Proceedings of the 1989 Winter Simulation Conference
E.A. MacNair, K.J. Musselman, P. Heidelberger (eds.)

A FLEXIBLE METHOD FOR ESTIMATING INVERSE DISTRIBUTION FUNCTIONS IN SIMULATION EXPERIMENTS

Athanassios Avramidis
James R. Wilson
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.

ABSTRACT

To generate random variates from an unknown continuous distribution via the inverse transform method, we present a flexible, computationally tractable procedure for estimating the associated inverse distribution function based on sample data. Previously proposed methods for estimating inverse distribution functions can fail in either the distribution-fitting or variate-generation stages of application. To avoid these difficulties, we have developed the procedure IDPF for estimating an Inverse Distribution with a Polynomial Filter. After a first-cut or reference distribution has been obtained by some standard technique, a front-end polynomial filter for the inverse of the reference distribution is estimated by constrained nonlinear regression so that the resulting inverse distribution has minimum "distance" from the empirical inverse distribution. The constraints on the regression ensure that the fitted inverse distribution function is nondefective and monotonically nondecreasing. A specific implementation of this procedure is based on well-known techniques for obtaining a reference fit from the Johnson translation system of distributions. We present the results of a Monte Carlo study to demonstrate the effectiveness of the method. Compared to the reference fit, procedure IDPF yields significantly better approximations not only to the empirical inverse distribution function but also to the underlying theoretical inverse distribution function.

1. INTRODUCTION

In the development of discrete-event simulation models, one frequently needs to generate independent observations of a continuous random variable X having an unknown cumulative distribution function (CDF) $F()$. Typically a random sample $\{X_1, X_2, ..., X_n\}$ from $F()$ is available, and this sample defines the associated empirical distribution function $F_n()$. 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 number of distributional shapes and thus a limited capability for approximating the distribution of the sample. Moreover, variate generation is troublesome or expensive for many well-known families of distributions.

Hora (1983) proposed a method for simulation input modeling that uses the inverse of a known continuous CDF $F_0()$ (the so-called *reference distribution*) as the starting point for estimating the target inverse CDF $F^{-1}()$. Hora assumed that $F^{-1}()$ has the functional form

$$F^{-1}(p) = F_0^{-1}\left\{p^{\alpha_0}\exp\left[\sum_{j=1}^{t}\frac{\alpha_j(p^j - 1)}{j}\right]\right\}, \quad p \in (0, 1), \qquad (1.1)$$

where t and $F_0()$ are to be chosen by the modeler. Hora's method attempts to reduce the problem of fitting an inverse CDF to that of choosing a reference distribution and then performing linear regression to estimate the parameters $\{\alpha_j : j = 0, 1, ..., t\}$ in (1.1). 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 serious drawbacks: (a) The linear statistical model that underlies the procedure for estimating the $\{\alpha_j\}$ has an exponentially distributed, multiplicative error term rather than a normally distributed, additive error term; and this invalidates all of the usual inferential procedures based on normal linear regression theory. (b) The fitted inverse CDF of the form (1.1) may fail to be monotonic; and in addition to being nonsensical, this condition can destroy the effectiveness of standard variance reduction techniques such as common random numbers and antithetic variates. (c) The fitted inverse CDF of the form (1.1) 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. See Avramidis (1989) for a detailed analysis of the properties of Hora's method.

Using Hora's formulation (1.1) as a point of departure, we propose a new method for fitting an Inverse Distribution with a Polynomial Filter (IDPF). Given a reference distribution $F_0()$ representing a first-cut estimate of the unknown continuous distribution $F()$ that is to be sampled by the inverse transform method, we seek an improved estimate of $F^{-1}()$ based on the assumption that $F^{-1}()$ can be adequately modeled with the functional form

$$F^{-1}(p) = F_0^{-1}\left[\sum_{j=1}^{r-1}\beta_j p^j + \left(1 - \sum_{j=1}^{r-1}\beta_j\right)p^r\right], \quad p \in (0, 1). \tag{1.2}$$

The coefficient estimates $\{\hat{\beta}_j\}$ are selected to minimize an appropriate function of the corresponding estimation errors $\{\hat{F}^{-1}(p) - F_n^{-1}(p) : p \in (0, 1)\}$ subject to the constraint that the polynomial within the square brackets on the right-hand side of (1.2) is monotonically nondecreasing.

This paper is organized as follows. In Section 2 we discuss the basis for procedure IDPF, and we present an implementation of this procedure using Johnson's translation system of distributions for the reference fits. Section 3 shows a typical application of the procedure, and Section 4 summarizes the results of a Monte Carlo study of the performance of the procedure. The main conclusions of this work are recapitulated in Section 5.

