ORSA Journal on Computing
Vol. 6, No. 4, Fall 1994

0899-1499/94/0604-0342 $01.25

A Flexible Method for Estimating Inverse Distribution Functions in Simulation Experiments

ATHANASSIOS N. AVRAMIDIS / *SABRE Decision Technologies, Dallas / Fort Worth Airport, TX 75261-9616; Email: avramidi@sdt.com*

JAMES R. WILSON / *Department of Industrial Engineering, North Carolina State University, Raleigh, NC 27695-7906, Email: jwilson@eos.ncsu.edu*

(Received: March 1990; revised: April 1992, March 1993; accepted: August 1993)

To generate random variates from an unknown continuous distribution, we present procedure IDPF — a flexible technique for estimating the associated inverse distribution function from sample data and for generating variates from the fitted distribution by inversion. To motivate IDPF, first we examine a predecessor due to Hora, and we explain how Hora's method can fail in either the distribution-fitting or variate-generation stage of application. We apply IDPF as follows. After selecting an initial inverse distribution function by a standard technique, we estimate a polynomial "filter" for the random-number input by constrained nonlinear regression to achieve minimum "distance" between the empirical inverse distribution and the final fitted inverse distribution obtained by composition of the initial inverse distribution with the polynomial "filter." The regression constraint ensures that the fitted inverse distribution function is nondefective and monotonically nondecreasing. A portable, public-domain implementation of IDPF is based on well-known techniques for selecting initial distributions from the Johnson translation system. A Monte Carlo study illustrates the effectiveness of IDPF. Compared to initial Johnson distributions selected by matching moments, IDPF-based fits are closer on the average to the corresponding empirical and theoretical inverse distribution functions. Similar conclusions apply to other initial distributions selected by other methods.

In the development of discrete-event simulation models, we frequently need to generate independent observations of a continuous random variable X having an unknown cumulative distribution function (c.d.f.) $F(\cdot)$. To facilitate the use of standard variance reduction techniques, we seek to sample X by inversion; and this approach requires a suitable approximation to the inverse $F^{-1}(\cdot)$ of the target c.d.f. Typically a random sample $\{X_1, X_{i_n}, \cdots, X_n\}$ from $F(\cdot)$ is available, and this sample defines the associated empirical c.d.f. $F_n(\cdot)$. The conventional approach to simulation input modeling involves (a) identifying an appropriate family of distributions to model the behavior of X; (b) estimating the corresponding parameter values that yield the "best" fit to the sample data set; and (c) invoking some standard sampling scheme to generate observations from the fitted distribution. Most of the well-known families of distributions have a fixed number of parameters, which implies a limited variety of distributional shapes and thus a limited capability for approximating the target empirical or theoretical distributions.

Hora[13] proposed an alternative method for simulation input modeling which has received substantial attention because of its flexibility and simplicity.[17, 24] Hora's method uses the inverse of a known continuous c.d.f. $F_0(\cdot)$ (the so-called *reference distribution*) as the starting point for estimating the target inverse c.d.f. $F^{-1}(\cdot)$. Hora assumed that $F^{-1}(\cdot)$ has the functional form

$$F^{-1}(p) = F_0^{-1}\left\{\exp\left[\alpha_0 \log(p) + \sum_{k=1}^{t} \frac{\alpha_k(p^k - 1)}{k}\right]\right\} \quad \text{for all } p \in (0,1), \qquad (1)$$

where t and $F_0(\cdot)$ are suitably chosen by the modeler. This method attempts to reduce the problem of estimating an inverse c.d.f. to that of selecting a reference distribution and then performing linear regression to estimate the parameters $\{\alpha_k : k = 0, 1, \ldots, t\}$ in (1). The obvious advantage of this approach is that the statistical theory for linear regression is well known and widely applied. Hora's method is also highly flexible since it allows the introduction of an arbitrarily large number of parameters to compensate for any inadequacies in the reference fit. On the other hand, Hora's method has some drawbacks: (a) The statistical model that underlies the procedure for estimating the $\{\alpha_k\}$ has an exponentially distributed, multiplicative error term; and thus we cannot apply standard inferential procedures based on classical linear regression analysis. (b) The fitted inverse c.d.f. of the form (1) may fail to be monotonically nondecreasing; and in addition to being illogical, this condition can destroy the effectiveness of standard variance reduction techniques such as common random numbers and antithetic variates.[28] (c) The fitted inverse c.d.f. of the form (1) may be undefined for some values of p in the unit interval (0, 1); and this means that the fitted distribution is defective (dishonest) and that the inverse-transform method of variate generation will ultimately fail for this distribution.

Subject classifications: Simulation, random variable generation: method of inversion.

Starting from Hora's original concept of adjusting a reference distribution based on regression analysis of a sample data set and taking the formulation (1) as a point of departure, we propose a method for fitting an Inverse Distribution with a Polynomial "Filter" (IDPF). Given a reference c.d.f. $F_0(\cdot)$ representing an initial estimate of the unknown c.d.f. $F(\cdot)$ that is to be sampled by inversion, we seek an improved estimate of $F^{-1}(\cdot)$ based on the assumption that this inverse c.d.f. can be adequately modeled with the functional form

$$F^{-1}(p) = F_0^{-1}\left[\sum_{k=1}^{r-1} \beta_k p^k + \left(1 - \sum_{k=1}^{r-1} \beta_k\right) p^r\right] \quad \text{for all } p \in (0,1). \tag{2}$$

In the approximation (2), the coefficient estimates $\{\hat{\beta}_k : k = 1, \ldots, r-1\}$ are computed so that the corresponding inverse c.d.f. estimate $\hat{F}^{-1}(\cdot)$ minimizes an appropriately weighted sum of squared deviations of the form $[\hat{F}^{-1}(p) - F_n^{-1}(p)]^2$ taken over selected values of $p \in (0, 1)$. For a given degree r of the polynomial "filter" within the square brackets on the right-hand side of (2), the least-squares estimates $\{\hat{\beta}_k : k = 1, \ldots, r-1\}$ of the filter coefficients are computed subject to the constraint that the corresponding filter is a strictly increasing function of p for $p \in (0, 1)$. The degree r of the filter is estimated by a likelihood ratio test.

This paper is organized as follows. Section 1 contains a detailed analysis of Hora's method and its potential problems. In Section 2 we discuss the basis for procedure IDPF, and we describe a portable, public-domain implementation of this procedure using reference distributions from the Johnson translation system.[14] In Section 3 we present an example illustrating the use of IDPF for simulation input modeling. In Section 4 we summarize a comprehensive experimental performance evaluation of IDPF when reference distributions from the Johnson translation system and the triangular distribution family are obtained by matching moments. The main conclusions of this work are recapitulated in Section 5. Although this paper is based on [2], some of our results were originally presented in [3].

1. Hora's Method for Estimating a Continuous Inverse C.D.F.

1.1. Basis for Hora's Method

Hora's method for estimating $F^{-1}(\cdot)$ requires several key assumptions. A fundamental requirement is that the target c.d.f. $F(\cdot)$ and the reference c.d.f. $F_0(\cdot)$ must respectively possess continuous densities $f(\cdot)$ and $f_0(\cdot)$ with the same support; moreover $f(\cdot)$ must be differentiable except possibly at the end points of its support. Let $\{X_1, X_2, \ldots, X_n\}$ denote a random sample from $F(\cdot)$, and let $x_p \equiv F^{-1}(p)$ denote the pth quantile of this distribution for $p \in (0, 1)$. The main assumption underlying Hora's method is that x_p can be adequately represented by the functional form (1), where t and $F_0(\cdot)$ are appropriately chosen by the user. To provide a basis for his input-modeling technique, Hora defined the function

$$\gamma(p) \equiv p \frac{d}{dp} \log[F_0(x_p)] \text{ for all } p \in (0,1); \tag{1.1}$$

and it is easy to see that (1) is equivalent to the condition

$$\gamma(p) = \sum_{k=0}^{t} \alpha_k p^k \text{ for all } p \in (0,1). \tag{1.2}$$

To estimate the coefficients $\{\alpha_k : k = 0, 1, \ldots, t\}$ in (1.2) by classical linear regression analysis, Hora formulated a dependent (response) variable with appropriate asymptotic properties as $n \to \infty$. Let $X_{(i)}$ denote the ith order statistic associated with the random sample $\{X_1, \ldots, X_n\}$ so that $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$. The dependent variable derived by Hora is

$$W_i \equiv i(\log\{F_0[X_{(i+1)}]\} - \log\{F_0[X_{(i)}]\}) \quad \text{for } i = 1, 2, \ldots, n-1.$$

It can be shown that for a *fixed* set of points $\{p_j : j = 1, 2, \ldots, m\}$ in (0, 1),

$$W_{\lfloor (n+1)p_j \rfloor} \xrightarrow[n \to \infty]{D} \gamma(p_j)\epsilon_j \text{ for } j = 1, 2, \ldots, m, \tag{1.3}$$

where $\xrightarrow{D}$ denotes convergence in distribution and $\{\epsilon_j : j = 1, \ldots, m\}$ are independent exponential random variables with mean 1. See Avramidis and Wilson[4] for a complete justification of (1.3). Hora also asserted that if $\gamma(\cdot)$ is a bounded function, then the analogue of (1.3) for convergence of first-order moments also holds:

$$\lim_{n \to \infty} E[W_{\lfloor (n+1)p_j \rfloor}] = \gamma(p_j) \text{ for } j = 1, 2, \ldots, m. \tag{1.4}$$

Taking $m = n - 1$ with $p_j = j/(n+1)$ for $j = 1, 2, \ldots, m$, Hora concluded that (1.2) and (1.4) provide some basis for the statistical model

$$W_j = \left[\sum_{k=0}^{t} \alpha_k \left(\frac{j}{n+1}\right)^k\right] \epsilon_j \text{ for } j = 1, 2, \ldots, n-1; \tag{1.5}$$

and he recommended that the coefficients $\{\alpha_k\}$ in (1.5) should be estimated by ordinary least-squares regression to obtain an estimate of $F^{-1}(\cdot)$ having the form (1).

