Proceedings of the 1985 Winter Simulation Conference
D. Gantz, G. Blais, S. Solomon (eds.)

# FITTING JOHNSON DISTRIBUTIONS USING LEAST SQUARES: SIMULATION APPLICATIONS

James J. Swain
School of Industrial and
Systems Engineering
Georgia Tech
Atlanta, GA 30332

James R. Wilson
School of Industrial Engineering
Purdue University
Pritsker & Associates, Inc.
W. Lafayette, IN 47907

ABSTRACT

A weighted least squares regression method is proposed for fitting cumulative probability distributions to data. This technique is illustrated for the Johnson translation system of distributions. The least squares procedure minimizes the distance between the vector of uniformized order statistics and its corresponding expected value to identify the Johnson distribution that provides the best fit. This least squares procedure is shown to be numerically robust and to provide a good fit of the data when compared to the empirical distribution. Two examples illustrate the use of the procedure.

## 1. Introduction

A common problem encountered by simulation practititoners is the choice and estimation of probability distributions to describe random variates. While theoretical models are sometimes preferred where they are sufficiently accurate, many "real world" data sets are not readily fit by the standard one- and two- parameter theoretical distributions. Flexible systems of distributions such as the *Johnson* translation system [1] are often used for fitting in those cases where fidelity to the data and parsimonious representation are desired, and an exact theoretical distribution is not required. In the case of the Johnson family, the distribution can be specified using one of four functional forms and up to four parameters. Representation by an empirical distribution, even using grouped data, would require many more parameters to specify the distribution.

A least squares method is proposed for fitting Johnson distributions to data. Fitting of distributional families is most often performed by matching sample moments to the moments of the distributional family to obtain a particular representation. However, even when the sample is sufficiently large to ensure that the sample moments are accurate, fitting may be difficult and the resulting fit may be *infeasible*: the fitted distribution may have a lower (upper) endpoint which is larger (smaller) than the corresponding extreme value observed in the sample [2]. Let $X_1, \ldots, X_n$ denote a random sample from the distribution $F_X$ that is to be estimated. We propose using weighted least squares to fit each uniformized order statistic of the sample, $U_{(j)}=F_X(X_{(j)})$, to the expected value of the associated uniform order statistic, $E[U_{(j)}]=j/(n+1)$ for $j=1,\ldots,n$ . This method is shown to be numerically robust and to provide an accurate approximation of the both the empirical and exact sampling distributions.

The Johnson family is based on four transformations to the normal distribution, so that $F_X(x)$ can be represented by

$$F_X(x) \approx \Phi\{ \psi_1 + \psi_2 f[(x-\psi_4)/\psi_3] \}, \qquad (1)$$

where $\Phi(z)$ is the cumulative distribution function of the standard normal and $f(u)$ is one of the following transformations

$$f(u) = \begin{cases} \log(u) & \text{for } S_L \text{ family} \\ \sinh^{-1}(u) & \text{for } S_U \text{ family} \\ \log[u/(1-u)] & \text{for } S_B \text{ family} \\ u & \text{for } S_N \text{ family} \end{cases}. \qquad (2)$$

The four families represent the lognormal, the "unbounded" distribution, the "bounded" distribution, and the normal. As shown in Figure 1, these four families cover the entire $(\beta_1,\beta_2)$ plane, assigning a unique distribution to each point. Points in the plane fall either on the normal point $(0,3)$, on the lognormal line, above the lognormal line for bounded distributions, or below the lognormal line for unbounded distributions. The parameters $\psi=(\psi_1,\psi_2,\psi_3,\psi_4)$ complete the specification of the distribution. Note that generating random variates from each Johnson distribution, where required, is straightforward [3].

