists the average of the sample size of each independent run. The *stdev N* row lists the standard deviation of the sample size. The *x* row lists the point where we want to estimate density. The *f(x)* row lists the true density. The values after each the estimate method are the  $\overline{IMSE}$  and the standard error of integrated squared error. The  $\hat{f}(x)$  row lists the grand mean of all density estimator from these 1000 replications. The *coverage* row lists the percentage of the c.i.'s that cover the true  $f(x)$ . The *avg  $\gamma$*  row lists the average of the relative precision of the density estimators. Here, the relative precision is defined as  $\gamma = |\hat{f}(x) - f(x)|/f(x)$ . The *stdev  $\gamma$*  row lists the standard deviation of the relative precision of the density estimators. The *avg hw* row lists the average of the c.i. half-width. The *stdev hw* row lists the standard deviation of the c.i. half-width.

As expected, the  $\overline{IMSE}$  from the kernel estimator is better than from the histogram estimator. However, the kernel estimator requires more computation. In these experiments, no relative or absolute precisions were specified, so the half-width of the c.i. is the result of the default precision. In general, the histogram estimator has larger variance, so better c.i. coverage. However, the histogram estimators are biased high around the tail area. This is because the histogram estimators often result in a bounded distribution, i.e., the tail of the distribution is truncated. With  $\alpha = 0.10$ , the independent sequences will fail the test of independence 10% of the times. The average sample sizes, 666 and 667, are close to the theoretical value, i.e.,  $\sum_{i=0}^{\infty} n_0 \alpha^i$ , where  $n_0 = 600$ .

Figures 2 and 3, respectively, show the empirical and true densities of the tri-modal and exponential distributions, generated from the first run of our experiments. The exponential distribution has a steep slope, so has a smaller window width  $h$  and has a more ragged empirical distribution curve.

We tested the following correlated sequences:

- Observations are from the AR1 (first-order autoregressive) process:

$$X_i = \mu + \varphi(X_{i-1} - \mu) + \varepsilon_i \text{ for } i = 1, 2, \dots,$$

where

$$E(\varepsilon_i) = 0, \quad E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Table 1: Coverage of 90% Confidence Density Estimators for the Tri-modal Distribution

avg <i>N</i>	666			
stdev <i>N</i>	119			
<i>x</i>	0.5	5	7.5	10
<i>f(x)</i>	0.1226	0.0665	0.0363	0.1360
Histogram Estimator (0.001856, 0.000756)				
$\hat{f}(x)$	0.1193	0.0661	0.0373	0.1308
coverage	88.1%	91.5%	91.0%	84.3%
avg $\gamma$	0.0515	0.0677	0.1007	0.0536
stdev $\gamma$	0.0389	0.0503	0.0749	0.0391
avg hw	0.0186	0.0150	0.0115	0.0192
stdev hw	0.0098	0.0078	0.0061	0.0106
Kernel Estimator (0.001207, 0.000530)				
$\hat{f}(x)$	0.1155	0.0651	0.0392	0.1252
coverage	78.7%	88.7%	86.6%	63.3%
avg $\gamma$	0.0621	0.0574	0.1003	0.0808
stdev $\gamma$	0.0395	0.0418	0.0731	0.0442
avg hw	0.0147	0.0118	0.0090	0.0151
stdev hw	0.0076	0.0060	0.0048	0.0117

Table 2: Coverage of 90% Confidence Density Estimators for the Expon(1) Distribution

avg <i>N</i>	667			
stdev <i>N</i>	124			
<i>x</i>	0.5	2.0	3.0	5.0
<i>f(x)</i>	0.6065	0.1353	0.0498	0.0067
Histogram Estimator (0.007100, 0.003581)				
$\hat{f}(x)$	0.6085	0.1359	0.0498	0.0088
coverage	90.4%	91.5%	90.0%	73.2%
avg $\gamma$	0.0443	0.0950	0.1643	0.3612
stdev $\gamma$	0.0340	0.0699	0.1240	0.7372
avg hw	0.0885	0.0426	0.0263	0.0057
stdev hw	0.0491	0.0226	0.0137	0.0141
Kernel Estimator (0.003114, 0.001654)				
$\hat{f}(x)$	0.6116	0.1363	0.0502	0.0076
coverage	88.7%	90.4%	89.1%	86.5%
avg $\gamma$	0.0316	0.0759	0.1304	0.4824
stdev $\gamma$	0.0248	0.0563	0.0989	2.9082
avg hw	0.0627	0.0339	0.0207	0.0100
stdev hw	0.0343	0.0182	0.0109	0.0572

$$-1 < \varphi < 1.$$

The  $\varepsilon_i$ 's are commonly called *error terms*.