There are several gaps in the basis for this development. First, Hora provided no justification for (1.3) when both m and the points $\{p_j : j = 1, \ldots, m\}$ are allowed to vary arbitrarily with the sample size n. Second, Hora gave no proof that boundedness of $\gamma(\cdot)$ is sufficient to ensure the validity of (1.4) even in the case that m and $\{p_j : j = 1, \ldots, m\}$ are fixed for all values of n—and in general for each value of j with p_j fixed, a moment-convergence property like (1.4) requires uniform integrability of the family of random variables $\{W_{\lfloor (n+1)p_j \rfloor} : n = 1, 2, \ldots\}$ on the left-hand side of display (1.4) rather than boundedness of the limiting function $\gamma(\cdot)$ on the right-hand side of that display.[6] Finally, it is inappropriate to apply ordinary least-squares regression to the statistical model (1.5) since the error term ϵ_j in this model is nonnormal and multiplicative, while standard inferential procedures of classical regression analysis (specifically, procedures for testing hypotheses and constructing confidence regions) are based on the assumption that the error term is normal and additive.[8] As discussed in the next section, difficulties can also arise in practical applications of this method.

1.2. Potential Problems with Hora's Method

To illustrate the difficulties that can occur in practice, we discuss the application of Hora's method to a data set that arose in a simulation study of medical decision making. A random sample of $n = 80$ glucose (blood sugar) levels (expressed in milligrams per deciliter) was taken from a population of elderly diabetics enrolled in a monitoring program of a general medicine clinic. First we fitted a Johnson distribution[14] to this data set by matching the first four sample moments using a modified version of the moment-matching algorithm AS 99 of Hill, Hill and Holder.[12] (In §2.2 below, we give a brief overview of the Johnson translation system of distributions as well as the modified moment-matching procedure that was used to fit reference distributions from the Johnson system.) In Figure 1 the empirical c.d.f. $F_n(\cdot)$ for this data set is plotted as a step function, and the moment-matching Johnson c.d.f. is plotted as a dashed curve. Visual inspection of Figure 1 reveals that the moment-matching fit to this data set is reasonably close. Sample goodness-of-fit statistics also support this conclusion—the Kolmogorov-Smirnov statistic is 0.069, and the chi-squared statistic (with 4 degrees of freedom) is 8.65. We then applied Hora's method to this data set, taking the reference c.d.f. $F_0(\cdot)$ to be the Johnson c.d.f. estimated by matching moments. In Figures 1, 2, and 3, the solid smooth curve represents the fitted c.d.f. $\hat{F}(\cdot)$ obtained with Hora's method using polynomials of degree $t = 0$, 1, and 2 respectively. In each of these figures, the corresponding inverse c.d.f.'s $F_n^{-1}(\cdot)$, $F_0^{-1}(\cdot)$, and $\hat{F}^{-1}(\cdot)$ can be seen by rotating the figure counterclockwise by 90°.

By any reasonable criterion for measuring goodness of fit, none of the distributions obtained with Hora's method

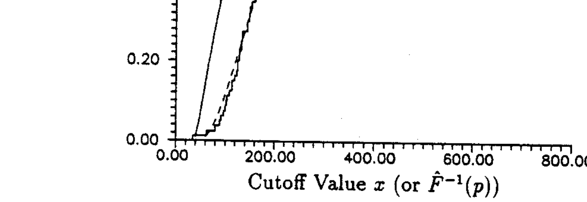

Figure 1. Glucose data set—empirical c.d.f. (step function), reference fit (dashed curve), and Hora fit with a polynomial of degree $t = 0$ (solid curve).

describes the glucose data set as well as the reference distribution; moreover, some of the fits based on Hora's method are simply unacceptable. In particular, the following anomalies are apparent:

1. With the fit of degree $t = 0$, the estimated inverse c.d.f. $\hat{F}^{-1}(p)$ differs substantially from the empirical inverse c.d.f. $F_n^{-1}(p)$ for all $p \in (0, 1)$.
2. With the fit of degree $t = 1$, the estimated inverse c.d.f. $\hat{F}^{-1}(p)$ is not even defined for all $p \in (0, 1)$. The long-dashed vertical line in Figure 2 indicates that for $0.6 < p < 1$ (roughly), the expression $\exp[\hat{\alpha}_0 \log(p) + \sum_{k=1}^{t} \hat{\alpha}_k (p^k - 1)/k]$ in display (1) is greater than one; and thus $\hat{F}^{-1}(p)$ is not defined for $p \in (0.6, 1)$. In effect the fitted Hora distribution is *defective* (or *dishonest*) so that $\lim_{x \to \infty} \hat{F}(x) < 1$; and this is clearly an unacceptable result.
3. With the fit of degree $t = 2$, the estimated inverse c.d.f. $\hat{F}^{-1}(p)$ fails to be monotonically nondecreasing for all $p \in (0, 1)$; and thus the corresponding c.d.f. $\hat{F}(x)$ is not even a single-valued function of the cutoff value x for all real x. Since monotonicity is an essential property of a legitimate inverse c.d.f., the fitted Hora distribution of degree $t = 2$ is also unacceptable.

Although plots of the fitted Hora c.d.f.'s of degrees 3 through 5 were omitted to conserve space, we remark that all of these c.d.f.'s are defective; moreover, all of the corresponding inverse c.d.f.'s fail to be monotonically nondecreasing. Kline, Bender and Nieber[17] observed similar lack of monotonicity in some of the estimated inverse c.d.f.'s they obtained by Hora's method when $n < 50$ and $t > 3$; and they attempted to avoid this problem by imposing the constraint $t \leqslant 3$ in small samples (that is, when $n < 50$). Our example demonstrates the need for a thorough examination of the potential problems with Hora's method.

To diagnose the cause of the poor performance of Hora's method in the glucose data set, we examined Figures 4, 5, and 6, which show plots of the sample points $\{[i/(n+1), W_i]^{\mathrm{T}}: i = 1, \ldots, n-1\}$ superimposed on fitted polynomials of degree t for $t = 0$, 1, and 2 respectively. Thus the solid curve in each of these figures represents the corresponding estimate of the function $\gamma(p)$ for $p \in (0, 1)$. Clearly the point $[1/81, W_1]^{\mathrm{T}}$ is an outlier. The effect of this outlier is to "pull up" the left-hand end of the estimated regression curve. In the case that $t = 0$, this outlier causes long runs of negative residuals to alternate with short runs of positive residuals as shown in Figure 4. In the cases that $t = 1$ and $t = 2$, this outlier causes long runs of negative residuals to alternate with long runs of positive residuals as shown in Figures 5 and 6 respectively. Standard techniques for analyzing the residuals in a regression suggest that all of these polynomial models for the function $\gamma(\cdot)$ are inadequate[8]; and increasing the degree of the polynomial approximation to $\gamma(\cdot)$ does *not* guarantee that a better approximation to the target inverse c.d.f. $F^{-1}(\cdot)$ will be obtained. A popular remedy for this problem is to discard outliers and perform a new regression. However, there is no clear-cut procedure for discarding outliers in this context; and the following

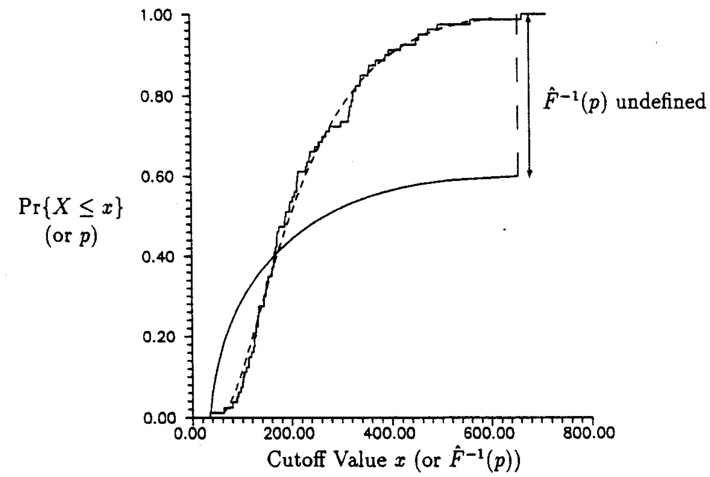

Figure 2. Glucose data set—empirical c.d.f. (step function), reference fit (dashed curve), and Hora fit with a polynomial of degree $t = 1$ (solid curve).

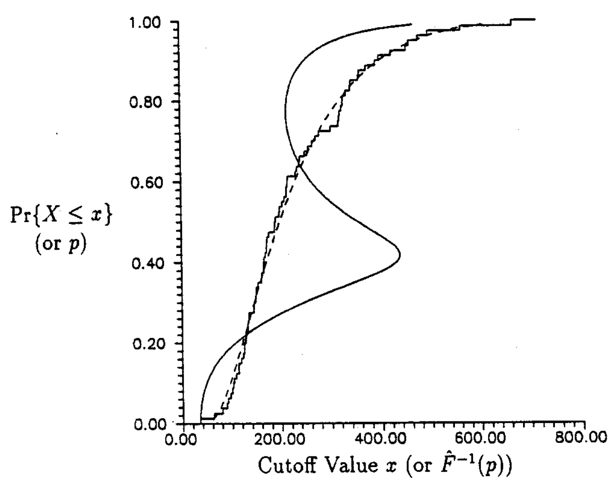

Figure 3. Glucose data set—empirical c.d.f. (step function), reference fit (dashed curve), and Hora fit with a polynomial of degree $t = 2$ (solid curve).