## 2. Least Squares Distribution Fitting

The least squares method for fitting a Johnson distribution seeks to minimize the distance (in n-dimensional space) between the uniformized order statistics and their corresponding expected values. This procedure is based on the observation that the sample order statistics $X_{(j)}$,

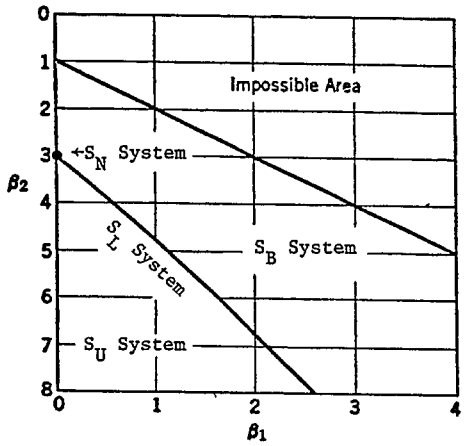


Figure 1: Regions Defining the Johnson Distributions in the $\beta_1$, $\beta_2$ plane.

$$X_{(1)} < X_{(2)} < \ldots < X_{(j)} < \ldots < X_{(n)},$$

from a sample of n observations can be converted to the (approximately) uniformized order statistics

$$R_j(\psi) = \Phi\{\psi_1 + \psi_2 f[(X_{(j)}-\psi_4)/\psi_3]\}, \qquad (3)$$

so that $R_j(\psi) \sim U_{(j)}$, the $j^{th}$ smallest in a sample of n random numbers. These variates can be represented for fitting as $R_j(\psi)=\rho_j+\varepsilon_j$, where $\rho_j=E[U_{(j)}]=j/(n+1)$, $j=1,\ldots,n$. The random error $\varepsilon_j$ is a translated uniform order statistic with mean $E[\varepsilon_j]=0$. The covariance between the errors $\varepsilon_j$ and $\varepsilon_k$ is given by

$$\mathrm{Cov}(\varepsilon_j,\varepsilon_k) = \frac{j(n-k+1)}{(n+1)^2(n+2)}, \quad j \le k \le n. \qquad (4)$$

Letting $R(\psi)=[R_1(\psi),R_2(\psi),\ldots R_n(\psi)]^T$, $\rho=(\rho_1,\rho_2,\ldots\rho_n)^T$, $\varepsilon=R(\psi)-\rho$ and letting $V=\|\mathrm{Cov}(\varepsilon_j,\varepsilon_k)\|$ denote the n x n covariance matrix of $\varepsilon$, we see that the least squares estimation problem becomes

$$\left.\begin{array}{l} \underset{\psi}{\text{minimize}}\ s(\psi) = \varepsilon^T(\psi)V^{-1}\varepsilon(\psi) \\ \text{subject to:} \\ \psi_2 \ge 0 \\ \psi_3 \begin{cases} \ge 0 & \text{for } S_U \\ \ge X_{(n)}-\psi_4 & \text{for } S_B \\ =1 & \text{for } S_N \text{ and } S_L \end{cases} \\ \psi_4 \begin{cases} \le X_{(1)} & \text{for } S_L \text{ and } S_B \\ =0 & \text{for } S_U \end{cases} \end{array}\right\} \qquad (5)$$

Note that mimimization by weighted least squares is reasonable, since the errors $\varepsilon_j$ are all approximately normally distributed except for the extreme order statistics, $j \to 1$ and $j \to n$. The correlation among the errors presents no problems, since the variance matrix is known and can be used to form the weighting matrix for the least squares estimation.

The general problem can be simplified somewhat since the weighting matrix $V^{-1}$ has a tridiagonal form

$$V^{-1} = (n+1)(n+2)\begin{bmatrix} 2 & -1 & 0 & 0 & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 \\ & \ldots & & & \ldots & \\ 0 & \ldots & & -1 & 2 & -1 \\ 0 & \ldots & & 0 & -1 & 2 \end{bmatrix}$$

and can be readily factored using the Cholesky decomposition $LL^T$, where L is lower bidiagonal matrix given by

$$L = \begin{bmatrix} \tau_{11} & 0 & 0 & 0 \ldots & 0 & 0 \\ \tau_{21} & \tau_{22} & 0 & 0 \ldots & 0 & 0 \\ 0 & \tau_{32} & \tau_{33} & 0 \ldots & 0 & 0 \\ & \ldots & & \ldots & & \\ 0 & 0 & 0 & 0 \ldots & \tau_{n,n-1} & \tau_{nn} \end{bmatrix}$$

where

$$\tau_{ij} = \begin{cases} \{(n+1)(n+2)(i+1)/i\}^{1/2} & 1 \le i=j \le n \\ -\{(n+1)(n+2)(i-1)/i\}^{1/2} & 2 \le i=j+1 \le n \\ 0 & \text{otherwise} \end{cases}$$

