Proceedings of the 1986 Winter Simulation Conference
J. Wilson, J. Henriksen, S. Roberts (eds.)

SAMPLE-BASED NON-UNIFORM RANDOM VARIATE GENERATION

Luc Devroye
School of Computer Science
McGill University

ABSTRACT.

A sample of n iid random variables with a given unknown density is given. We discuss several issues related to the problem of generating a new sample of iid random variables with almost the same density. In particular, we look at sample independence, consistency, sample indistinguishability, moment matching and generator efficiency. We also introduce the notion of a replacement number, the minimum number of observations in a given sample that have to be replaced to obtain a sample with a given density.

1. INTRODUCTION.

Assume that we are given a sample $X_1, \ldots, X_n$ of iid R^d-valued random vectors with common unknown density f, and that we are asked to generate a new independent sample $Y_1, \ldots, Y_m$ of independent random vectors with the same density f. This is quite an impossible task since f is usually not known. The purpose of this note is to discuss just what can be done, and how close we can come to generating a perfect sample.

What one can do is construct a **density estimate** $f_n(x) = f_n(x, X_1, \ldots, X_n)$ of $f(x)$, and then generate a sample of size m from f_n. This procedure has several drawbacks: first of all, f_n is typically not equal to f. Also, the new sample depends upon the original sample. Yet, we have very few options available to us. Ideally, we would like the new sample to appear to be distributed as the original sample. This will be called sample indistinguishability. This and other issues will be discussed in this section. Some of this material appeared originally in Devroye and Gyorfi (1985, chapter 8) and Devroye (1986).

2. SAMPLE INDEPENDENCE.

There is little that can be done about the dependence between $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ except to hope that for n large enough, some sort of asymptotic independence is obtained. In some applications, sample independence is not an issue at all.

Since the Y_i's are conditionally independent given $X_1, \ldots, X_n$, we need only consider the dependence between Y_1 and $X_1, \ldots, X_n$. A measure of the dependence is

$$D_n \overset{\Delta}{=} \sup_{A,B} | P(Y \in A, X \in B) - P(Y \in A) P(X \in B) | ,$$

where the supremum is with respect to all Borel sets A of R^d and all Borel sets B of R^{nd}, and where $Y = Y_1$ and X is our shorthand notation for $(X_1, \ldots, X_n)$. We say that the samples are asymptotically independent when

$$\lim_{n \to \infty} D_n = 0 .$$

In situations in which $X_1, \ldots, X_n$ is used to design or build a system, and $Y_1, \ldots, Y_m$ is used to test it, the sample dependence often causes optimistic evaluations. Without the asymptotic independence, we can't even hope to diminish this optimistic bias by increasing n.

The inequality in Theorem 1 below provides us with a sufficient condition for asymptotic independence. First, we need the following Lemma.

Lemma 1. (Scheffe, 1947).

For all densities f and g on R^d,

$$\int | f - g | = 2 \sup_B | \int_B f - \int_B g |$$

where the supremum is with respect to all Borel sets B of R^d.

Scheffe's lemma tells us that if we assign probabilities to sets (events) using two different densities, then the maximal difference between the probabilities over all sets is equal to one half of the L_1 distance between the densities. From Lemma 1, we obtain

Theorem 1.

Let f_n be a density estimate, which itself is density. Then

$$D_n \leq E(\int | f_n - f |) .$$

Proof of Theorem 1.

See Devroye and Gyorfi (1985). ■

We see that for the sake of asymptotic sample independence, it suffices that the expected L_1 distance between f_n and f tends to zero with n. This is also called **consistency**. Asymptotic independence does not imply consistency: just let f_n be the uniform density in all cases, and observe that $D_n \equiv 0$, yet $\int | f_n - f |$ is a positive constant for all n and all nonuniform f.