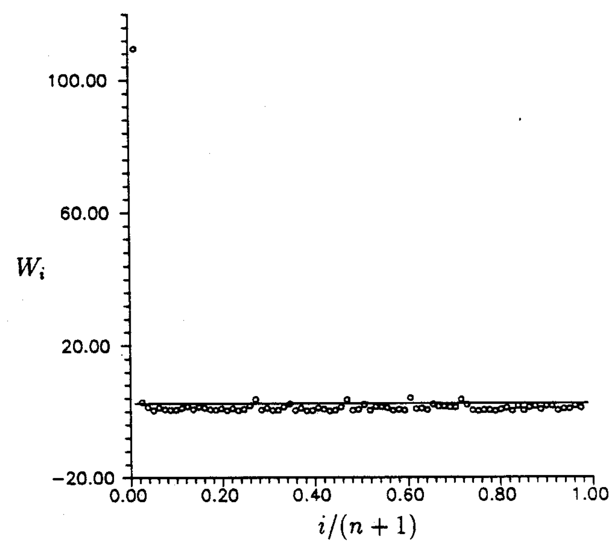

Figure 4. Polynomial of degree $t = 0$ fitted to the points $\{[i/(n+1), W_i]^{T}: i = 1, 2, \ldots, n-1\}$ for the glucose data set.

analysis shows that even after outliers have been deleted, the anomalies observed in this example can still occur.

The difficulties encountered in applying Hora's method to the glucose data set ultimately stem from the way that errors in the regression-based estimate of the function $\gamma(\cdot)$ are transformed into errors in the final estimate of the target inverse c.d.f. $F^{-1}(\cdot)$. To describe the relevant characteristics of this transformation, we let $\hat{\gamma}(p) = \sum_{k=0}^{t} \hat{\alpha}_k p^k$ denote the regression-based estimate of $\gamma(p)$ for $p \in (0, 1)$, where $\{\hat{\alpha}_k : k = 0, 1, \ldots, t\}$ are the ordinary least-squares estimates of the coefficients in the regression model (1.5). Suppose that in some nonempty open interval $(p_1, p_2) \subset (0, 1)$, we have $\hat{\gamma}(p) < 0$ for all $p \in (p_1, p_2)$. The estimate $\hat{x}_p = \hat{F}^{-1}(p)$ of the inverse c.d.f. is implicitly defined by the differential equation

$$\hat{\gamma}(p) = p \frac{d}{dp} \log[F_0(\hat{x}_p)] \text{ for all } p \in (0,1); \quad (1.6)$$

and since the functions $\log(\cdot)$ and $F_0(\cdot)$ are monotonically nondecreasing and differentiable on their respective domains, it follows from (1.6) and the chain rule that $\hat{x}_p$ is a strictly decreasing function of p in the interval (p_1, p_2). Thus the failure of Hora's method to yield an estimator of $\hat{x}_p$ that is monotonically nondecreasing for all $p \in (0, 1)$ is a direct consequence of allowing $\hat{\gamma}(p)$ to be negative over nonempty subintervals of (0, 1).

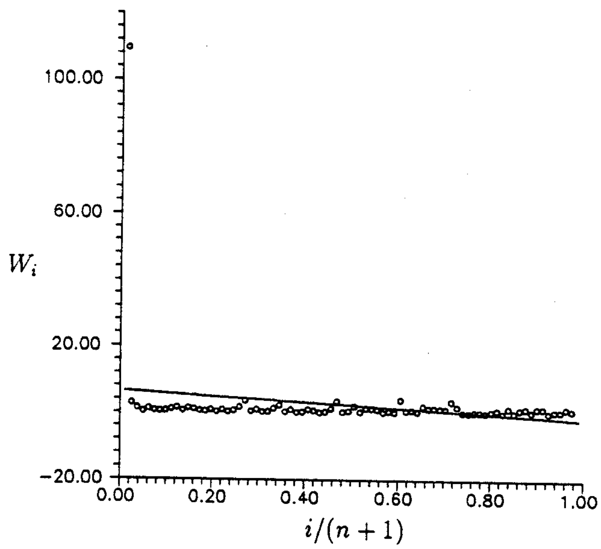

Figure 5. Polynomial of degree $t = 1$ fitted to the points $\{[i/(n+1), W_i]^{\mathrm{T}}: i = 1, 2, \ldots, n-1\}$ for the glucose data set.

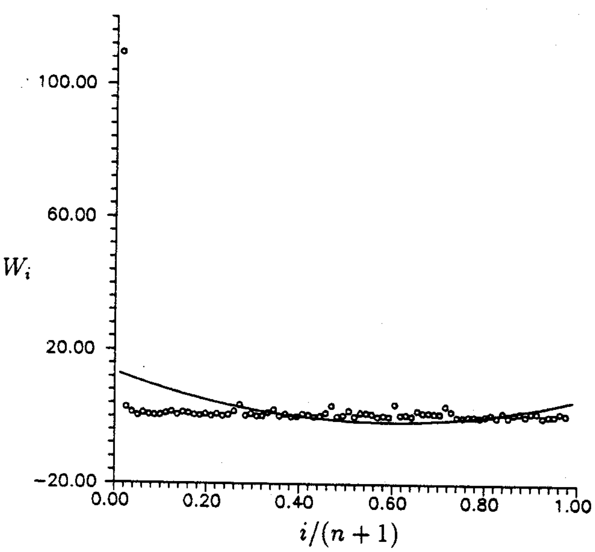

Figure 6. Polynomial of degree $t = 2$ fitted to the points $\{[i/(n+1), W_i]^{\mathrm{T}}: i = 1, 2, \ldots, n-1\}$ for the glucose data set.

The failure of Hora's method to yield a nondefective (honest) c.d.f. $\hat{F}(x)$ is also a direct consequence of allowing $\hat{\gamma}(p)$ to be negative for some $p \in (0, 1)$. Solving for $\hat{x}_p$ in equation (1.6), we have

$$\hat{F}^{-1}(p) \equiv \hat{x}_p = F_0^{-1}\left[\exp\left(-\int_p^1 \frac{\hat{\gamma}(u)}{u}\,du\right)\right] \text{ for all } p \in (0,1), \tag{1.7}$$

which yields equation (1) after carrying out the integration. If $\int_p^1 [\hat{\gamma}(u)/u]\,du < 0$ for some $p \in (0, 1)$, then the argument of $F_0^{-1}(\cdot)$ in (1.7) will be greater than one for this value of p. Thus the estimated inverse c.d.f. $\hat{F}^{-1}(p)$ will be undefined for all values of p such that $\int_p^1 [\hat{\gamma}(u)/u]du < 0$.

In several applications of Hora's method, we have observed that the closeness of the fit to a sample data set can deteriorate significantly as the degree t of the polynomial approximation (1.2) increases. This is a highly surprising drawback for a regression-based estimation method. Even when the estimate $\hat{\gamma}(\cdot)$ of the function $\gamma(\cdot)$ improves as the degree of the polynomial increases, this does not necessarily imply a corresponding improvement in the estimate $\hat{F}^{-1}(\cdot)$ of the target inverse c.d.f. $F^{-1}(\cdot)$. This paradoxical behavior can be explained partially as follows. Suppose that in some appropriate metric, the estimated degree-d polynomial $\hat{\gamma}_d(\cdot)$ is closer to the function $\gamma(\cdot)$ than the estimated degree-c polynomial $\hat{\gamma}_c(\cdot)$, where $d > c$. For concreteness, we consider a situation in which

$$\int_0^1 [\gamma(p) - \hat{\gamma}_d(p)]^2 dp < \int_0^1 [\gamma(p) - \hat{\gamma}_c(p)]^2 dp < \infty; \tag{1.8}$$

specifically we assume $\gamma(p) \equiv 1$ (so the reference distribution provides a perfect fit to the underlying distribution), $d = 1$, and $c = 0$ with

$$\hat{\gamma}_d(p) = \sqrt{5} - \sqrt{5}\,p \text{ and } \hat{\gamma}_c(p) = 2 \text{ for all } p \in (0,1). \tag{1.9}$$

If we take $u_0 \equiv (\sqrt{5} - 2)/\sqrt{5} \simeq 0.106$, then it follows from (1.9) that

$$\gamma(u) < \hat{\gamma}_c(u) < \hat{\gamma}_d(u) \text{ for all } u \in (0, u_0). \tag{1.10}$$

If we take $p_0 \equiv 7.70 \times 10^{-5}$, then it also follows from (1.9) that

$$-\int_p^1 \frac{\hat{\gamma}_d(u)}{u}du < -\int_p^1 \frac{\hat{\gamma}_c(u)}{u}du < -\int_p^1 \frac{\gamma(u)}{u}du \quad \text{for all } p \in (0, p_0). \tag{1.11}$$

Equation (1.7) shows that the corresponding inverse c.d.f.'s $\hat{F}_d^{-1}(\cdot)$, $\hat{F}_c^{-1}(\cdot)$, and $F^{-1}(\cdot)$ are obtained respectively from (1.11) by applying the monotonically nondecreasing transformation F_0^{-1} [exp($\cdot$)] to the three expressions in (1.11) so that we have

$$\hat{F}_d^{-1}(p) < \hat{F}_c^{-1}(p) < F^{-1}(p) \text{ for all } p \in (0, p_0); \tag{1.12}$$

thus $\hat{F}_d^{-1}(\cdot)$ is worse than $\hat{F}_c^{-1}(\cdot)$ as an estimate of $F^{-1}(\cdot)$ on a nonempty subinterval of (0, 1) even though condition (1.8) holds. This example illustrates how in the chain of operations required by Hora's method to recover $\hat{F}^{-1}(\cdot)$ from $\hat{\gamma}(\cdot)$, the estimation errors $\{\gamma(p) - \hat{\gamma}(p): p \in (0, 1)\}$ can be greatly distorted as they are transformed into the corresponding deviations $\{F^{-1}(p) - \hat{F}^{-1}(p): p \in (0, 1)\}$; and an improved estimate of $\gamma(p)$ is not necessarily transformed into an improved estimate of $F^{-1}(p)$ for any $p \in (0, 1)$.