The weighted sums of squares function $s(\psi)$ can be expressed as an unweighted least squares problem in terms of the weighted residuals, $s(\psi)=U(\psi)^TU(\psi)$, where the weighted residuals are given by $U(\psi)=L^T\varepsilon(\psi)$. In either formulation, $s(\psi)$ is easily evaluated with computing formula

$$s(\psi)=(n+1)(n+2)\left[\sum_{j=1}^{n}\varepsilon_j^2(\psi) - \sum_{j=2}^{n}\varepsilon_j(\psi)\varepsilon_{j-1}(\psi)\right]$$

which can be minimized using any general purpose optimization package.

The transformation function f(u) can be specified in a number of ways. A common procedure is to compute sample statistics for $\sqrt{\beta_1}$ and $\beta_2$ to determine which family is to be used. Given the variability of higher order sample moments, it is usually preferable to fit the data using each of the transformations and to make the choice based upon the minimum sum of squares or perhaps upon the minimum chi-square goodness-of-fit statistic.

## 3. Examples

Two examples are presented to show the effectiveness of the Johnson translation system when estimation is performed by least squares estimation. The first problem is a risk assessment in structural reliability and the second involves a description of the sampling distribution of parameter estimators arising in a nonlinear statistical model. In each case, the results from the Johnson system are compared to empirical results and to the analytic solution, when it is known.

The sampled results were obtained using FORTRAN V on a CDC CYBER 855 computer, with random numbers obtained using IMSL subroutines [4]. Each set of runs was independently seeded. The numerical results for the fitted Johnson distribution were all performed on an IBM PC using Microsoft FORTRAN version 2.0 . The Nelder-Mead Simplex search algorithm [5] was used to solve the minimization problem (5). The Johnson curve fitting package [2] is a portable, ANSI FORTRAN 77 program design to perform all aspects of fitting, including data summary, fitting by moment matching to sample or exact moments, estimation through weighted least squares, ordinary least squares, or percentile matching, and comparison between empirical and fitted distributions.

### 3.1 Structural Reliability Problem

Ayyub and Haldar [6] present a several methodologies for evaluating structural reliability and give an example for a particular case. The general problem is to determine from mechanical considerations and wind loadings whether a structure is likely to fail. Both the external wind loadings and several of the structural parameters are treated as random variates. In the latter case, nonhomogeneities of the materials and uncertainties in exact parameters are modeled through random variation.

For the particular case of a 120 ft high, 6 ft diameter cylindrical pressure vessel, the equation for the *failure surface* Z is given by

$$Z = 144\,\pi\,X_1\,F_y - .0241344\,C_f\,X_2\,V_{30}^2,$$

in which $X_1$ and $X_2$ are lognormally distributed, $F_y$ and $C_f$ are normally distributed, and $V_{30}$ is Weibull distributed. The random variates $X_1$, $F_y$, $C_f$, $X_2$, and $V_{30}$ are also assumed to be mutually independent. The structure is reliable if Z>0. It is desired to estimate the risk probability, P(Z<0).

Ayyub and Haldar present a number of approximate and simulation results, including direct simulation and simulation with variance reduction using antithetic variates and conditional expectations, alone and in combination, to get precise estimates of the reliability of the structure as a function of the nominal thickness of the vessel walls of the pressure vessel. It is important to note that direct simulation does not work well as the thickness of the vessel increases: since as the risk decreases, fewer observations of Z<0 will be generated. For instance, in two of the four cases presented, no negative Z-values are observed in random samples of size 5000. The direct empirical estimates that P(Z<0)=0 offer limited information about the actual magnitude of the risk in these cases.

To illustrate the Johnson distribution fit using weighted least squares, two Monte Carlo samples of n=500 observations of Z were generated. The two fits are compared to the empirical distributions in Figures 2 and 3. The first sample was fit using the $S_U$ family of distributions. The second sample was fit using the $S_B$ distribution with a lower bound of -42,242 and a range of 97,975. As illustrated, the fit in both cases appears to be quite good. This is confirmed numerically by the $\chi^2$ goodness-of-fit statistics for the two cases, 18.85 (p=17%) for the first case and 7.1 (p=93%) in the second case.