3. CONSISTENCY OF DENSITY ESTIMATES.

A density estimate f_n is **consistent** if for all densities f,

$$\lim_{n \to \infty} E (\int | f_n - f |) = 0 .$$

Consistency guarantees that the expected value of the maximal error committed by replacing probabilities defined with f with probabilities defined with f_n tends to 0. Many estimates are consistent, see e.g. Devroye and Gyorfi (1985). Parametric estimates, i.e. estimates in which the form of f_n is fixed up to a finite number of parameters, which are estimated from the sample, cannot be consistent because f_n is required to converge to f for all f, not a small subclass. Perhaps the best known and most widely used consistent density estimate is the **kernel estimate**

$$f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K (\frac{x - X_i}{h}) ,$$

where K is a given density (or kernel), chosen by the user, and $h > 0$ is a smoothing parameter, which typically depends upon n or the data (Rosenblatt, 1956; Parzen, 1962). For consistency it is necessary and sufficient that $h \to 0$ and $nh^d \to \infty$ in probability as $n \to \infty$ (Devroye and Gyorfi, 1985). How one should choose h as a function of n or the data is the subject of a lot of controversy. Usually, the choice is made based upon the approximate minimization of an error criterion. Sample independence (Theorem 1) and sample indistinguishability (next section) suggest that we try to minimize

$$E (\int | f_n - f |) .$$

But even after narrowing down the error criterion, there are several strategies. One could minimize the supremum of the criterion where the supremum is taken over a class of densities. This is called a **minimax strategy**. If f has compact support on the real line and possesses one absolutely continuous derivative and an absolutely integrable second derivative, then the best choices for individual f (i.e., not in the minimax sense) are

$$h = Cn^{-\frac{1}{5}} ,$$

$$K(x) = \frac{3}{4}(1-x^2) \quad (| x | \leq 1) ,$$

where C is a constant depending upon f only:

$$C = \left(\sqrt{\frac{15}{2\pi}} \frac{\int \sqrt{f}}{\int | f'' |} \right)^{\frac{2}{5}} .$$

The optimal kernel coincides with the optimal kernel for L_2 criteria (Bartlett, 1963). The optimal formula for h, which depends upon the unknown density f, can be estimated from the data. Alternatively, as suggested by Deheuvels (1977), one could compute the formula for a given parametric density, a rough guess of sorts, and then estimate the parameters from the data. For example, if this is done with the normal density as initial guess, we obtain the recommendation to take

$$h = \left(\frac{15e \sqrt{2\pi}}{8n} \right)^{\frac{1}{5}} \hat{\sigma} ,$$

where $\hat{\sigma}$ is a robust estimate of the standard deviation of the normal density (Devroye and Gyorfi, 1985). A typical robust estimate is the so-called quick-and-dirty estimate

$$\hat{\sigma} = \frac{X_{(np)} - X_{(nq)}}{x_p - x_q} ,$$

where x_p, x_q are the p-th and q-th quantiles of the standard normal density, and $X_{(np)}$ and $X_{(nq)}$ are the p-th and q-th quantiles in the data, i.e. the (np)-th and (nq)-th order statistics.

The construction given here with the kernel estimate is simple, and yields fast generators. Other constructions have been suggested in the literature with random variate generation in mind. Often, the explicit form of f_n is not given or needed. Constructions often start from an empirical distribution function based upon $X_1, \ldots, X_n$, and a smooth approximation of this distribution function (obtained by interpolation), which is directly useful in the inversion method. Guerra, Tapia and Thompson (1978) use Akima's (Akima, 1970) quasi-Hermite piecewise cubic interpolation to obtain a smooth monotone function coinciding with the empirical distribution function at the points X_i. Recall that the empirical distribution is the distribution which puts mass $\frac{1}{n}$ at point X_i. Butler (1970) on the other hand uses Lagrange's quadratic interpolation on the inverse empirical distribution function to speed random variate generation up even further.

4. SAMPLE INDISTINGUISHABILITY. THE REPLACEMENT NUMBER.

In simulations, one important qualitative measure of the goodness of a method is the indistinguishability of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_m$ for the given sample size m. Note that we have forced both sample sizes to be the same, although for the construction of f_n we keep on using n points. Let us try to quantify this notion by means of the following imbedding technique. Let A be a fixed Borel set of R^d, and let (Ω, F, P) be a probability space with the property that $(Y_1(\omega), \ldots, Y_m(\omega))$ and $(Z_1(\omega), \ldots, Z_m(\omega))$ are two sequences of iid R^d-valued random vectors with common density f_n and f respectively. The sequences **are allowed to be dependent.** For a fixed set A of R^d, let N_A and M_A be the cardinalities of A induced by the first and second sample respectively.