- Observations are from the M/M/1 queueing model.

The AR1 process shares many characteristics observed in simulation output processes, including asymptotic first- and second-order stationarity, and autocorrelations that decline exponentially with increasing lag. If we make the additional assumption that the  $\varepsilon_i$ 's are normally distributed, since we have already assumed that they are uncorrelated,

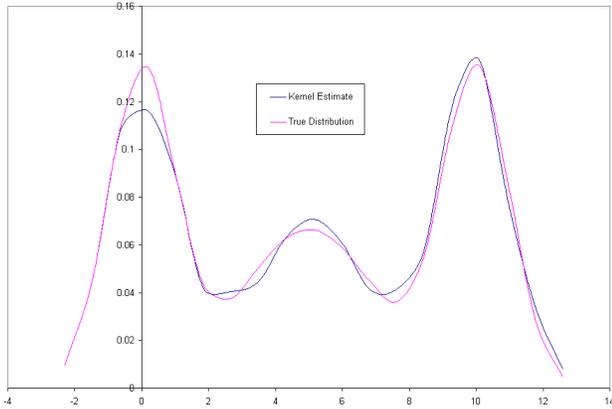


Figure 2: Empirical Density of the Tri-modal Distribution

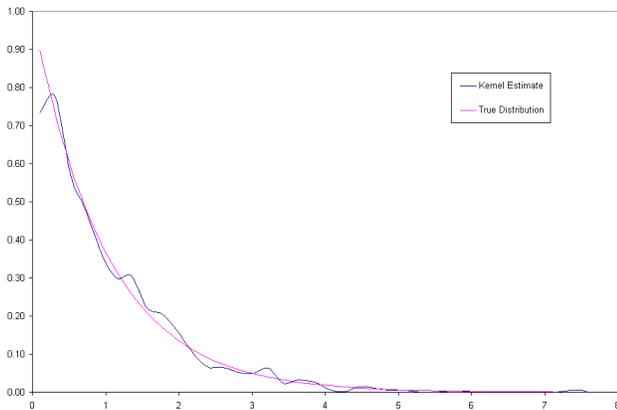


Figure 3: Empirical Density of the Exponential Distribution

they will now be independent as well, i.e., the  $\varepsilon_i$ 's are i.i.d.  $\mathcal{N}(0, 1)$ , where  $\mathcal{N}(\mu, \sigma^2)$  denotes a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . It can be shown that  $X$  has asymptotically a  $\mathcal{N}(0, \frac{1}{1-\phi^2})$  distribution, and the steady-state variance constant of the AR(1) process is  $1/(1-\phi)^2$ . We set  $\phi$  to 0.90 and set  $\mu$  to zero for this experiment. In order to eliminate the initial bias,  $X_0$  is set to a random variate drawn from the steady-state distribution.

Table 3 lists the experimental results of the AR1 process. The c.i. coverage of these four design points are around the specified 90% confidence level for both estimators. The simulation run length generally increases as the correlation coefficient  $\phi$  of the AR(1) process increases. The run length of the AR1 process with  $\phi = 0.9$  is much larger than independent sequences and consequently much smaller  $\overline{IMSE}$ .

The waiting-time density of the stationary M/M/1 delay in queue is  $f(x) = (v - \lambda) \frac{\lambda}{v} e^{-(v-\lambda)x}$  for  $x \geq 0$  and is discontinuous at  $x = 0$ , where  $\lambda$  is the arrival rate and  $v$  is the service rate. Summary of our experimental results of the M/M/1 delay-in-queue process is in Table 4. Except

for the kernel density estimate of  $x = 0.5$ , the c.i. coverages are above or close to the specified 90%. The kernel method encounters difficulty when estimating  $f(0.5)$  for the M/M/1 queueing process, because the value 0.5 is close to the discontinuity point 0. To deal with this difficulty, various adaptive methods have been proposed; see Silverman (1986, pp. 19-29).

Figures 4 and 5, respectively, show the empirical distributions of the AR1 process with  $\phi = 0.9$  and the M/M/1 delay-in-queue process with  $\rho = 0.90$ , generated from the first run of our experiments. The theoretical steady-state distribution of this AR1 process and this M/M/1 queueing process are, respectively,  $N(0, 1/0.18)$  and  $1 - 0.9e^{-0.1x}$ , where  $x \geq 0$ . Again, our experimental results show that these density estimates provide an excellent approximation of the underlying steady-state distributions. However, the kernel estimator over-smoothes the density around the discontinuity point.