The Johnson distribution can also be fit using ordinary least squares, where the covariances among the errors are not taken into account. Since the weighted formulation assigns more importance to the extreme tails, there is reason to think that the ordinary least squares fit could give a better overall fit, except possibly in the tails. This is illustrated for the first case in Figure 4. Note that the larger residuals near the center of the distribution in Figure 2 are reduced when ordinary least squares is used as shown in Figure 4.

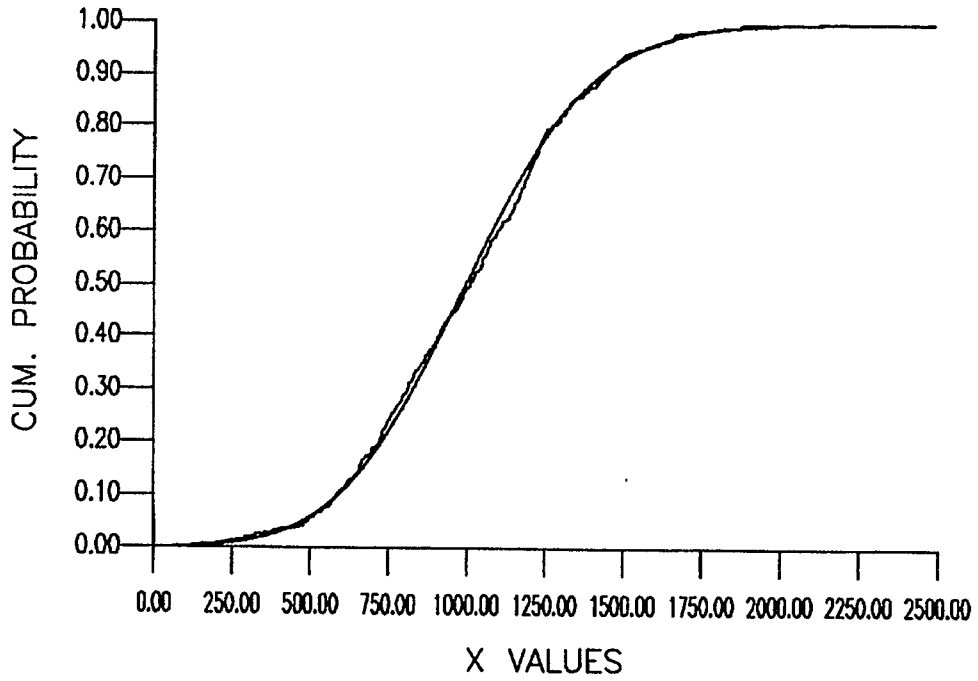


Figure 2: Weighted Least Squares Fit for Reliability Data Set 1 (Smooth line is fitted curve).