An appropriate measure of closeness is **the replacement number**

$$\Delta = \inf_{(\Omega, F, P)} E \left(\sup_A | N_A - M_A | \right) .$$

Here E is the conditional expectation given $X_1, \ldots, X_n$. This is different from, and stronger than, the approach taken in Devroye and Gyorfi (1985). Indeed, Δ is small if the cardinalities of all sets A are nearly equal for all A. We can consider Δ as the (conditional) expected value of the minimum number of Y_i's that should be altered and replaced by other random variables to make the sample into one that can be considered as an iid sample drawn from f. The crucial result needed here is

Theorem 2.

For any f and f_n, we have

$$\Delta = \frac{m}{2} \int | f_n - f | .$$

Proof of Theorem 2.

First, we note that

$$E\left(\sup_A | N_A - M_A | \right)$$
$$\geq \sup_A E\left(| N_A - M_A | \right)$$
$$\geq \sup_A | E(N_A) - E(M_A) |$$

(Jensen's inequality)

$$= m \sup_A | \int_A f_n - \int_A f |$$
$$= \frac{m}{2} \int | f_n - f |$$

(Scheffe's theorem) .

For the inequality in the other direction, we will use an embedding argument. The object here is to construct two dependent samples of size m each, one drawn from f, and one drawn from f_n, such that

$$\Delta \leq \frac{m}{2} \int | f_n - f | .$$

Observe that there is no hope of obtaining this with two independent samples, for $\sup_A | N_A - M_A | = 2m$ for any independent samples with densities, even if the densities are identical. The construction of the samples can be done as follows (see Devroye, 1985): define

$$\delta = \int | f_n - f | .$$

Then define the following densities:

$$f_{\min} = \frac{\min(f, f_n)}{1-\delta} ,$$
$$f_0 = \frac{f - \min(f, f_n)}{\delta} ,$$
$$g_0 = \frac{f_n - \min(f, f_n)}{\delta} .$$

Three independent samples of iid random vectors are considered:

$$U_1, U_2, \ldots, U_m \sim f_{\min}$$
$$V_1, V_2, \ldots, V_m \sim f_0$$
$$W_1, W_2, \ldots, W_m \sim g_0 .$$

In addition, let N be binomial (m, δ) and let $(\sigma_1, \ldots, \sigma_m)$ be a random permutation of $(1, \ldots, m)$, and let both N and the random permutation be independent of the three samples. Then, define

$$(Z_1, \ldots, Z_m)$$
$$= (U_1, \ldots, U_{m-N}, V_1, \ldots, V_N) ,$$
$$(Y_1, \ldots, Y_m)$$
$$= (U_1, \ldots, U_{m-N}, W_1, \ldots, W_N) .$$

We claim that

$$(Z_{\sigma_1}, \ldots, Z_{\sigma_m})$$

is an iid sample drawn from f, and that

$$(Y_{\sigma_1}, \ldots, Y_{\sigma_m})$$

is an iid sample drawn from f_n. This is based upon the mixture decomposition

$$f = (1-\delta) f_{\min} + \delta f_0 .$$

What matters is that the Z_i's and the Y_i's agree except in N components, where N is binomial (m, δ). Let E be the expected value with respect to the probability measure defined above. Then

$$E\left(\sup_A | N_A - M_A | \right)$$
$$= \frac{1}{2} E\left(\sum_{i=1}^{m} | N_{Z_i} - M_{Z_i} | \right)$$

(Scheffe's theorem)

$$\leq \frac{1}{2} E\left(\sum_{i=1}^{m} I_{Z_i \neq Y_i} \right)$$
$$\leq \frac{1}{2} E\left(N \right)$$
$$= \frac{m\ \delta}{2} . \blacksquare$$

The fact that Δ is precisely equal to $\frac{m}{2} \int | f_n - f |$ will allow us to associate numbers with Δ. It also shows the importance of taking a density estimate f_n which is close to f in the L_1 sense. This is why we have concentrated thus far on the kernel estimate, and not on its ancestor, the histogram estimate. It should be noted that the kernel estimate is very flexible, and can be adapted to many situations. However, there are certain limitations. To cite two typical (negative) results, we have