We concluded that a basic shortcoming of Hora's method is its attempt to estimate the nonnegative function $\gamma(\cdot)$ as an unrestricted polynomial based on classical linear regression. Thus we sought an alternative procedure for estimating $F^{-1}(\cdot)$ that is based on direct consideration of the squared deviations $\{[F_n^{-1}(p) - \hat{F}^{-1}(p)]^2 : p \in (0, 1)\}$ and that is appropriately constrained to yield a legitimate estimated inverse c.d.f.

2. The Estimation Procedure IDPF

2.1. Basis for Procedure IDPF

The assumptions underlying procedure IDPF parallel the assumptions underlying Hora's method. As with Hora's method, we require that the target c.d.f. $F(\cdot)$ and the reference c.d.f. $F_0(\cdot)$ must respectively possess continuous densities $f(\cdot)$ and $f_0(\cdot)$ with the same support; moreover, $f(\cdot)$ must be differentiable except possibly at a finite number of points in its support. It is also highly desirable that $F_0^{-1}(\cdot)$ should be selected from a functional family which is sufficiently flexible to allow a close approximation to the empirical inverse c.d.f. $F_n^{-1}(\cdot)$. (In §2.2.1 below, we describe a specific family of reference distributions possessing these properties.) We assume that the target inverse c.d.f. $F^{-1}(\cdot)$ can be adequately represented by the functional form

$$F^{-1}(p) = F_0^{-1}[q(p)] \text{ for all } p \in (0,1), \tag{2.1}$$

where $q(\cdot)$ is a polynomial function of p. In the spirit of Hora's original concept of adjusting a reference distribution based on sample data, we view the polynomial $q(\cdot)$ as a "filter" for the random-number input to our scheme for sampling X, where the coefficients of $q(\cdot)$ are estimated from sample data using an appropriate least-squares procedure. Thus the target random variate X is generated as $F_0^{-1}[q(U)]$, where U is a random number uniformly distributed on the unit interval (0, 1). (In §2.2.3 below, we describe a specific implementation of this variate-generation scheme for the family of reference distributions introduced in §2.2.1.)

For the right-hand side of (2.1) to define a legitimate inverse c.d.f., we require that

$$0 \leqslant q(p) \leqslant 1 \text{ for all } p \in (0,1) \tag{2.2}$$

and

$$q(p) \text{ is strictly increasing in } p \text{ for } p \in (0,1). \tag{2.3}$$

Condition (2.2) guarantees that $F_0^{-1}[q(p)]$ is defined for all $p \in (0, 1)$, while condition (2.3) ensures that $F_0^{-1}[q(p)]$ is monotonically nondecreasing in p for $p \in (0, 1)$. In addition, since $F(\cdot)$ and $F_0(\cdot)$ have the same support, we must have the following boundary conditions for $q(\cdot)$:

$$q(0) = 0, \text{ and } q(1) = 1. \tag{2.4}$$

Observe that (2.3) and (2.4) imply (2.2). Our specification of the inverse c.d.f. will be complete after we choose the degree r and the coefficients $\{\beta_j\}$ of the polynomial $q(\cdot)$. To satisfy (2.4), we take $q(\cdot)$ to have the general form

$$q(p) = \sum_{j=1}^{r-1} \beta_j p^j + \left(1 - \sum_{j=1}^{r-1} \beta_j\right) p^r \text{ for all } p \in (0,1). \tag{2.5}$$

Procedure IDPF uses a nonlinear least-squares procedure to estimate the coefficients $\{\beta_j\}$ of the polynomial (2.5), and the degree r of this polynomial is determined by a variant of a likelihood ratio test due to Gallant.[9] Assuming that r is given and using the well-known approximation[10]

$$E[X_{(i)}] \simeq F^{-1}\left(\frac{i-0.5}{n}\right) \text{ for } i = 1, \ldots, n, \tag{2.6}$$

we formulate the problem of least-squares estimation of $\{\beta_j : j = 1, 2, \ldots, r-1\}$ as

$$\min_{\{\beta_j\}_{j=1}^{r-1}} \sum_{i=1}^{n} \left\{ X_{(i)} - F_0^{-1}\left[q\left(\frac{i-0.5}{n}\right)\right]\right\}^2, \tag{2.7}$$

where $q(\cdot)$ is given by (2.5) and is subject to (2.3). The function to be minimized in (2.7) does not account for the variance of each $X_{(i)}$; thus we will refer to this version of procedure IDPF as the ordinary least-squares (OLS) estimation procedure.

To incorporate the variability of the order statistics $\{X_{(i)}\}$ into the estimation procedure IDPF, we exploit a key asymptotic property of these variates. Let p denote a fixed quantity in (0, 1). As an estimator of the pth quantile x_p, the statistic $X_{(i_n)}$ is asymptotically normal with mean x_p and variance $[p(1-p)]/[nf^2(x_p)]$ provided that $i_n/n \to p$ sufficiently fast as $n \to \infty$. Formally this property is summarized in the relation[25]

$$\frac{n^{1/2} f(x_p)[X_{(i_n)} - x_p]}{[p(1-p)]^{1/2}} \xrightarrow[n \to \infty]{D} N(0,1) \quad \text{if } i_n/n = p + o(n^{-1/2}). \tag{2.8}$$

Using $f_0[X_{(i)}]$ to approximate $f(x_p)$ for $p = (i - 0.5)/n$, we obtain the weighted least-squares (WLS) estimation problem

$$\min_{\{\beta_j\}_{j=1}^{r-1}} \sum_{i=1}^{n} \left\{ X_{(i)} - F_0^{-1}\left[q\left(\frac{i-0.5}{n}\right)\right]\right\}^2 \times \frac{n f_0^2[X_{(i)}]}{\dfrac{i-0.5}{n}\left(1 - \dfrac{i-0.5}{n}\right)}, \tag{2.9}$$

where again $q(\cdot)$ has the form (2.5) and is subject to (2.3). In §2.2.2, we detail a specific implementation of the least-squares estimation procedures prescribed by (2.7) and (2.9).

The degree r of the polynomial filter is determined by a likelihood ratio test that has been adapted to constrained nonlinear regression.[9] For concreteness, we discuss this test in the context of WLS estimation of the parameter vector $(\beta_1, \ldots, \beta_{r-1})$; a parallel development applies in the case of OLS estimation. At the outset, we assume that (2.1) and (2.5) hold for some value of r to be determined. Starting with the degree $r = 2$ and computing the optimal solution $(\hat{\beta}_1^{(r)}, \ldots, \hat{\beta}_{r-1}^{(r)})$ to (2.9) with the associated objective-function value $Q^{(r)}$, we seek to test the null hypothesis that

$$\sum_{j=1}^{r-1} \beta_j = 1 \tag{2.10}$$

(so that the degree of the polynomial filter is $r - 1$) versus the alternative hypothesis that $\sum_{j=1}^{r-1} \beta_j \neq 1$ (so that the degree of the filter is at least r). When $r = 2$, this is equivalent to testing the null hypothesis $\beta_1 = 1$ so that the inverse of the reference c.d.f. coincides with the inverse of the target c.d.f. If (2.1), (2.5), and (2.10) hold and the sample size n is large, then the analysis of Gallant[9] provides a partial basis for the approximation

$$\Pr\left\{ \frac{Q^{(r-1)}}{Q^{(r)}} > 1 + \frac{F_{1-a}(1, n-r+1)}{n-r+1} \right\} \simeq a, \tag{2.11}$$

where $a \in (0, 1)$ and $F_{1-a}(1, n-r+1)$ denotes the quantile of order $1 - a$ for the F-distribution with 1 degree of freedom in the numerator and $n - r + 1$ degrees of freedom in the denominator. Given a significance level a for the likelihood ratio test procedure, we determine the degree of the polynomial filter according to

$$r = \min\left\{ t = 2, 3, \ldots : \frac{Q^{(t-1)}}{Q^{(t)}} \leqslant 1 + \frac{F_{1-a}(1, n-t+1)}{n-t+1} \right\} - 1; \tag{2.12}$$

and we deliver the corresponding vector of least-squares parameter estimates $(\hat{\beta}_1^{(r)}, \ldots, \hat{\beta}_{r-1}^{(r)})$.

2.2. An Implementation of Procedure IDPF Using Johnson's Translation System

2.2.1. System of Reference Distributions. In this study we focused primarily on the Johnson translation system of distributions[14] as a source for the reference fits. We say that $F_0(\cdot)$ belongs to the Johnson translation system if

$$F_0(x) = \Phi\left[\gamma + \delta \cdot g\left(\frac{x - \xi}{\lambda} \right) \right], \tag{2.13}$$

where $\Phi(\cdot)$ is the standard normal c.d.f., γ and δ are shape parameters ($\delta > 0$), ξ is a location parameter, λ is a scale parameter ($\lambda > 0$), and $g(\cdot)$ is one of the following functions:

$$g(y) = \begin{cases} \log(y), & \text{for the } S_L \text{ (lognormal) family,} \\ \log[y + \sqrt{y^2 + 1}\,], & \text{for the } S_U \text{ (unbounded) family,} \\ \log[y/(1-y)], & \text{for the } S_B \text{ (bounded) family,} \\ y, & \text{for the } S_N \text{ (normal) family.} \end{cases} \tag{2.14}$$

The corresponding density function is

$$f_0(x) = \frac{\delta}{\lambda\sqrt{2\pi}} g'\left(\frac{x - \xi}{\lambda} \right) \exp\left\{ -\frac{1}{2}\left[\gamma + \delta \cdot g\left(\frac{x - \xi}{\lambda} \right) \right]^2 \right\} \quad \text{for all } x \in \mathbf{H}, \tag{2.15}$$

where $\mathbf{H}$ is the (closed) support of the distribution

$$\mathbf{H} = \begin{cases} [\,\xi, +\infty) & \text{for the } S_L \text{ (lognormal) family,} \\ (-\infty, +\infty) & \text{for the } S_U \text{ (unbounded) family,} \\ [\,\xi, \xi + \lambda] & \text{for the } S_B \text{ (bounded) family,} \\ (-\infty, +\infty) & \text{for the } S_N \text{ (normal) family,} \end{cases} \tag{2.16}$$

and $g'(\cdot)$ is the derivative of the function $g(\cdot)$ in (2.14) so that

$$g'(y) = \begin{cases} 1/y, & \text{for the } S_L \text{ (lognormal) family,} \\ 1/\sqrt{y^2 + 1}\,, & \text{for the } S_U \text{ (unbounded) family,} \\ 1/[\,y(1-y)], & \text{for the } S_B \text{ (bounded) family,} \\ 1, & \text{for the } S_N \text{ (normal) family.} \end{cases} \tag{2.17}$$

These four families of the Johnson system can fit any distribution to its first four moments, and in practice the Johnson system has been used successfully in a broad range of disciplines.[26] Moreover, a multivariate extension of the Johnson system is relatively straightforward.[15] These two properties motivated the use of the Johnson system in our implementation of procedure IDPF.