## 5 CONCLUSIONS

We have presented two algorithms for estimating the density  $f(x)$  of a stationary process. Since, to obtain the kernel estimator, the procedure needs to compute the histogram estimator, a prudent course is to choose any reasonable estimate based on these two estimates that are consistent with prior belief about the true sampling density. However, the histogram procedure is more suitable as a generic density-estimation procedure since it requires less computation, delivers a valid c.i., and has no difficulty estimating the density around a bounded tail or discontinuity point, though the  $\overline{IMSE}$  is generally larger.

Some density estimates require more observations than others before the asymptotics necessary for density estimates become valid. Our algorithm works well in determining the required simulation run length for the asymptotic approximation to become valid. The results from our empirical experiments show that the procedure is excellent in achieving the pre-specified accuracy. Our proposed histogram-approximation algorithm computes quantiles only at grid points and uses Lagrange interpolation to estimate the density at certain points. The algorithm also generates an empirical distribution (histogram) of the output sequence, which can provide insights into the underlying stochastic process.

Our approach has the desirable properties that it is a sequential procedure and it does not require users to have *a priori* knowledge of values that the data might assume. This allows the user to apply this method without having to execute a separate pilot run to determine the range of values to be expected, or guess and risk having to re-run the simulation. The main advantage of our approach is that by using a straightforward test of independence to determine the simulation run length and obtain quantiles at grid points,

we can apply classical statistical techniques directly and do not require more advanced statistical theory, thus making it easy to understand, and simple to implement.

Table 3: Coverage of 90% Confidence Density Estimators for the AR1(0.9) Process

avg $N$	20965			
stdev $N$	3836			
$x$	-0.5	0.0	1.0	2.0
$f(x)$	0.1698	0.1739	0.1581	0.1189
Histogram Estimator (0.000221, 0.000133)				
$\hat{f}(x)$	0.1697	0.1737	0.1579	0.1190
coverage	89.6%	89.7%	89.1%	90.4%
avg $\gamma$	0.0138	0.0132	0.0147	0.0192
stdev $\gamma$	0.0107	0.0105	0.0111	0.0148
avg hw	0.0075	0.0076	0.0077	0.0074
stdev hw	0.0041	0.0042	0.0041	0.0038
Kernel Estimator (0.000209, 0.000128)				
$\hat{f}(x)$	0.1696	0.1737	0.1578	0.1189
coverage	88.7%	88.9%	90.7%	90.4%
avg $\gamma$	0.0142	0.0133	0.0151	0.0190
stdev $\gamma$	0.0107	0.0103	0.0110	0.0142
avg hw	0.0074	0.0076	0.0077	0.0074
stdev hw	0.0041	0.0042	0.0041	0.0038

Table 4: Coverage of 90% Confidence Density Estimators for the MM1(0.9) Process

avg $N$	418496			
stdev $N$	92729			
$x$	0.5	2.5	5.0	10.0
$f(x)$	0.0856	0.0701	0.0546	0.0331
Histogram Estimator (0.000021, 0.000020)				
$\hat{f}(x)$	0.0833	0.0704	0.0547	0.0331
coverage	85.0%	90.8%	92.6%	91.2%
avg $\gamma$	0.0295	0.0102	0.0082	0.0093
stdev $\gamma$	0.0232	0.0083	0.0064	0.0070
avg hw	0.0050	0.0023	0.0015	0.0010
stdev hw	0.0034	0.0013	0.0008	0.0006
Kernel Estimator (0.0000021, 0.000017)				
$\hat{f}(x)$	0.1657	0.0701	0.05460	0.0331
coverage	0.0%	91.2%	91.3%	90.4%
avg $\gamma$	0.9355	0.0086	0.0082	0.0096
stdev $\gamma$	0.0291	0.0063	0.0063	0.0076
avg hw	0.0066	0.0020	0.0015	0.0010
stdev hw	0.0037	0.0011	0.0008	0.0006

REFERENCES

Billingsley, P. 1999. *Convergence of probability measures*. 2nd ed. New York: John Wiley & Sons.