2. THE ESTIMATION PROCEDURE IDPF

2.1. Basis for Procedure IDPF

As with Hora's method, at the outset we require an *admissible* reference distribution $F_0()$; this means that $F_0()$ must have the same support as $F()$. It is of course highly desirable that $F_0^{-1}()$ should also provide a reasonably close approximation to the empirical inverse distribution $F_n^{-1}()$. We assume that the target inverse CDF $F^{-1}()$ has the form

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

where $q()$ is a polynomial function of p. In order for (2.1) to define a legitimate inverse CDF, we require that

$$0 \le q(p) \le 1, \quad p \in (0, 1) \quad \text{and} \tag{2.2a}$$

$$q(p) \text{ is nondecreasing in } p \text{ for } p \in (0, 1). \tag{2.2b}$$

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

$$q(0) = 0, \quad \text{and} \quad q(1) = 1. \tag{2.2c}$$

Note that (2.2b) and (2.2c) imply (2.2a). Our specification of the inverse CDF will be complete after we choose the degree r and the coefficients $\{\beta_j\}$ of the polynomial $q()$. To satisfy (2.2c), $q()$ must have the general form

$$q(p) = \sum_{j=1}^{r-1}\beta_j p^j + \left(1 - \sum_{j=1}^{r-1}\beta_j\right)p^r, \quad p \in (0, 1). \tag{2.3}$$

Procedure IDPF is based on a nonlinear least-squares procedure for selecting the degree r of the polynomial (2.3) and for estimating the coefficients $\{\beta_j\}$ of this polynomial. Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote the order statistics corresponding to the given sample. Using the well-known approximation

$$E[X_{(i)}] \simeq F^{-1}\left(\frac{i - \frac{1}{2}}{n}\right), \quad i = 1, \ldots, n \tag{2.4}$$

(Hahn and Shapiro 1967), we formulate the problem of least-squares estimation of r and $\{\beta_j : j = 1, 2, \ldots, r-1\}$ as

$$\min_{r, \{\beta_j\}} \sum_{i=1}^{n}\left\{X_{(i)} - F_0^{-1}\left[q\left(\frac{i - \frac{1}{2}}{n}\right)\right]\right\}^2, \tag{2.5}$$

where $q()$ is given by (2.3) and is subject to (2.2b). The function to be minimized in (2.5) does not account for the variance of $X_{(i)}$; thus we will refer to this variant of procedure IDPF as ordinary least-squares (OLS) estimation.

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 α denote a fixed percentage in (0, 1). As an estimator of the order-α quantile $x_\alpha \equiv F^{-1}(\alpha)$, the statistic $X_{(i_n)}$ satisfying $(i_n - 1)/n \le \alpha \le i_n/n$ for every sample size n is asymptotically normal with mean x_α and variance $[\alpha(1 - \alpha)]/[nf^2(x_\alpha)]$, where $f()$ is the density corresponding to the unknown CDF $F()$. Formally this property is summarized in the relation

$$\frac{n^{1/2} f(x_\alpha)\left[X_{(i_n)} - x_\alpha\right]}{[\alpha(1-\alpha)]^{1/2}} \xrightarrow[n \to \infty]{D} N(0, 1) \tag{2.6}$$

(see p. 94 of Serfling (1980)). Using the reference density $f_0()$ to approximate $f()$ and using $F_0^{-1}\{q[(i-\frac{1}{2})/n]\}$ to approximate x_α for $\alpha = (i - \frac{1}{2})/n$, we obtain the weighted least-squares (WLS) estimation problem:

$$\min_{r, \{\beta_j\}} \sum_{i=1}^{n} \left\{X_{(i)} - F_0^{-1}\left[q\left(\frac{i-\frac{1}{2}}{n}\right)\right]\right\}^2 \frac{n f_0^2[X_{(i)}]}{\frac{i-\frac{1}{2}}{n}\left(1 - \frac{i-\frac{1}{2}}{n}\right)}, \tag{2.7}$$

where again $q()$ has the form (2.3) and is subject to (2.2b).