In the initial phase of procedure IDPF, we compute a specific reference distribution in the Johnson system using a noninteractive version of the software package FITTR1[26] developed by Venkatraman and Wilson.[27] Although FITTR1 incorporates a variety of methods for fitting Johnson distributions to sample data, throughout this paper all reference distributions are obtained by the method of moment matching—that is, the reference distribution is chosen to yield the same first four moments as the given sample data set. Moment matching is a popular method for fitting Johnson distributions to sample data;[12] however, this technique can yield *infeasible* parameter estimates such that some of the sample observations lie outside the support of the fitted distribution. FITTR1 incorporates a modified version of the moment-matching algorithm of Hill, Hill and Holder[12] to avoid such infeasibility in the reference distribution.

2.2.2. Numerical Methods Implemented in Procedure IDPF. Given the reference distribution obtained from the initial phase of IDPF and a candidate value for the degree r of the polynomial filter (2.5), we must estimate the parameters of the filter by invoking appropriate numerical methods to perform the minimization indicated in (2.7) or (2.9) subject to (2.3). The feasible region is a complicated subset of $(r - 1)$-dimensional Euclidean space which cannot be described conveniently in geometric or analytic terms; consequently we must also invoke appropriate numerical methods to check the feasibility of each trial solution $(\hat{\beta}_1^{(r)}, \ldots, \hat{\beta}_{r-1}^{(r)})$. These considerations motivated the use of a search technique for finding the minimum of (2.7) or (2.9), where each infeasible point is assigned a large penalty to force the search away from the infeasible region.

To check the feasibility of a trial solution $(\hat{\beta}_1^{(r)}, \ldots, \hat{\beta}_{r-1}^{(r)})$ for the minimization problem (2.7) or (2.9), we observe that the condition (2.3) is equivalent to

$$q'(p) = \sum_{j=1}^{r-1} j\hat{\beta}_j^{(r)} p^{j-1} + \left[1 - \sum_{j=1}^{r-1} \hat{\beta}_j^{(r)}\right] r p^{r-1} > 0 \quad \text{for all } p \in (0,1). \tag{2.18}$$

Since $q(1) > q(0)$, the derivative $q'(\cdot)$ must be positive in some nonempty subinterval of (0, 1). Thus (2.18) is satisfied if and only if the following equation in the variable p

$$\sum_{j=1}^{r-1} j\hat{\beta}_j^{(r)} p^{j-1} + \left[1 - \sum_{j=1}^{r-1} \hat{\beta}_j^{(r)}\right] r p^{r-1} = 0 \tag{2.19}$$

has no roots in (0, 1). To verify this condition, we applied Müller's method for finding the roots of a polynomial as implemented by Conte and de Boor.[7] If a root of (2.19) is found in the unit interval, then the corresponding trial solution $(\hat{\beta}_1^{(r)}, \ldots, \hat{\beta}_{r-1}^{(r)})$ is infeasible for the minimization problem (2.7) or (2.9); and in this case a large positive value is assigned to the objective function.

The minimization of (2.7) or (2.9) is performed using the Nelder-Mead simplex search procedure as implemented by Olsson and Nelson.[19] The objective function is evaluated at the vertices of a simplex representing alternative solutions to the minimization problem, and the search moves in a direction of declining objective-function values through a sequence of reflections, expansions, and contractions of the simplex until either (a) the simplex is sufficiently small so that the trial solutions represented by the simplex vertices are sufficiently close together in value, or (b) the differences between the objective-function values at the simplex vertices are sufficiently small. The search procedure has been used successfully in a wide variety of applications.[18]

In the computer implementation of the likelihood ratio test procedure (2.12) to determine the degree of the polynomial filter, we used the significance level $a = 0.2$. Because of the generally good quality of the reference fits provided by the initial phase of IDPF, we imposed the upper bound $r \leqslant 6$ on the degree of the fitted filter (2.5). To execute this version of procedure IDPF on a computer, we developed a portable FORTRAN 77 program which is in the public domain and is available from the authors upon request.

2.2.3. Variate-Generation Scheme. To generate variates from the inverse c.d.f. $\hat{F}^{-1}(\cdot)$ fitted to a data set $\{X_1, \ldots, X_n\}$ by procedure IDPF, we use the following scheme. The reference distribution (2.13) is defined by the quantities $\hat{\gamma}, \hat{\delta}, \hat{\xi}$, and $\hat{\lambda}$ and the function $g(\cdot)$ computed in the initial phase of IDPF as described in §2.2.1. The polynomial filter (2.5) is defined by the quantities r and $(\hat{\beta}_1^{(r)}, \ldots, \hat{\beta}_{r-1}^{(r)})$ computed in the final phase of IDPF as described in §2.2.2.

To generate a sample X from the fitted distribution, we perform the following steps:

1. Generate a random number U from the uniform distribution on (0, 1).
2. Evaluate the fitted polynomial filter at U

$$q(U) = \sum_{j=1}^{r-1} \hat{\beta}_j^{(r)} U^j + \left[1 - \sum_{j=1}^{r-1} \hat{\beta}_j^{(r)}\right] U^r. \tag{2.20}$$

3. Deliver the sample value

$$X = \hat{\xi} + \hat{\lambda} \cdot g^{-1}\left\{\frac{\Phi^{-1}[q(U)] - \hat{\gamma}}{\hat{\delta}}\right\}, \tag{2.21}$$

where

$$g^{-1}(z) = \begin{cases} e^z, & \text{for the } S_L \text{ (lognormal) family,} \\ \frac{1}{2}(e^z - e^{-z}), & \text{for the } S_U \text{ (unbounded) family,} \\ 1/(1 + e^{-z}), & \text{for the } S_B \text{ (bounded) family,} \\ z, & \text{for the } S_N \text{ (normal) family.} \end{cases} \tag{2.22}$$

To implement (2.21) in step 3 above, we recommend using an approximation to the inverse standard normal distribution function $\Phi^{-1}(\cdot)$ given in Equation 26.2.23, p. 933 of [1].

This variate-generation scheme is not difficult to implement; and although the setup time to compute all of the parameters can be large in comparison to some conventional distribution-fitting schemes, we believe that the main applications of IDPF are in the following situations where increased setup time is not the user's main concern: (a) Conventional schemes yield unacceptable fits to the sample data set, and there is a strong motivation to obtain an improved fit regardless of the increased setup time. (b) The relevant data sets to be used are so large and so numerous that sampling from the empirical inverse c.d.f. is cumbersome. We have applied procedure IDPF to large-scale simulation experiments involving numerous data sets with sample sizes exceeding 500; and in all of these situations, the variate-generation scheme (2.20)–(2.22) proved to be much more convenient to use than sampling schemes such as REMPIR[5] that are based on some variant of the empirical inverse c.d.f. Moreover, in our experience, the increased computational cost of the variate-generation scheme (2.20)–(2.22) (versus REMPIR or sampling schemes based on the unadorned empirical inverse c.d.f.) is usually negligible compared to the computational cost of the overall simulation experiment. Substantial evidence supporting this observation is given in a recent study by Klein and Baris[16] of the computation times required to generate Johnson S_B variates by inversion (that is, by the scheme (2.20)–(2.22) for the S_B family with $r = 1$) in the context of large-scale health-care simulation experiments. The following example provides some indication of the advantages of procedure IDPF in large-scale simulation input modeling.

3. An Application of Procedure IDPF

To illustrate the input-modeling problems that procedure IDPF has been designed to solve, we discuss a simulation application that arose in entomology. Researchers collected a random sample of $n = 467$ times for a certain species of Costa Rican wasps to complete the water-collecting cycle in the process of constructing a nest. This sample was called the "LWF" data set. As depicted by the dashed curve in Figure 7, the reference distribution (selected from the Johnson system by matching moments in the initial phase of IDPF) was taken to be an S_B distribution with estimated parameters $\hat{\gamma} = 5.840$, $\hat{\delta} = 1.426$, $\hat{\lambda} = 1615.$, and $\hat{\xi} = 10.99$. For this reference fit, the Kolmogorov-Smirnov goodness-of-fit statistic has the value 0.2058 and the Mann-Wald chi-squared goodness-of-fit statistic with 13 degrees of freedom has the value 140.7.