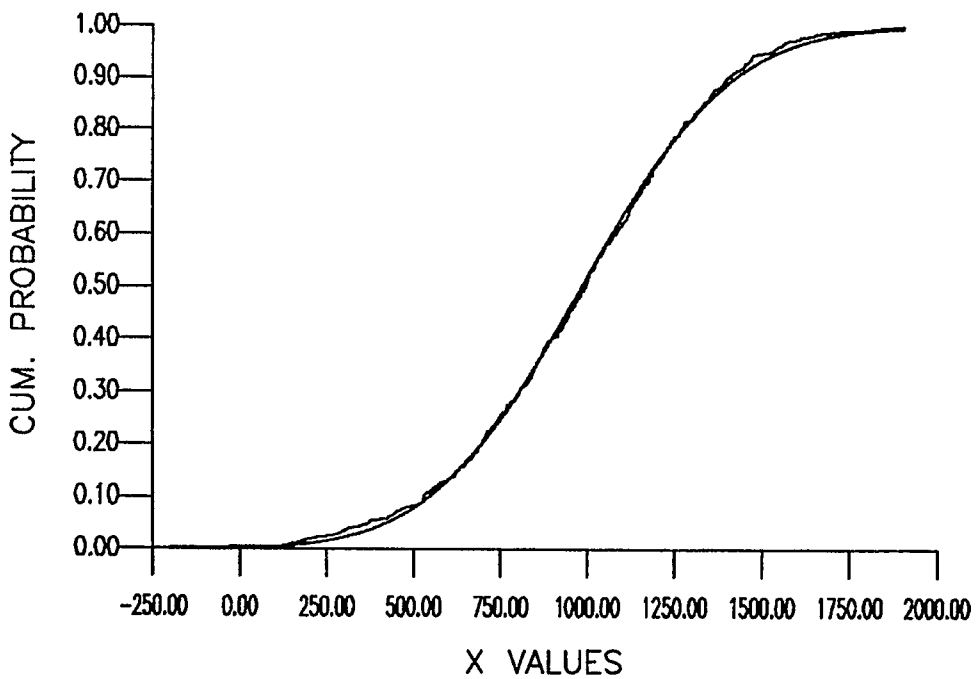


Figure 3: Weighted Least Squares Fit for Reliability Data Set 2 (Smooth line is fitted curve).

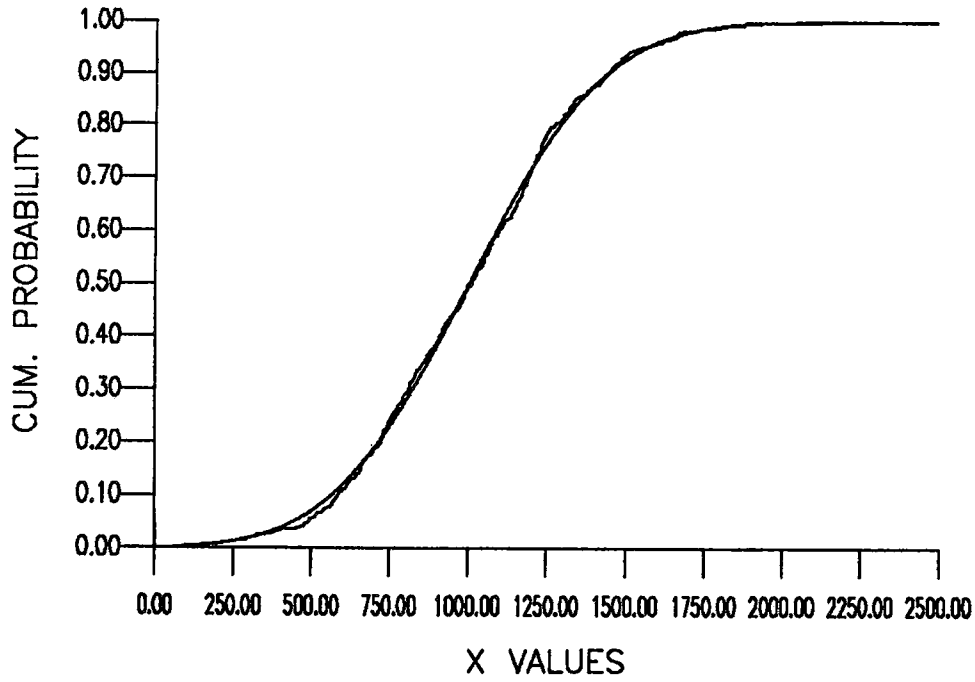


Figure 4: Ordinary Least Squares Fit for Reliability Data Set 1 (Smooth line is fitted curve).

The least-squares fits can be contrasted to the fit by moment matching. This is the usual method for choosing and fitting a Johnson distribution. One selects the appropriate family from the sample moments using Figure 1. Once the family is selected, the particular distribution is chosen by matching its moments to the sample moments. The standard algorithm of Hill, Hill, and Holder [7] converged in the first case but required coversion to double precision operation to converge for the sample moments in the second case.

The two cases also illustrate the use of the Johnson fitted distributions for making estimates of the risk probability P(Z<0). The empirical estimate in the first case is 0, since the smallest observation in this case is 65.7 . The risk estimate using the fitted Johnson distribution is 0.00108. In the second case the empirical estimate is 0.004, while the fitted Johnson distribution estimate is 0.00205. A direct estimate of the risk probability based on 20,000 Monte Carlo samples is 0.00505 (s.e. .000997).