Next we consider the problem of minimizing (2.5) or (2.7) subject to (2.2b). We assume for the moment that r is fixed; rules for choosing r will be discussed later. Note that the feasible region is a complicated subset of $(r-1)$-dimensional Euclidean space which cannot be described conveniently in geometric or analytic terms; instead we must resort to numerical techniques for checking the feasibility of each trial solution $(\beta_1, \beta_2, \ldots, \beta_{r-1})$. This motivates the use of a search technique for finding the minimum of (2.5) or (2.7), where each infeasible point is assigned a large penalty to force the search away from the infeasible region. Taking the derivative of $q()$ with respect to p, we see that (2.2b) is equivalent to

$$q'(p) = \sum_{j=1}^{r-1} j\beta_j p^{j-1} + \left(1 - \sum_{j=1}^{r-1}\beta_j\right) r p^{r-1} \geq 0, \quad p \in (0, 1). \tag{2.8}$$

Since $q(1) > q(0)$, the derivative $q'()$ must be positive in some subinterval of $(0, 1)$. Thus (2.8) is satisfied if and only if the equation

$$\sum_{j=1}^{r-1} j\beta_j p^{j-1} + \left(1 - \sum_{j=1}^{r-1}\beta_j\right) r p^{r-1} = 0 \tag{2.9}$$

has no roots in $(0, 1)$. To verify this condition, we apply a standard algorithm for finding the roots of a polynomial. If a root of (2.9) is found in the unit interval, then the corresponding polynomial $q()$ is infeasible for the minimization problem (2.5) or (2.7); and in this case a very large positive value must be assigned to the objective function.

It is not difficult to show that at least in principle, we can always find a polynomial $q()$ for which the objective function in (2.5) or (2.7) is arbitrarily close to zero. However, this polynomial might have an arbitrarily high degree r, making it unusable in simulation applications. Thus in practice, procedure IDPF should start with a reference distribution that provides a fairly good fit to the sample distribution so that a manageable upper limit can be set for the degree r; moreover a reasonable heuristic rule should be used to identify the optimal value of r.

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

In this study we concentrate on using the Johnson translation system of distributions as a source for the reference fit. We say that $F_0()$ belongs to the Johnson translation system if

$$F_0(x) = \Phi\left[\gamma + \delta \cdot g\left(\frac{x-\xi}{\lambda}\right)\right] \quad \text{for all } x \in \mathbf{H}, \tag{2.10}$$

where $\Phi()$ is the standard normal CDF, γ and δ are shape parameters, ξ is a location parameter, λ is a scale parameter, $\mathbf{H}$ is the (closed) support of the distribution

$$\mathbf{H} = \begin{cases} [\xi, +\infty) & \text{for the } S_L \text{ (lognormal) family,} \\ (-\infty, +\infty) & \text{for the } S_U \text{ (unbounded) family,} \\ [\xi, \xi + \lambda] & \text{for the } S_B \text{ (bounded) family,} \\ (-\infty, +\infty) & \text{for the } S_N \text{ (normal) family,} \end{cases} \tag{2.11}$$

and $g()$ is one of the following functions:

$$g(y) = \begin{cases} \ln(y), & \text{for the } S_L \text{ (lognormal) family,} \\ \ln\left[y + \sqrt{y^2+1}\right], & \text{for the } S_U \text{ (unbounded) family,} \\ \ln[y/(1-y)], & \text{for the } S_B \text{ (bounded) family,} \\ y, & \text{for the } S_N \text{ (normal) family.} \end{cases} \tag{2.12}$$

For all $x \in \mathbf{H}$, the corresponding density function is

$$f_0(x) = \frac{\delta}{\lambda\sqrt{2\pi}} g'\left(\frac{x-\xi}{\lambda}\right) \exp\left\{-\frac{1}{2}\left[\gamma + \delta \cdot g\left(\frac{x-\xi}{\lambda}\right)\right]^2\right\}, \tag{2.13}$$

where

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

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. Moreover, a multivariate extension of the Johnson system is relatively straightforward (Johnson (1949b)). These two properties motivated the use of the Johnson system in our implementation of procedure IDPF.

As a front end for procedure IDPF that computes a specific reference distribution in the Johnson system, we used a noninteractive version of the software package FITTR1 developed by Venkatraman and Wilson (1987). Although FITTR1 incorporates a variety of methods for fitting Johnson distributions to sample data, the example discussed in the next section is based on 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. Although moment matching is a popular method for fitting Johnson distributions to sample data, this technique can yield *infeasible* parameter estimates such that some of the sample observations lie outside the support of the fitted distribution. We have modified the moment-matching algorithm of FITTR1 to avoid such infeasibility in the reference distribution.

The minimization of (2.5) or (2.7) is performed using the Nelder-Mead simplex search procedure as implemented by Olsson and Nelson (1974). 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 or (b) the differences between the objective-function values at the simplex vertices are sufficiently small. This search procedure has been used successfully in a wide variety of applications (Olsson 1974).