Since the reference fit showed substantial departures from the lower tail of the empirical c.d.f. where we would expect smaller values of both the target density $f(\cdot)$ and the reference density $f_0(\cdot)$, we chose to apply the OLS version of procedure IDPF to compensate for the obvious inadequacies of the reference fit. (Recall that the OLS and WLS versions of IDPF refer to the method for estimating the parameters of the filter (2.5) in the final phase of procedure IDPF.) Starting from the value $Q^{(1)} = 23.44$ for the OLS objective function (2.7) based on the reference distribution without a filter, procedure IDPF fitted a polynomial filter of degree $r = 4$ with the objective-function value $Q^{(4)} = 4.95$ and the associated coefficient estimates $\hat{\beta}_1 = 3.223$, $\hat{\beta}_2 = -7.988$, and $\hat{\beta}_3 = 9.292$. The solid curve in Figure 7 depicts the resulting c.d.f. $\hat{F}(\cdot)$. The inverse c.d.f.'s $F_n^{-1}(\cdot)$, $F_0^{-1}(\cdot)$, and $\hat{F}^{-1}(\cdot)$ can be seen by rotating Figure 7 counterclockwise by 90°. Figure 8 displays the related probability density functions (p.d.f.'s) $f_0(\cdot)$ and $\hat{f}(\cdot)$ as well as a histogram based on the sample data set. Visual inspection of Figures 7 and 8 clearly reveals the superiority of the IDPF-based fit to

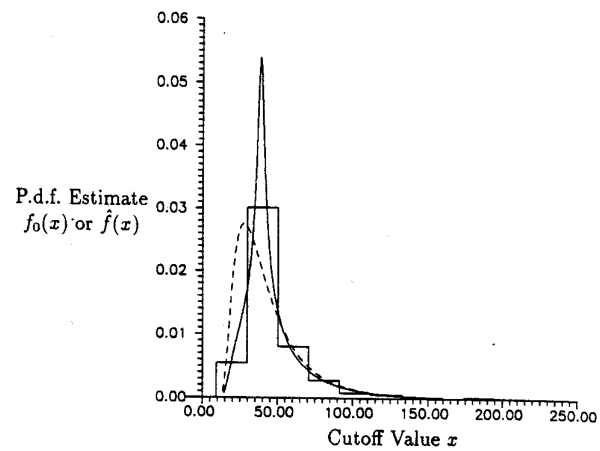

Figure 8. P.d.f.'s for the "LWF" data set—histogram (step function), reference fit (dashed curve), and OLS fit (solid curve).

the empirical distribution of the "LWF" data set compared to the reference fit.

We have observed that whereas the WLS version of IDPF is generally more effective in adjusting the central portion of the reference fit, the OLS version of IDPF is usually more effective in compensating for discrepancies in the tails of the reference fit. Although not depicted here, the WLS fit to the "LWF" data set was barely distinguishable from the OLS fit. It is also interesting to note that the OLS method yielded an estimated density $\hat{f}(\cdot)$ with a substantially larger ordinate at the mode and a markedly different shape than the reference density $f_0(\cdot)$. We believe procedure IDPF provides an open-ended mechanism for extending the basic types of distributional shapes that are achievable with a given family of reference distributions. See [2] and [3] for other applications of procedure IDPF.

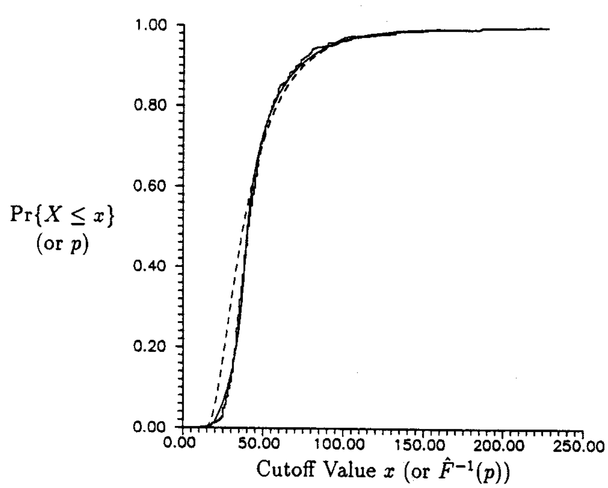

Figure 7. C.d.f.'s for the "LWF" data set—empirical c.d.f. (step function), reference fit (dashed curve), and OLS fit (solid curve).

4. Monte Carlo Evaluation of Procedure IDPF

4.1. Layout of the Monte Carlo Experiments

The two basic goals of the Monte Carlo analysis were:

1. To evaluate procedure IDPF as a data-reduction device —that is, as a means of obtaining a simplified analytic representation of a *specific* set of data. This can be done by measuring how well the fitted inverse c.d.f. approximates the empirical inverse c.d.f.
2. To evaluate procedure IDPF as a means for estimating the inverse of the underlying c.d.f. from which the sample data set has been taken. This involves measuring how well the fitted inverse c.d.f. approximates the underlying theoretical inverse c.d.f.

Although these goals coincide asymptotically as the sample size $n \to \infty$, the extent to which they agree in small samples is not clear. This consideration motivated the formulation of the separate goals 1 and 2.

In designing the Monte Carlo experiments to evaluate procedure IDPF, we selected all of the target distributions from the generalized lambda family of distributions.[21, 22] This selection was based on the flexibility of the generalized lambda family and on the simplicity of its inverse c.d.f.:

$$F^{-1}(p) = \lambda_1 + [p^{\lambda_3} - (1-p)^{\lambda_4}]/\lambda_2 \text{ for all } p \in (0,1), \tag{4.1}$$

where λ_1 is a location parameter, λ_2 is a scale parameter, and λ_3 and λ_4 are shape parameters. For further discussion of this family, see Ramberg et al.[23]

We performed six basic experiments, each with a different target distribution from the generalized lambda family. The goal here was to test procedure IDPF for a diversity of underlying distributional shapes and to identify the factors that significantly affect the performance of the procedure. All six target distributions used in our study had mean zero, variance one, skewness α_3, and kurtosis α_4 as shown in Table I. We designed a complete factorial-type experiment with low, medium, and high values for the factor α_3 and with low and high values for the factor α_4. The values of the parameters λ_1, λ_2, λ_3, and λ_4 corresponding to each of the six experiments were obtained from tables given in [23] and are also displayed in Table I.

Within each of the six basic experiments, we performed two subexperiments using the sample size n as an additional factor. The levels $n = 20$ and $n = 100$ were used. Each of the twelve resulting subexperiments consisted of the following steps:

1. Generate a sample of the selected size n from the target inverse c.d.f. of the form (4.1).
2. Select a reference c.d.f. $F_0(\cdot)$ from the Johnson system by the method of moment matching in the initial phase of IDPF.
3. Compute the estimated inverse c.d.f. $\hat{F}^{-1}(\cdot)$ by the specified filter-estimation method (OLS or WLS) in the final phase of procedure IDPF.
4. Compute the relevant performance measures that gauge the difference in quality between the reference inverse c.d.f. $F_0^{-1}(\cdot)$ and the IDPF-fitted inverse c.d.f. $\hat{F}^{-1}(\cdot)$ as estimates of the empirical inverse c.d.f. $\hat{F}_n^{-1}(\cdot)$ and the theoretical inverse c.d.f. $F^{-1}(\cdot)$.
5. Generate 400 independent replications of the protocol defined by steps 1–4 and compute appropriate summary statistics.

4.2. Formulation of the Performance Measures

To accommodate both of the stated goals of the Monte Carlo analysis, we found it necessary to formulate separate performance measures for each goal. The most natural performance measure for the first goal seems to be the final computed value of the objective function that procedure IDPF was designed to minimize. To compare performance across different sample sizes, we chose to standardize the objective function through division by the sample size n. Thus for OLS estimation, the appropriate performance measure is

$$Q_1(\tilde{F}) \equiv \frac{1}{n}\sum_{i=1}^{n}\left[X_{(i)} - \tilde{F}^{-1}\left(\frac{i-0.5}{n}\right)\right]^2, \tag{4.2}$$

where $\tilde{F}^{-1}(\cdot)$ denotes either the reference inverse c.d.f. $F_0^{-1}(\cdot)$ or the IDPF-fitted inverse c.d.f. $\hat{F}^{-1}(\cdot)$. For WLS estimation, the relevant figure of merit is

$$Q_2(\tilde{F}) \equiv \sum_{i=1}^{n}\left[X_{(i)} - \tilde{F}^{-1}\left(\frac{i-0.5}{n}\right)\right]^2 \frac{f_0^2[X_{(i)}]}{\dfrac{i-0.5}{n}\left(1-\dfrac{i-0.5}{n}\right)}. \tag{4.3}$$

We emphasize that (4.2) and (4.3) depend only on $\hat{F}_n^{-1}(\cdot)$, $F_0^{-1}(\cdot)$, and $\hat{F}^{-1}(\cdot)$; neither $Q_1(\cdot)$ nor $Q_2(\cdot)$ depends on knowledge of the true underlying inverse c.d.f. Other standard goodness-of-fit statistics might also be appropriate here—for example, the Kolmogorov-Smirnov statistic or the chi-squared statistic could be used. However, aside from the fact that these statistics have been developed to test goodness-of-fit for the c.d.f. rather than for the inverse c.d.f., we think that the performance of IDPF should be measured by the same quantity that the procedure was designed to minimize.

To gauge the success of procedure IDPF in satisfying the second distribution-fitting goal discussed in §4.1, we introduce the following quantity analogous to (4.2) and (4.3):

$$Q_3(\tilde{F}) \equiv \frac{1}{n}\sum_{i=1}^{n}\left[F^{-1}\left(\frac{i-0.5}{n}\right) - \tilde{F}^{-1}\left(\frac{i-0.5}{n}\right)\right]^2. \tag{4.4}$$

Notice that in (4.4), $F^{-1}(\cdot)$ represents the relevant inverse c.d.f. of the form (4.1); and $\tilde{F}^{-1}(\cdot)$ denotes either the reference inverse c.d.f. $F_0^{-1}(\cdot)$ or the IDPF-fitted inverse c.d.f. $\hat{F}^{-1}(\cdot)$.