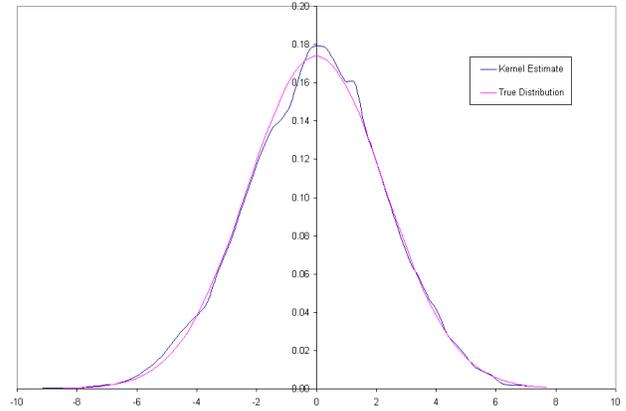


Figure 4: Empirical Density of the AR1 Process

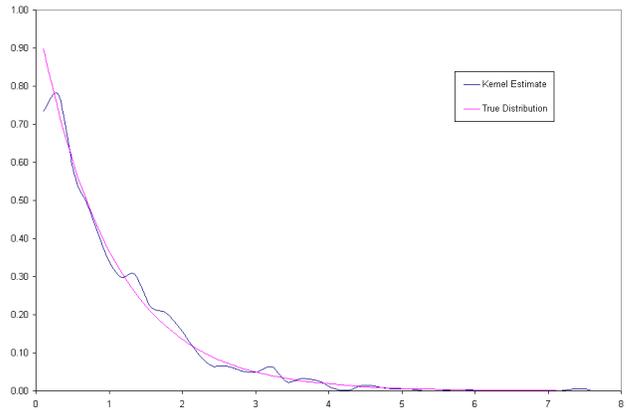


Figure 5: Empirical Density of the M/M/1 Process

Chen, E. J. 2001. Proportion estimation of correlated sequences. *Simulation* 76 (5): 273-276, 301-304.

Chen, E. J., and W. D. Kelton. 2003. Determining simulation run length with the runs test. *Simulation Modelling Practice and Theory* 11 (3-4): 237-250.

Chen, E. J., and W. D. Kelton. 2006. Estimating steady-state distributions via simulation-generated histograms. *Computer and Operations Research*. To Appear.

Devroye, L., and L. Györfi. 1985. *Nonparametric density estimation: the  $L_1$  view*. New York: John Wiley & Sons.

Hogg, R. V., and A. T. Craig. 1995. *Introduction to mathematical statistics*. Fifth Edition. Englewood Cliffs, New Jersey: Prentice Hall.

Knuth, D. E. 1998. *The art of computer programming*. Vol. 2. 3rd ed. Reading, Mass.: Addison-Wesley.

Rosenblatt, M. 1971. Curve estimates. *Annals of Mathematical Statistics* 42:1815-1842.

- Scott, D. W., and L. E. Factor. 1981. Monte carlo study of three data-based nonparametric probability density estimators. *Journal of the American Statistical Association* 76 (373):9–15.
- Sen, P. K. 1972. On the Bahadur representation of sample quantiles for sequences of  $\phi$ -mixing random variables. *Journal of Multivariate Analysis* 2 (1):77–95.
- Silverman, B. W. 1986. *Density estimation for statistics and data analysis*. New York: Chapman and Hall.
- von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference and the variance. *Annals of Mathematical Statistics* 12:367–395.

#### AUTHOR BIOGRAPHIES

**E. JACK CHEN** is a Senior Staff Specialist with BASF Corporation. He received a Ph.D. from the University of Cincinnati. His research interests are in the area of computer simulation. His email address is <e.jack.chen@basf.com>.

**W. DAVID KELTON** is a Professor in the Department of Quantitative Analysis and Operations Management at the University of Cincinnati. He received a B.A. in mathematics from the University of Wisconsin-Madison, an M.S. in mathematics from Ohio University, and M.S. and Ph.D. degrees in industrial engineering from Wisconsin. His research interests and publications are in the probabilistic and statistical aspects of simulation, applications of simulation, and stochastic models. Currently, he serves as Editor-in-Chief of the *INFORMS Journal on Computing*, and has been Simulation Area Editor for *Operations Research*, the *INFORMS Journal on Computing*, and *IIE Transactions*, as well as Associate Editor for *Operations Research*, the *Journal of Manufacturing Systems*, and *Simulation*. From 1991 to 1999 he was the INFORMS co-representative to the Winter Simulation Conference Board of Directors and was Board Chair for 1998. In 1987 he was Program Chair for the WSC, and in 1991 was General Chair. His email and web addresses are <david.kelton@uc.edu> and <www.cba.uc.edu/faculty/keltonwd>.