To find the roots of equation (2.9), we employ Müller's method as implemented by Conte and de Boor (1980). Given approximations p_{i-2}, p_{i-1}, and p_i to a zero of the function $h(p)$ defined by the left-hand side of (2.9) for all real p, we take the next approximation p_i to be a zero of the parabola that passes through the three points $[p_{i-2}, h(p_{i-2})]$, $[p_{i-1}, h(p_{i-1})]$, and $[p_i, h(p_i)]$. Once a zero z_1 of $h(p)$ has been located to a prespecified accuracy, the procedure is repeated with the deflated function $h_1(p) \equiv h(p)/(p - z_1)$ to find the next zero. The algorithm returns all complex roots of a polynomial with real coefficients; of course we only use the real roots for the purposes of IDPF.

In the absence of a formal statistical test for identifying the degree r of the polynomial filter, we use the following heuristic rule. Starting with the degree $r = 2$ and the associated coefficient estimates $\{\hat{\beta}_j^{(r)}\}$ and objective-function value $Q^{(r)}$, we perform another iteration of the fitting procedure with degree $r+1$ to obtain the coefficient estimates $\{\hat{\beta}_j^{(r+1)}\}$ and the objective-function value $Q^{(r+1)}$. So long as $Q^{(r+1)} < 0.95Q^{(r)}$, additional iterations of the fitting procedure are performed. When the last objective-function value $Q^{(r+1)}$ is within 5% of the next-to-last value $Q^{(r)}$, the next-to-last parameter values r and $\{\hat{\beta}_j^{(r)} : j = 1, ..., r-1\}$ are delivered and the procedure stops. Because of the generally good quality of the reference fits provided by FITTR1, we imposed the additional limit $r \leq 6$ in this implementation of procedure IDPF. To execute this version of procedure IDPF on a computer, we developed a portable FORTRAN 77 program which is available upon request.

3. AN APPLICATION OF PROCEDURE IDPF

To illustrate the input-modeling problems that procedure IDPF was designed to solve, we discuss a simulation application that arose in the field of medical decision making. The sample data set labelled SGOT is a random sample of blood levels of serum glutamic oxylacetic transaminase (expressed in units per deciliter) taken from a population of elderly diabetics enrolled in a monitoring program of a general medicine clinic. As shown by the dashed curve in Figure 1, the reference distribution $F_0()$ determined by the moment-matching option of FITTR1 was an S_B (bounded) distribution with estimated parameters $\hat{\gamma} = 3.154$, $\hat{\delta} = 0.914$, $\hat{\lambda} = 706.9$, and $\hat{\xi} = 6.000$. For this reference fit, the Kolmogorov-Smirnov goodness-of-fit statistic had the value 0.163; and the Mann-Wald chi-square goodness-of-fit statistic with 4 degrees of freedom had the value 19.09. Note that the step function in Figure 1 is the empirical CDF $F_n()$.

Since the reference fit showed substantial departures from the central portion of the empirical distribution where we would expect the largest values of both the target density $f()$ and the reference density $f_0()$, we chose to apply the WLS version of procedure IDPF to compensate for the obvious inadequacies of the reference fit. Starting from the reference value $Q^{(0)} = 4.694$ of the WLS objective function (2.7), procedure IDPF fitted a polynomial filter of degree $r = 4$ with the objective-function value $Q^{(4)} = 0.3123$ and the associated coefficient estimates $\hat{\beta}_1 = 2.686$, $\hat{\beta}_2 = -5.543$, and $\hat{\beta}_3 = 5.400$. The solid curve in Figure 1 shows the resulting CDF $\hat{F}()$. The corresponding inverse distribution functions $F_n^{-1}()$, $F_0^{-1}()$, and $\hat{F}^{-1}()$ can be seen by