Table I. Layout of the Monte Carlo Experiments

Experiment	Skewness α_3	Kurtosis α_4	Parameters of $F^{-1}(\cdot)$ in (4.1) λ_1	λ_2	λ_3	λ_4
1	0.2	3.0	−0.237	0.1983	0.1065	0.1672
2	0.2	9.0	−0.034	−0.3168	−0.1306	−0.1387
3	0.8	3.0	−1.225	0.1996	0.0068	0.3356
4	0.8	10.0	−0.141	−0.3033	−0.1129	−0.1454
5	2.0	9.0	−0.993	−0.0011	−0.0407	−0.0011
6	2.0	15.0	−0.428	−0.2380	−0.0592	−0.1415

Since procedure IDPF is based on a reference distribution, its performance should be measured by the improvement in the quality of the fit that IDPF yields relative to the reference fit. We therefore define the differences

$$\Delta Q_j \equiv Q_j(F_0) - Q_j(\hat{F}) \text{ for } j = 1,2,3. \quad (4.5)$$

For any statistic ΔQ_j in the tables to follow, we let $\overline{\Delta Q_j}$ denote the grand mean of the ΔQ_j-values across all 400 replications of the relevant Monte Carlo experiment; and we let $\mathrm{SE}(\overline{\Delta Q_j})$ denote the standard error of $\overline{\Delta Q_j}$. Finally, we define the standardized statistics

$$Z_j \equiv \frac{\overline{\Delta Q_j}}{\mathrm{SE}(\overline{\Delta Q_j})} \text{ for } j = 1,2,3. \quad (4.6)$$

Under the respective hypotheses that the differences ΔQ_1, ΔQ_2, or ΔQ_3 have expected values equal to zero, we see that Z_1, Z_2, and Z_3 respectively have asymptotic standard normal distributions. Thus the Z-values in (4.6) can be used to test the corresponding hypotheses at any desired level of significance.

As an additional means of measuring the ability of procedure IDPF to yield an improved estimate of the theoretical inverse c.d.f. $F^{-1}(\cdot)$, we collected some limited statistics on the probability distribution of the performance measure ΔQ_3. On each replication of procedure IDPF within a given subexperiment, we let B (respectively, W) denote the event that the IDPF-based fit is better (respectively, worse) than the reference fit with respect to criterion (4.4):

$$B \equiv \{\Delta Q_3 > 0\} \text{ and } W \equiv \{\Delta Q_3 < 0\}.$$

In the tables that follow, $\hat{P}(B)$ and $\hat{P}(W)$ respectively denote the estimated probability of occurrence for the events B and W based on 400 independent replications of procedure IDPF within each subexperiment.

4.3. Discussion of the Experimental Results

Tables II–V contain the results of our Monte Carlo study. The subexperiments have been renumbered {1a, 1b, 2a, ..., 6b}, with the suffix "a" denoting the sample size $n = 20$ and the suffix "b" denoting the sample size $n = 100$. We start by discussing the results for the ordinary least-squares (OLS) version of procedure IDPF. Table II displays the values of the statistics $\overline{Q_1(F_0)}$, $\overline{\Delta Q_1}$, $\mathrm{SE}(\overline{\Delta Q_1})$, and Z_1 for each subexperiment. The values of the statistic Z_1 indicate that the average differences $\overline{\Delta Q_1}$ are statistically significant at the 0.05 level for all subexperiments, with the exception of subexperiment 1b. In addition, a comparison of each mean difference $\overline{\Delta Q_1}$ with the corresponding baseline value $\overline{Q_1(F_0)}$ indicates that the OLS version of procedure IDPF yields *practically* significant improvements in fit as well as *statistically* significant improvements. The relative reduction in the value of $\overline{Q_1}$ is in the range 0–72% for the twelve subexperiments.

Table III displays the values of the statistics $\overline{Q_3(F_0)}$, $\overline{\Delta Q_3}$, $\mathrm{SE}(\overline{\Delta Q_3})$, Z_3, $\hat{P}(B)$, and $\hat{P}(W)$ for the OLS version of procedure IDPF. The values of the statistic Z_3 indicate that all of the average differences $\overline{\Delta Q_3}$ are statistically significant at the 0.05 level, with the exception of subexperiment 1b. However, the relative reduction in the grand average $\overline{Q_3}$ is somewhat smaller than the corresponding reduction in the grand average $\overline{Q_1}$ for each subexperiment; in particular, this reduction is in the range 0–70% for the twelve subexperiments. Moreover, the values of $\hat{P}(B)$ and $\hat{P}(W)$ show that most of the time IDPF yields an approximation to the theoretical inverse c.d.f. that is at least as accurate as the inverse reference c.d.f. Finally, we observe that the benefit from the use of IDPF is much larger (in terms of the relative reductions $\overline{\Delta Q_3}/\overline{Q_3(F_0)}$ and the probabilities $\hat{P}(B)$ of obtaining an improved fit) at the high levels for the factors α_3, α_4, and n.

Tables IV and V are the counterparts of Tables II and III for the weighted least-squares (WLS) version of procedure IDPF. The remarks about Tables II and III apply to Tables IV and V respectively, with the following exceptions: (a) In subexperiment 1b, the average difference $\overline{\Delta Q_3}$ has a statistically significant *negative* value whereas the average difference $\overline{\Delta Q_2}$ has a statistically significant positive value;

Table II. Goodness-of-Fit Statistics $\overline{Q_1(F_0)}$, $\overline{\Delta Q_1}$, $\mathrm{SE}(\overline{\Delta Q_1})$ and Z_1 for the OLS Version of Procedure IDPF Using a Johnson Reference Distribution

Experiment	$\overline{Q_1(F_0)}$	$\overline{\Delta Q_1}$	$\mathrm{SE}(\overline{\Delta Q_1})$	Z_1
1a	0.031202	0.004442	0.002226	1.995336
1b	0.007787	0.000001	0.000001	1.000000
2a	0.118910	0.061598	0.013712	4.492316
2b	0.076258	0.036061	0.016075	2.243273
3a	0.023577	0.004089	0.001077	3.796257
3b	0.007868	0.001993	0.000553	3.604807
4a	0.107228	0.049504	0.011215	4.414129
4b	0.186513	0.110178	0.037865	2.909756
5a	0.038363	0.009516	0.001530	6.217970
5b	0.035363	0.021986	0.002905	7.568177
6a	0.106231	0.049056	0.008228	5.962364
6b	0.373475	0.293896	0.028027	10.486024

Table III. Goodness-of-Fit Statistics $\overline{Q_3(F_0)}$, $\overline{\Delta Q_3}$, SE($\overline{\Delta Q_3}$), Z_3, $\hat{P}(B)$, and $\hat{P}(W)$ for the OLS Version of Procedure IDPF Using a Johnson Reference Distribution

Experiment	$\overline{Q_3(F_0)}$	$\overline{\Delta Q_3}$	SE($\overline{\Delta Q_3}$)	Z_3	$\hat{P}(B)$	$\hat{P}(W)$
1a	0.104595	0.004753	0.002604	1.824922	0.02	0.0025
1b	0.019628	0.000006	0.000006	1.000000	0.0025	0.0
2a	0.223423	0.053360	0.014508	3.677963	0.095	0.0275
2b	0.093924	0.034289	0.015462	2.217625	0.015	0.0
3a	0.106236	0.007226	0.002186	3.306053	0.0575	0.005
3b	0.022207	0.002798	0.000903	3.099211	0.06	0.005
4a	0.211356	0.047679	0.012719	3.748666	0.105	0.015
4b	0.205547	0.100708	0.035111	2.868305	0.0525	0.0
5a	0.165995	0.008929	0.001941	4.600364	0.1375	0.035
5b	0.058574	0.021198	0.002519	8.414308	0.3325	0.0375
6a	0.233651	0.045928	0.008443	5.439571	0.1575	0.0125
6b	0.394187	0.282496	0.027338	10.333402	0.37	0.01

Table IV. Goodness-of-Fit Statistics $\overline{Q_2(F_0)}$, $\overline{\Delta Q_2}$, SE($\overline{\Delta Q_2}$) and Z_2 for the WLS Version of Procedure IDPF Using a Johnson Reference Distribution

Experiment	$\overline{Q_2(F_0)}$	$\overline{\Delta Q_2}$	SE($\overline{\Delta Q_2}$)	Z_2
1a	0.265553	0.031452	0.012199	2.578146
1b	0.283225	0.007773	0.002066	3.762244
2a	0.552648	0.177689	0.029733	5.976159
2b	1.104506	0.450259	0.087352	5.154536
3a	0.294844	0.057471	0.011905	4.827642
3b	0.562669	0.213551	0.047910	4.457309
4a	0.537475	0.181215	0.027265	6.646435
4b	1.616831	0.767682	0.150712	5.093719
5a	0.492144	0.187897	0.018389	10.218063
5b	2.118439	1.532813	0.132652	11.555175
6a	0.668121	0.254093	0.030946	8.210784
6b	8.358417	4.963116	0.349668	14.193811

Table V. Goodness-of-Fit Statistics $\overline{Q_3(F_0)}$, $\overline{\Delta Q_3}$, SE($\overline{\Delta Q_3}$), Z_3, $\hat{P}(W)$, and $\hat{P}(B)$ for the WLS Version of Procedure IDPF Using a Johnson Reference Distribution

Experiment	$\overline{Q_3(F_0)}$	$\overline{\Delta Q_3}$	SE($\overline{\Delta Q_3}$)	Z_3	$\hat{P}(B)$	$\hat{P}(W)$
1a	0.104595	0.004848	0.002822	1.717952	0.0225	0.01
1b	0.019628	−0.000201	0.000072	−2.783028	0.02	0.04
2a	0.223423	0.035767	0.008837	4.047259	0.105	0.0325
2b	0.093924	0.044423	0.015792	2.813056	0.31	0.16
3a	0.106236	0.006405	0.001979	3.236514	0.07	0.0175
3b	0.022207	0.002714	0.000856	3.170968	0.1225	0.0525
4a	0.211356	0.051581	0.012376	4.167773	0.1125	0.025
4b	0.205547	0.101777	0.034189	2.976870	0.26	0.145
5a	0.165995	0.016111	0.003681	4.376300	0.2475	0.0975
5b	0.058574	0.021557	0.002484	8.677993	0.47	0.1225
6a	0.233651	0.046802	0.010557	4.433092	0.18	0.02
6b	0.394187	0.276697	0.026654	10.381077	0.52	0.195

however, the practical significance of these quantities is questionable. (b) The quantities $\hat{P}(B)$ and $\hat{P}(W)$ are larger than the corresponding quantities for OLS estimation, and the ratios $\hat{P}(B)/\hat{P}(W)$ are smaller than the corresponding ratios for OLS estimation. Item (b) implies that compared to the OLS version of IDPF, the WLS version is more likely to yield a nontrivial filter; and the resulting estimate $\hat{F}^{-1}(\cdot)$ of the underlying inverse c.d.f. $F^{-1}(\cdot)$ is also more likely to be worse than the reference inverse c.d.f. $F_0^{-1}(\cdot)$.