### 3.2 A Nonlinear Estimator Problem

Swain [8] presents a number of Monte Carlo investigations of the sampling distribution of estimators t for models nonlinear in the parameters θ. The sampling distribution is difficult to characterize for finite samples when the model is nonlinear in the model parameters. For instance, even when errors are normal, the sampling distribution is typically only asymptotically normal, and may be distinctly nonnormal in small samples.

As an example, consider the problem of estimating θ when observations are of the form

$$y_i = x_i^{\theta} + \varepsilon_i \qquad i=1,2$$

and $X=(x_1,x_2)=(1,\xi)$. The errors ε are independent random variates with zero means and common variance $\sigma^2$. Least squares solutions can be readily computed in this problem, since $t=\log(y_2)/\log(\xi)$. Numerical solutions for the distribution of t can be obtained for a number of error distributions, including the (truncated) normal, gamma, and uniform.

The versatility of the Johnson family of distributions is illustrated by three sets of samples for this nonlinear problem. Each set contains 500 observations for a particular error distribution for ε: the first set has (truncated) normal errors, the second has gamma errors, and the third has uniform errors. The empirical distribution function and Johnson fits using weighted least squares are compared for the three sets of observations in Figures 5, 6,

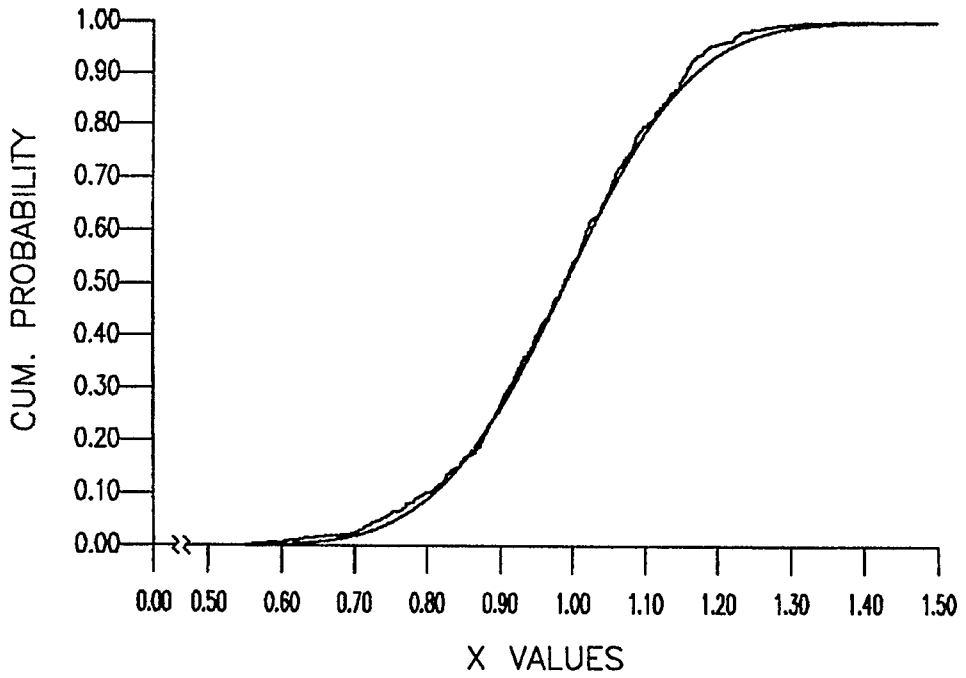


Figure 5: Weighted Least Squares Fit for Nonlinear Model with Normal Errors (Smooth line is fitted curve).

and 7. Somewhat better fits were obtained overall using ordinary (unweighted) least squares (not illustrated here).