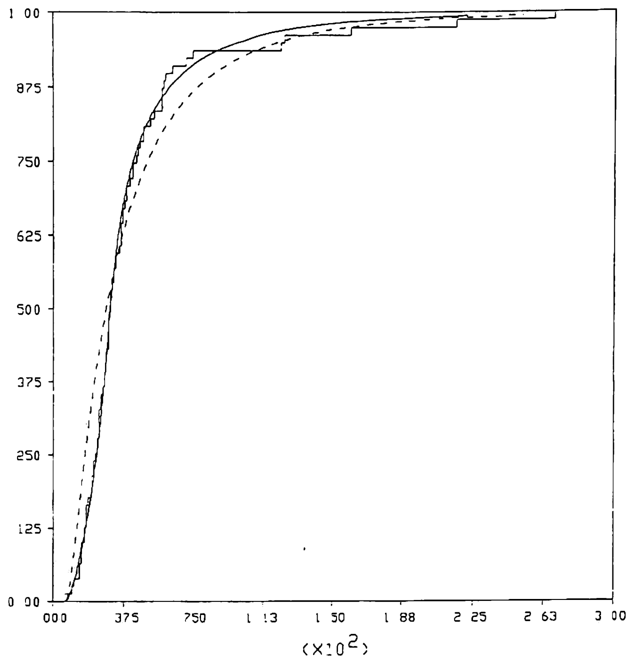

Figure 1: CDFs for SGOT Data – Empirical CDF (Stepped), Reference Fit (Dashed Curve), and WLS Fit (Solid Curve)

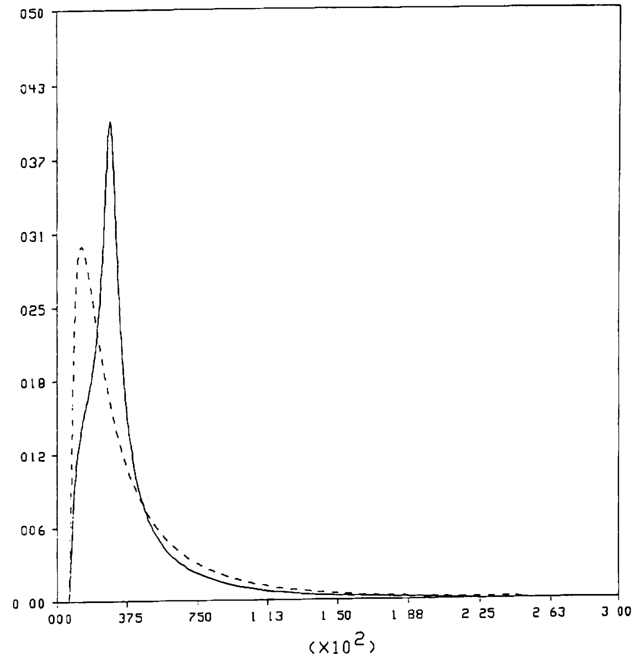

Figure 2: PDFs for SGOT Data – Reference Fit (Dashed Curve) and WLS Fit (Solid Curve)

rotating Figure 1 counterclockwise by 90°. Figure 2 shows the related probability density functions (PDFs) $f_0()$ and $\hat{f}()$.

We have observed that whereas the WLS version of procedure IDPF is more effective in adjusting the central portion of the reference fit, the OLS version is more effective in compensating for discrepancies in the tails of the reference fit. Although not depicted here, the OLS fit to the SGOT data set was barely distinguishable from the reference fit. It is also interesting to note that the WLS method yielded a density $\hat{f}()$ with a substantially larger ordinate at the mode and a markedly different shape than the reference density $f_0()$. 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 Avramidis (1989) for other applications of procedure IDPF.

4. MONTE CARLO EVALUATION OF PROCEDURE IDPF

4.1. Layout of the Monte Carlo Experiments

The two basic goals of the Monte Carlo analysis are:

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 CDF approximates the empirical inverse CDF.
2. To evaluate procedure IDPF as a means for estimating the inverse of the underlying distribution from which the sample data set has been taken. This involves measuring how well the fitted inverse CDF approximates the underlying theoretical inverse CDF.

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 (Ramberg and Schmeiser 1972, 1974). This selection was based on the flexibility of the generalized lambda family and on the simplicity of its inverse CDF:

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

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. (1979).

We performed four 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 four target distributions used in our study have mean 0, variance 1, skewness α_3, and kurtosis α_4 as shown in Table 1. We designed a complete factorial-type experiment with high and low values for the factors α_3 and α_4. The values of the parameters λ_1, λ_2, λ_3, and λ_4 corresponding to each of the four experiments were obtained from tables given in Ramberg et al. (1979) and are also displayed in Table 1.

Table 1: Layout of the Monte Carlo Experiments.