Our tentative conclusion is that the OLS version of procedure IDPF is more stable than the WLS version, at least when the method of moment matching is used in the initial phase of IDPF. A definitive comparison of these two versions will require a more extensive Monte Carlo study.

4.4. Experimental Results for Triangular Reference Distributions

To illustrate the performance of procedure IDPF with other families of reference distributions, we carried out a second set of Monte Carlo experiments using a triangular reference distribution estimated by the method of moments. This approach to obtaining an initial input model is often used by simulation practitioners.[20] We tabulate the complete results for this second set of experiments in [4]; here we merely summarize our overall findings. Our main conclusion is that IDPF is generally able to achieve greater improvements in the quality of the fit when starting from a reference distribution in the triangular family rather than the Johnson translation system. This is reasonable, since the triangular distribution family has three parameters while most distributions in the Johnson system have four parameters; thus virtually any method for selecting a reference distribution should yield a better initial fit using a Johnson distribution rather than a triangular distribution. All of the other observations of §4.3 for the case of Johnson reference distributions selected by matching moments remained the same when a triangular reference distribution was used instead. Thus we conclude that the potential benefits of using procedure IDPF are not dependent on using a Johnson reference distribution; IDPF can be used effectively with many types of reference distributions selected by a broad diversity of estimation methods.

5. Summary and Conclusions

We believe that procedure IDPF can be a useful tool for input modeling in simulation experiments. Our Monte Carlo performance evaluation provides evidence that procedure IDPF can yield improvements over the reference fit with respect to each of the following input-modeling objectives: (a) approximating the empirical inverse c.d.f., and (b) approximating the theoretical inverse c.d.f. The appealing feature of IDPF is that it can be superimposed on *any* distribution-fitting scheme and thus yield further improvement on the original fit. This is achieved at the expense of an increase in setup time due to the estimation of the filter as well as an increase in variate-generation time due to the evaluation of the filter. We believe that there are situations where the improvement in the fit is worth such a price.

The other main advantage of procedure IDPF is that the estimated inverse c.d.f. $\hat{F}^{-1}(\cdot)$ preserves all the smoothness properties of the reference inverse c.d.f. $F_0^{-1}(\cdot)$ as a consequence of the infinite differentiability of the filter $q(\cdot)$. In instances where the underlying c.d.f. is known to be smooth, we believe that procedure IDPF can give the user greater insight into the fundamental shape of the target distribution than procedures based on the empirical inverse c.d.f., which is a step function. In addition, IDPF is designed to avoid the reliability problems that have been observed with other methods for estimating inverse c.d.f.'s in simulation experiments.

Future research on procedure IDPF should focus on the method (2.12) for determining the degree r of the polynomial filter (2.5). In particular, we have not developed a complete justification for the key asymptotic property (2.11) under appropriate regularity conditions on the underlying c.d.f. $F(\cdot)$ and the reference c.d.f. $F_0(\cdot)$. Although our Monte Carlo results provide some experimental evidence of the approximate validity of (2.11) and although the theoretical and experimental results of Gallant[9] make (2.11) plausible, a more complete analysis is required. We should also examine alternative methods for determining the degree of the polynomial filter based on the asymptotic properties of nonlinear least-squares estimators.

Acknowledgments

We thank the Area Editor, the Associate Editor, and the three Referees for numerous suggestions which substantially improved this paper. We thank Bruce Schmeiser of Purdue University for many enlightening discussions on simulation input modeling. We also thank Yuehwern Yih of Purdue University and Arne Thesen of the University of Wisconsin for providing the "LWF" data set discussed in Section 3. This research was supported in part by a David Ross Grant from the Purdue Research Foundation and by the National Science Foundation under Grant No. DMS-8717799.

References

1. M. ABRAMOWITZ and I. A. STEGUN, 1972. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* National Bureau of Standards Applied Mathematics Series No. 55, U.S. Government Printing Office, Washington, D.C.
2. A. AVRAMIDIS, 1989. *A Flexible Method for Estimating Inverse Distribution Functions in Simulation Experiments,* M.S.I.E. Thesis, School of Industrial Engineering, Purdue University, West Lafayette, IN.
3. A. AVRAMIDIS and J. R. WILSON, 1989. A Flexible Method for Estimating Inverse Distribution Functions in Simulation Experiments, *Proceedings of the 1989 Winter Simulation Conference,* Institute of Electrical and Electronics Engineers, Piscataway, NJ, 428–436.
4. A. AVRAMIDIS and J. R. WILSON, 1993. Regression-Based Methods for Estimating Inverse Distribution Functions in Simulation Experiments, Technical Report 93-3, Department of Industrial Engineering, North Carolina State University, Raleigh, NC.
5. P. BRATLEY, B. L. FOX, and L. E. SCHRAGE, 1987. *A Guide to Simulation, 2nd ed.,* Springer-Verlag, New York.
6. K. L. CHUNG, 1974. *A Course in Probability Theory, 2nd ed.,* Academic Press, New York.

7. S. D. CONTE and C. DE BOOR, 1980. *Elementary Numerical Analysis: An Algorithmic Approach,* McGraw-Hill, New York.
8. N. R. DRAPER and H. SMITH, 1981. *Applied Regression Analysis, 2nd ed.,* Wiley, New York.
9. A. R. GALLANT, 1975. The Power of the Likelihood Ratio Test of Location in Nonlinear Regression Models, *Journal of the American Statistical Association 70,* 198–203.
10. G. J. HAHN and S. S. SHAPIRO, 1967. *Statistical Models in Engineering,* Wiley, New York.
12. I. D. HILL, R. HILL and R. L. HOLDER, 1976. Fitting Johnson Curves by Moments, *Applied Statistics 25,* 180–189.
13. S. C. HORA, 1983. Estimation of the Inverse Function for Random Variate Generation, *Communications of the Association for Computing Machinery 26,* 590–594.
14. N. L. JOHNSON, 1949. Systems of Frequency Curves Generated by Methods of Translation, *Biometrika 36,* 149–176.
15. N. L. JOHNSON, 1949. Bivariate Distributions Based on Simple Translation Systems, *Biometrika 36,* 297–304.
16. R. W. KLEIN and P. M. BARIS, 1991. Selecting and Generating Variates for Modeling Service Times, *Computers and Industrial Engineering 20,* 27–33.
17. D. E. KLINE, D. A. BENDER and J. L. NIEBER, 1987. Modified Regression Approach for Modeling Probability Distributions, *Transactions of the American Society of Agricultural Engineers 30,* 697–702.
18. D. M. OLSSON, 1974. A Sequential Simplex Program for Solving Minimization Problems, *Journal of Quality Technology 6,* 53–57.
19. D. M. OLSSON and L. S. NELSON, 1975. The Nelder-Mead Simplex Procedure for Function Minimization, *Technometrics 17,* 45–51.
20. C. D. PEGDEN, R. E. SHANNON, and R. P. SADOWSKI, 1990. *Introduction to Simulation Using SIMAN,* McGraw-Hill, New York.
21. J. S. RAMBERG and B. W. SCHMEISER, 1972. An Approximate Method for Generating Symmetric Random Variables, *Communications of the Association for Computing Machinery 15,* 987–990.
22. J. S. RAMBERG and B. W. SCHMEISER, 1974. An Approximate Method for Generating Asymmetric Random Variables, *Communications of the Association for Computing Machinery 17,* 78–82.
23. J. S. RAMBERG, P. R. TADIKAMALLA, E. J. DUDEWICZ and E. F. MYKYTKA, 1979. A Probability Distribution and its Uses in Fitting Data, *Technometrics 21,* 201–214.
24. B. W. SCHMEISER, 1990. Simulation Experiments, in *Handbooks in OR & MS,* Volume 2, D. P. Heyman and M. J. Sobel (eds.), North-Holland, Amsterdam, 295–330.
25. R. J. SERFLING, 1980. *Approximation Theorems of Mathematical Statistics,* Wiley, New York.
26. J. J. SWAIN, S. VENKATRAMAN and J. R. WILSON, 1988. Least-Squares Estimation of Distribution Functions in Johnson's Translation System, *Journal of Statistical Computation and Simulation 29,* 271–297.
27. S. VENKATRAMAN and J. R. WILSON, 1987. Modeling Univariate Populations with Johnson's Translation System—Description of the FITTR1 Software, Research Memorandum 87–21, School of Industrial Engineering, Purdue University, West Lafayette, IN.
28. J. R. WILSON, 1984. Variance Reduction Techniques for Digital Simulation, *American Journal of Mathematical and Management Sciences 4,* 277–312.