A. $$\inf_{f, h, K} E\left(\int | f_n - f | \right)$$
$$\geq \frac{1}{\sqrt{512\ n} \sqrt{1 + \frac{1}{32\ n}}} ;$$

B. $$\inf_{f, h, K \geq 0} E\left(\int | f_n - f | \right)$$
$$\geq (0.86 + o(1)) n^{-\frac{2}{5}} .$$

The difference between these results is that in the former case, the infimum is over all integrable K, even kernels taking negative values, while in the second case, the infimum is over all kernels that are densities. Both bounds however are valid for all f. This makes them very useful for determining whether the sample size is large enough for the kernel estimate. As a rule of thumb, when $K \geq 0$, we have

$$E(\Delta) \geq 0.42\ m\ n^{-\frac{2}{5}}\ .$$

This gives an idea of what kind of accuracy we can expect. A small table of approximative lower bounds for $E(\Delta)$ is provided below.

n:	1	10	100	1000	10000	100000
m:						
1	0.42	0.167	0.066	0.026	0.010	0.0042
10	4.2	1.67	0.66	0.26	0.010	0.042
100	42	16.7	6.6	2.6	1	0.42
1000	420	167	66	26	10	4.2
10000	4200	1670	660	260	100	42
100000	42000	16700	6600	2600	1000	420

If we could attain this lower bound, then given an original sample of size $n=10000$, we could generate $m=10000$ Y_i's with the property that if we could alter about 100 of the Y_i values, we would in fact obtain a sample that is exactly distributed. Most of the time, tables of this nature can be used to determine whether the lower bound for $E(\Delta)$ is acceptable.

On the positive side, we should mention that for many densities, $E(\int | f_n - f\ |) = O(n^{-2/5})$. This is true whenever f has a finite $1+\epsilon$-th moment for some $\epsilon > 0$, and f and f' are absolutely continuous, and f'' is absolutely integrable. For precise information about the rates, consult Devroye and Gyorfi (1985).

We finally mention that Δ cannot oscillate a lot about its mean for any kernel estimate. We have for any boxed kernel (i.e., bounded kernel of compact support, integrating to one),

A. $\sup_{h,f} Var\left(\Delta\right) \leq \frac{Cm^2}{n}$ for some universal constant C depending upon K only.

B. $\sup_{h,f} P(\ |\Delta - E(\Delta)\ | > \frac{um}{\sqrt{n}})$ $\leq e^{-C^* u}$ for some constant C^* depending upon K only, and all $u > 0$.

Both results are valid uniformly over all densities f (Devroye, 1986). Together with (upper or lower) bounds for $E(\Delta)$ they can be used to derive distribution-free confidence intervals for Δ. They also imply that $\Delta / E(\Delta) \rightarrow 1$ in probability for most f_n (at least those for which $\sqrt{n} E(\int | f_n - f\ |) \rightarrow \infty$).

5. MOMENT MATCHING.

Some statisticians attach a great deal of importance to the moments of the densities f_n and f. For $d=1$, the i-th **moment mismatch** is defined as

$$M_{n,i} = \int x^i f_n - \int x^i f$$

$$(i=1,2,3,\ldots)\ .$$

Clearly, $M_{n,i}$ is a random variable. Assume that we employ the kernel estimate with a zero mean finite variance (σ^2) kernel K. Then, we have

$$M_{n,1} = \frac{1}{n} \sum_{i=1}^{n} (X_i - E(X_i))\ ,$$

$$M_{n,2} = \frac{1}{n} \sum_{i=1}^{n} (X_i^2 - E(X_i^2)) + h^2 \sigma^2\ .$$

This follows from the fact that f_n is an equiprobable mixture of densities K shifted to the X_i's , each having variance $h^2\sigma^2$ and zero mean. It is interesting to note that the distribution of $M_{n,1}$ is not influenced by h or K. By the weak law of large numbers, $M_{n,1}$ tends to 0 in probability as $n \rightarrow \infty$ when f has a finite first moment. The story is different for the second moment mismatch. Whereas $E(M_{n,1})=0$, we now have $E(M_{n,2}) = h^2\sigma^2$, a positive bias. Fortunately, h is usually small enough so that this is not too big a bias. Note further that the variances of $M_{n,1}$, $M_{n,2}$ are equal to

$$\frac{Var(X_1)}{n}\ ,\ \frac{Var(X_1^2)}{n}$$

respectively. Thus, h and K only affect the bias of the second order mismatch. Making the bias very small is not recommended as it increases the expected L_1 error, and thus the sample dependence and distinguishability.

6. GENERATORS FOR f_n.

For the kernel estimate, generators can be based upon the property that a random variate is distributed as an equiprobable mixture, as is seen from the following trivial algorithm.