Expt.	α_3	α_4	λ_1	λ_2	λ_3	λ_4
1	0.8	3.0	− 1.225	0.199	0.006	0.335
2	0.8	10.0	− 0.141	− 0.303	− 0.112	− 0.145
3	2.0	9.0	− 0.993	− 0.001	− 0.040	− 0.001
4	2.0	15.0	− 0.426	− 0.238	− 0.059	− 0.141

Within each of the four 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 eight resulting subexperiments consisted of the following steps:

1. Generate a sample of the selected size n from the target distribution of the form (4.1).
2. Determine a reference distribution $F_0()$ by selecting the best of the S_U and S_B fits computed with the least-squares estimation options of FITTR1 (Swain, Venkatraman, and Wilson 1988).
3. Determine the final estimate $\hat{F}^{-1}()$ of the target inverse CDF using the selected version of procedure IDPF.
4. Compute the relevant performance measures that gauge the difference in quality between the reference fit and the IDPF fit.
5. Generate 100 independent replications of the protocol defined by steps 1−4 and compute the grand mean and the standard error of the grand mean for each performance measure.

4.2. Formulation of the Performance Measures

To accomodate 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. For the OLS method this quantity is

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

where $\bar{F}^{-1}()$ denotes either the reference inverse $F_0^{-1}()$ or the IDPF-fitted inverse $\hat{F}^{-1}()$. For the WLS method, the relevant figure of merit is

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

It should be stressed that (4.2) and (4.3) depend only on $F_n^{-1}()$, $F_0^{-1}()$, and $\hat{F}^{-1}()$; neither $Q_1()$ nor $Q_2()$ depends on knowledge of the true underlying inverse CDF. Other standard goodness-of-fit statistics might also be appropriate here − for example, the Kolmogorov-Smirnov statistic or the chi-square statistic could be used. However, aside from the fact that these statistics have been developed to test goodness-of-fit for the CDF rather than for the inverse CDF, we think that the performance of procedure IDPF should be measured by precisely 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 Subsection 4.1, we introduce a quantity similar to the Kolmogorov-Smirnov statistic that has been adapted to estimation of an inverse CDF rather than a CDF. Specifically, we define

$$D(\bar{F}) \equiv \sup\{\,|\,\bar{F}^{-1}(p) - F^{-1}(p)\,|\; : p \in (0,1)\}, \tag{4.4}$$

where $F^{-1}()$ denotes the target inverse CDF (4.1); and again $\bar{F}^{-1}()$ denotes either the reference inverse $F_0^{-1}()$ or the IDPF-fitted inverse $\hat{F}^{-1}()$. Although (4.4) cannot be computed exactly, accurate approximations to it can be obtained by computing the maximum of $|\,\bar{F}^{-1}(p) - F^{-1}(p)\,|$ for a finite set of closely-spaced p-values between zero and one.

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 difference

$$\Delta Q_1 \equiv Q_1(F_0) - Q_1(\hat{F}), \tag{4.5}$$

where $Q_1(F_0)$ and $Q_1(\hat{F})$ respectively denote the Q_1-statistics for the reference fit and the IDPF fit. Similarly, we let

$$\Delta Q_2 \equiv Q_2(F_0) - Q_2(\hat{F}) \quad \text{and} \tag{4.6}$$

$$\Delta D \equiv D(F_0) - D(\hat{F}). \tag{4.7}$$

As a general rule for any statistic T in the tables to follow, we let $\overline{T}$ denote the grand mean of the T-values across all 100 replications of the relevant Monte Carlo experiment; and we let $\mathrm{SE}(\overline{T})$ denote the standard error of $\overline{T}$. Finally, we define the standardized statistics

$$Z_{\Delta Q_1} \equiv \frac{\overline{\Delta Q_1}}{\mathrm{SE}(\overline{\Delta Q_1})}, \; Z_{\Delta Q_2} \equiv \frac{\overline{\Delta Q_2}}{\mathrm{SE}(\overline{\Delta Q_2})}, \; Z_{\Delta D} \equiv \frac{\overline{\Delta D}}{\mathrm{SE}(\overline{\Delta D})}. \tag{4.8}$$

Under the respective hypotheses that the differences ΔQ_1, ΔQ_2, or ΔD have expected values equal to zero, $Z_{\Delta Q_1}$, $Z_{\Delta Q_2}$, and $Z_{\Delta D}$ respectively have asymptotic standard normal distributions. Thus the Z-values in (4.8) can be used to test the corresponding hypotheses at any desired level of significance.