Mixture method for kernel estimate

```
Generate $Z$, a random integer uniformly distributed on $\{1,2, . . . , n\}$.
Generate a random variate $W$ with density $K$.
RETURN $X_Z + hW$
```

For Bartlett's kernel $K(x) = \frac{3}{4}(1-x^2)_+$, we suggest either rejection or a method based upon properties of order statistics:

Generator based upon rejection for Bartlett's kernel

```
REPEAT
    Generate a uniform $[-1,1]$ random variate $X$ and an independent uniform $[0,1]$ random variate $U$.
UNTIL $U \leq 1-X^2$
RETURN $X$
```

The order statistics method for Bartlett's kernel

```
Generate three iid uniform [-1,1] random variates $V_1,V_2,V_3$.
IF $|V_3| > \max(|V_1|,|V_2|)$
    THEN RETURN $X \leftarrow V_2$
    ELSE RETURN $X \leftarrow V_3$
```

In the rejection method, X is accepted with probability 2/3, so that the algorithm requires on average three independent uniform random variates. However, we also need some multiplications. The order statistics method always uses precisely three independent uniform random variables, but the multiplications are replaced by a few absolute value operations.

Sometimes, K takes negative values, but integrates to one. The density estimate is

$$f_n(x) = c \left(\frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x-X_i}{h}) \right)_+$$

where c is a normalization constant. Since

$$f_n(x) \le g_n(x)$$
$$\overset{\Delta}{=} \frac{c}{nh^d} \sum_{i=1}^{n} K_+(\frac{x-X_i}{h}) ,$$

the rejection method can be altered slightly:

Generator based upon the rejection method

```
REPEAT
    Generate $Z$, a random integer uniformly distributed on $\{1,2, . . . , n\}$.
    Generate a random variate $W$ with density $K_+/\int K_+$.
    $Y \leftarrow X_Z + hW$.
    Generate a uniform [0,1] random variate $U$.
    Accept $\leftarrow [U g_n(Y) \le f_n(Y)]$ .
UNTIL Accept
RETURN $Y$.
```

The expected number of iterations is $\int K_+$. For fast evaluation of g_n / f_n, it is perhaps best to use a hash structure for the X_i's, or, when K is polynomial with compact support, to compute the piecewise polynomial forms of f_n and g_n, and to locate intervals by binary search trees in time $O(\log(n))$. In the latter case, the expected time of the algorithm is $O(\log(n))$.

We finally mention that for piecewise polynomial nonnegative K, the inversion method can be implemented without a lot of trouble. This has the advantage that the correlation between observations can be better controlled (see Bratley, Fox and Schrage (1983) for a discussion).

7. HISTOGRAMS.

Data-based histograms have been suggested for random variate generation by several authors. Bratley, Fox and Schrage (1983) use it to generalize data in a manner that would make the inversion method easily applicable. This is based upon the fact that the distribution function of every histogram is piecewise linear. The alias method can be used in general to obtain fast inversion algorithms (see Walker (1977), Chen and Asau (1974), Ahrens and Kohrt (1981), Kronmal and Peterson (1979) and Devroye and Gyorfi (1985)). Archer (1980) is mainly concerned with moment matching in his definition of a data-based histogram. Scott (1979) and others discuss the issue of choosing the bin width in equi-spaced histograms.

The difference between an ordinary histogram and a data-based histogram is related to the definition of the height of the histogram in each interval. For a data-based histogram, with intervals A_n, the height on the interval A_n is

$$\frac{\text{number of points in } A_n}{n \times \text{length of } A_n} .$$

Generators for these kinds of histograms are easy to define. Associate with each data point X_i the interval coordinates (L_i, R_i) of the interval to which X_i belongs. Thus, the storage is $2n$. Then proceed as follows:

Generator for a data-based histogram.

```
Generate a uniform $\{1, . . . , n\}$-valued random integer $Z$.
Generate a uniform [0,1] random variate $U$.
RETURN $L_Z + U(R_Z - L_Z)$.
```

The data points could be rearranged in a preprocessing step according to membership in the same intervals, e.g. by sorting. This can be used to reduce the storage. What is more important than storage and speed however is the consistency of the underlying density estimate. For example, if the bins are defined by the order statistics (so that each bin has precisely one data point), the estimate is not consistent for any f . The best one can hope for with a continuous density f is $E(\int | f_n - f |) = O(n^{-1/3})$ (which is worse than the best achievable rate with the kernel estimate). See e.g. Scott (1979) or Devroye and Gyorfi (1985).