4.3. Discussion of the Experimental Results

Tables 2 through 5 contain the results of our Monte Carlo experiments. The subexperiments have been renumbered {1a, 1b, 2a, ..., 4b}, 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 2 displays the values of the statistics $\overline{Q_1(F_0)}$, $\overline{\Delta Q_1}$, $\mathrm{SE}(\overline{\Delta Q_1})$, and $Z_{\Delta Q_1}$ for each subexperiment. The values of the statistic $Z_{\Delta Q_1}$ indicate that the observed differences $\overline{\Delta Q_1}$ are statistically significant for all eight subexperiments; and these results provide some evidence of the effectiveness of procedure IDPF. In addition, a comparison of each mean difference $\overline{\Delta Q_1}$ with the corresponding baseline value $\overline{Q_1(F_0)}$ indicates that in all eight subexperiments, the OLS version of procedure IDPF yields *practically* significant improvements in fit as well as *statistically* significant improvements. Neither the shape of the target distribution (that is, α_3 and α_4) nor the sample size appear to affect the performance of the method; the relative reduction in the value of $\overline{Q_1}$ is approximately 30-50% for all eight subexperiments.

Table 3 displays the values of the statistics $\overline{D(F_0)}$, $\overline{\Delta D}$, $\mathrm{SE}(\overline{\Delta D})$, and $Z_{\Delta D}$ for the OLS version of procedure IDPF. Again, the values of the statistic $Z_{\Delta D}$ indicate that the differences $\overline{\Delta D}$ are statistically significant for all subexperiments, with the exception of subexperiment 2a and possibly subexperiment 4a. However, the relative reduction in the grand average $\overline{D}$ is generally much smaller than the corresponding reduction in the grand average $\overline{Q_1}$ for each subexperiment. As in Table 2, Table 3 does not reveal any systematic effect on the performance of procedure IDPF that is due to the shape of the target distribution or the sample size; the reduction in $\overline{D}$ relative to the comparable figure for the reference distribution is approximately 0-20% for all eight subexperiments.

Tables 4 and 5 are the counterparts of Tables 2 and 3 for the weighted least-squares (WLS) version of procedure IDPF. All remarks about Tables 2 and 3 apply to Tables 4 and 5 respectively. The relative reductions in the values of the statistics $\overline{Q_2}$ and $\overline{D}$ are approximately the same as the corresponding figures for the OLS procedure. Thus no marked difference in the performance of the OLS and WLS versions of procedure IDPF can be detected from this study. A definitive comparison of these two procedures will require a more extensive Monte Carlo study.

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 some evidence that procedure IDPF can yield significant improvements relative to the reference fit with respect to each of the following input-modeling objectives: (a) approximating the empirical inverse CDF, and (b) approximating the true underlying inverse distribution. Procedure IDPF has been designed to avoid the reliability problems that have been observed with other methods for estimating inverse CDFs in simulation experiments. The main disadvantage of the procedure is the lack of a rigorously developed statistical theory for selecting the order of the polynomial filter. This is a topic for future research.

Table 2: Goodness-of-Fit Statistics $\overline{Q_1(F_0)}$, $\overline{\Delta Q_1}$, SE($\overline{\Delta Q_1}$), and $Z_{\Delta Q_1}$ for the OLS Version of Procedure IDPF

Expt.	$\overline{Q_1(F_0)}$	$\overline{\Delta Q_1}$	SE($\overline{\Delta Q_1}$)	$Z_{\Delta Q_1}$
1a	0.0531	0.0253	0.0067	3.7795
1b	0.0108	0.0023	0.0004	4.8640
2a	0.0957	0.0360	0.0083	4.2968
2b	0.0612	0.0250	0.0038	6.4861
3a	0.0850	0.0501	0.0159	3.1508
3b	0.0292	0.0079	0.0018	4.3046
4a	0.0931	0.0319	0.0069	4.6086
4b	0.0966	0.0420	0.0058	7.1744

Table 3: Goodness-of-Fit Statistics $\overline{D(F_0)}$, $\overline{\Delta D}$, SE($\overline{\Delta D}$), and $Z_{\Delta D}$ for the OLS Version of Procedure IDPF

Expt.	$\overline{D(F_0)}$	$\overline{\Delta D}$	SE($\overline{\Delta D}$)	$Z_{\Delta D}$
1a	1.0564	0.2030	0.0474	4.2774
1b	0.4172	0.0381	0.0103	3.7013
2a	1.3338	0.0010	0.0418	0.0248
2b	0.9928	0.1118	0.0458	2.4390
3a	1.8740	0.3109	0.0940	3.3060
3b	0.8576	0.0970	0.0307	3.1566
4a	1.5824	0.0831	0.0544	1.5255
4b	1.1101	0.2348	0.0575	4.0778

Table 4: Goodness-of-Fit Statistics $\overline{Q_2(F_0)}$, $\overline{\Delta Q_2}$, SE($\overline{\Delta Q_2}$), and $Z_{\Delta Q_2}$ for the WLS Version of Procedure IDPF