Some data-based histograms have interesting optimality properties. To illustrate this, consider Grenander's estimate (Grenander, 1956) for monotone densities on $[0,\infty)$. The monotone density f_n maximizing the likelihood product for $X_1, . . . , X_n$ is a data-based histogram with breakpoints at some order statistics. The order statistics correspond to the points at which the smallest concave majorant of the empirical distribution function touches the empirical distribution function. These are the points of contact obtained by putting a large elastic band around the empirical distribution function and letting go. It is known that for many monotone densities the expected L_1 error tends to zero as $n^{-1/3}$. For example, if f has two bounded continuous derivatives, it is about

$$0.82^{-\frac{1}{3}} \int (\frac{1}{2} f \; | f' |)^{\frac{1}{3}}$$

(Groeneboom, 1983).

References

J. H. Ahrens and K. D. Kohrt, "Computer methods for efficient sampling from largely arbitrary statistical distributions," *Computing*, vol. 26, pp. 19-31, 1981.

H. Akima, "A new method of interpolation and smooth curve fitting based on local procedures," *Journal of the ACM*, vol. 17, pp. 589-602, 1970.

N. P. Archer, "The generation of piecewise linear approximations of probability distribution functions," *Journal of Statistical Computation and Simulation*, vol. 11, pp. 21-40, 1980.

M. S. Bartlett, "Statistical estimation of density functions," *Sankhya Series A*, vol. 25, pp. 245-254, 1963.

P. Bratley, B. L. Fox, and L. E. Schrage, *A Guide to Simulation*, Springer-Verlag, New York, N.Y., 1983.

E. L. Butler, "Algorithm 370. General random number generator," *Communications of the ACM*, vol. 13, pp. 49-52, 1970.

H. C. Chen and Y. Asau, "On generating random variates from an empirical distribution," *AIIE Transactions*, vol. 6, pp. 163-166, 1974.

P. Deheuvels, "Estimation non parametrique de la densite par histogrammes generalises," *Revue de Statistique Appliquee*, vol. 25, pp. 5-42, 1977.

L. Devroye, "The equivalence of weak, strong and complete convergence in *Lsub* 1 for kernel density estimates," *Annals of Statistics*, vol. 11, pp. 896-904, 1983.

L. Devroye and C. S. Penrod, "Distribution-free lower bounds in density estimation," *Annals of Statistics*, vol. 12, pp. 1250-1262, 1984.

L. Devroye and L. Gyorfi, *Nonparametric Density Estimation. The L_1* View, John Wiley, New York, N.Y., 1985.

L. Devroye, "A note on the L_1 consistency of variable kernel estimates," *Annals of Statistics*, vol. 13, pp. 1041-1049, 1985.

L. Devroye, "A universal lower bound for the kernel estimate," , 1986. Submitted.

L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.

L. Devroye, "The kernel estimate is relatively stable," , 1986. Submitted.

U. Grenander, "On the theory of mortality measurement. Part II," *Skandinavisk Aktuarietidskrift*, vol. 39, pp. 125-153, 1956.

P. Groeneboom, "Estimating a monotone density," in *Proceedings of the Neyman-Kiefer Conference*, 1983.

V. O. Guerra, R. A. Tapia, and J. R. Thompson, "A random number generator for continuous random variables based on an interpolation procedure of Akima," *Proceedings of the 1978 Winter Simulation Conference*, pp. 228-230, 1978.

R. A. Kronmal and A. V. Peterson, "On the alias method for generating random variables from a discrete distribution," *The American Statistician*, vol. 33, pp. 214-218, 1979.

E. Parzen, "On the estimation of a probability density function and the mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065-1076, 1962.

M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, vol. 27, pp. 832-837, 1956.

H. Scheffe, "A useful convergence theorem for probability distributions," *Annals of Mathematical Statistics*, vol. 18, pp. 434-458, 1947.

D. W. Scott, "Optimal data-based histograms," *Biometrika*, vol. 66, pp. 605-610, 1979.

A. J. Walker, "An efficient method for generating discrete random variables with general distributions," *ACM Transactions on Mathematical Software*, vol. 3, pp. 253-256, 1977.