Expt.	$\overline{Q_2(F_0)}$	$\overline{\Delta Q_2}$	SE($\overline{\Delta Q_2}$)	$Z_{\Delta Q_2}$
1a	0.3793	0.1561	0.0151	10.3240
1b	0.4523	0.0629	0.0083	7.5775
2a	0.2839	0.1291	0.0120	10.7620
2b	0.2863	0.0512	0.0104	4.8961
3a	0.3449	0.1421	0.0121	11.7083
3b	0.5984	0.0665	0.0108	6.1633
4a	0.2827	0.1229	0.0095	12.9020
4b	0.3310	0.0952	0.0195	4.8853

Table 5: Goodness-of-Fit Statistics $\overline{D(F_0)}$, $\overline{\Delta D}$, SE($\overline{\Delta D}$), and $Z_{\Delta D}$ for the WLS Version of Procedure IDPF

Expt.	$\overline{D(F_0)}$	$\overline{\Delta D}$	SE($\overline{\Delta D}$)	$Z_{\Delta D}$
1a	1.0558	0.1607	0.0402	3.9891
1b	0.4203	0.0226	0.0061	3.6825
2a	1.3091	- 0.0082	0.0201	- 0.4069
2b	1.1040	0.0630	0.0214	2.9327
3a	1.8368	0.1968	0.0647	3.0406
3b	0.8854	0.0415	0.0188	2.2028
4a	1.5777	0.0729	0.0658	1.1075
4b	1.1739	0.0904	0.0240	3.7661

ACKNOWLEDGMENTS

This research is supported by the National Science Foundation under Grant No. DMS-8717799. The U.S. Government has certain rights in this material.

REFERENCES

Avramidis, A. (1989). A flexible method for estimating inverse distribution functions in simulation experiments. Unpublished M.S.I.E. Thesis, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

Conte, S. D. and de Boor, C. (1980). *Elementary Numerical Analysis: An Algorithmic Approach*, McGraw-Hill, New York.

Hahn, G. J. and Shapiro, S. S. (1967). *Statistical Models in Engineering*, John Wiley & Sons, New York.

Hora, S. C. (1983). Estimation of the inverse function for random variate generation. *Communications of the ACM* **26**, 590-594.

Johnson, N. L. (1949a). Systems of frequency curves generated by methods of translation. *Biometrika* **36**, 149-176.

Johnson, N. L. (1949b). Bivariate distributions based on simple translation systems. *Biometrika* **36**, 297-304.

Olsson, D. M. (1974). A sequential simplex program for solving minimization problems. *Journal of Quality Technology* **6**, 53-57.

Olsson, D. M. and Nelson, L. S. (1975). The Nelder-Mead simplex procedure for function minimization. *Technometrics* **17**, 45-51.

Ramberg, J. S. and Schmeiser, B. W. (1972). An approximate method for generating symmetric random variables. *Communications of the ACM* **15**, 987-990.

Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM* **17**, 78-82.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., and Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics* **21**, 201-214.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Swain, J. J., Venkatraman, S., and Wilson, J. R. (1988). Least-squares estimation of distribution functions in Johnson's translation system. *Journal of Statistical Computation and Simulation* **29**, 271-297.

Venkatraman, S. and Wilson, J. R. (1987). Modeling univariate populations with Johnson's translation system – Description of the FITTR1 software. Research Memorandum, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

AUTHORS' BIOGRAPHIES

ATHANASSIOS AVRAMIDIS is a Ph.D. student in the School of Industrial Engineering at Purdue University. He received a diploma in mechanical engineering from the University of Thessaloniki (Greece) in 1987, and he received an M.S. in industrial engineering from Purdue University in 1989. His current research interests include simulation input modeling and variance reduction techniques. He is a student member of ORSA and IIE.

Athanassios Avramidis
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.
(317) 494-0607
avramidi@gb.ecn.purdue.edu

JAMES R. WILSON is an Associate Professor in the School of Industrial Engineering at Purdue University. He received a B.A. in mathematics from Rice University, and he received M.S. and Ph.D. degrees in industrial engineering from Purdue University. His current research interests include simulation input modeling, variance reduction techniques, ranking-and-selection procedures, simulation output analysis, and medical decision analysis. He currently serves as President of TIMS/College on Simulation, Associate Editor of *IIE Transactions*, and Simulation Department Editor of *Management Science*.

James R. Wilson
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.
(317) 494-5408
wilsonj@gc.ecn.purdue.edu