## 4. Discussion

The figures illustrate the versatility and quality of approximation using the Johnson family of distributions when fit by weighted least squares. The weighted least squares method was also compared to the methods of moment matching, percentile matching [2], and ordinary least squares. Weighted least squares fitting had two noteworthy features: it always worked and the fits obtained were generally good. In this sense weighted least squares is a robust procedure, since it always converged to an acceptable solution. While other procedures sometimes did a little better, they were not uniformly superior; and in particular, they did not always converge to an answer. Ordinary least squares solutions, for instance, generally required many more iterations of the Simplex search procedure than the weighted least squares method; and sometimes ordinary least squares failed to converge. On the other hand, the ordinary least squares fit seemed qualitatively to be a better fit over the entire distribution when convergence was obtained. Several good fits were also obtained using either percentile matching or moment matching, but there appeared to be somewhat more variability in the quality of the fits, and convergence was not always obtained. Finally, it appears that the moment matching algorithm converges more reliably when the computations are performed in double precision.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the assistance of Jose Rodriquez of Pritsker & Associates in the preparation of the figures using the graphics capability of TESS.

## REFERENCES

1. Johnson, N. L., "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika* 36, 1949, pp. 149-176.

2. Wilson, J. R., "Modeling Multiattribute Populations with Johnson's Translation System," Technical Report, Mechanical Engineering Department, University of Texas, Austin, TX, 1983.

3. Hill, I. D., "Algorithm AS 100. Normal-Johnson and Johnson-Normal Transformations," *Applied Statistics* 25, 1976, pp. 190-192.

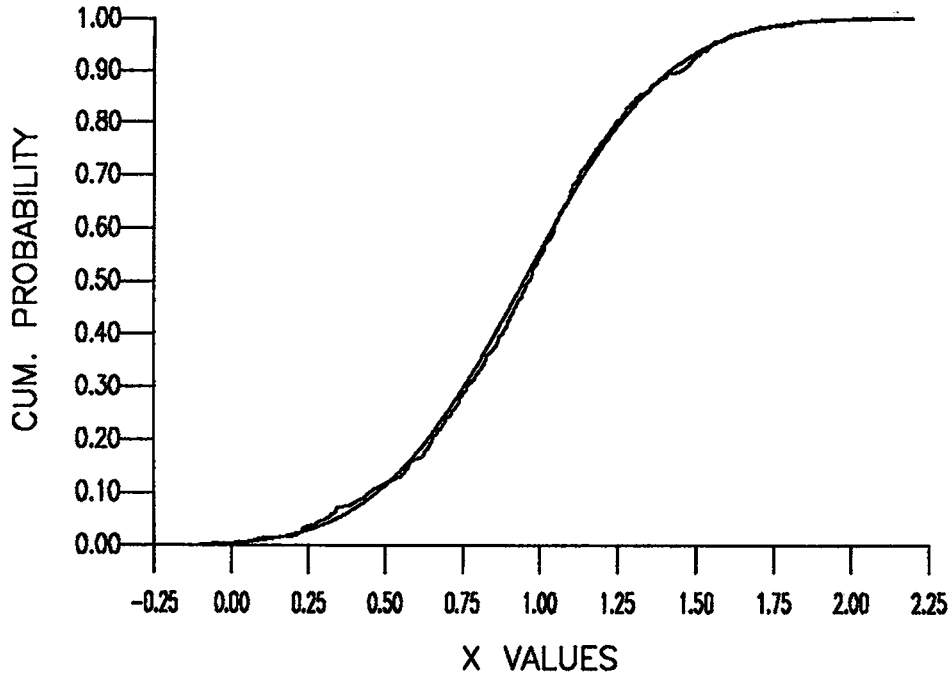


Figure 6: Weighted Least Squares Fit for Nonlinear Model with Gamma Errors (Smooth line is fitted curve).

4. International Mathematical and Statistical Library. *IMSL Library Reference* Manual, 8th edition, IMSL, Inc., Houston, Tx, 1980.

5. Olsson, D. M., "A Sequential Simplex Program for Solving Minimization Problems," *Journal of Quality Technology* 6, 1974, pp. 53-57.

6. Ayyub, B. M. and A. Haldar, "Practical Structural Reliability Techniques," *Journal* of Structural Engineering 110(8), 1984, pp. 1707-1724.

7. Hill, I. D., R. Hill, and R. L. Holder, "Fitting Johnson Curves by Moments," *Applied Statistics* 25, 1976, pp. 180-189.

8. Swain, J. J., *Monte Carlo Estimation of* Sampling Distributions Arising in Nonlinear Statistical Modeling, Unpublished Ph.D. Dissertation, Purdue University, W. Lafayette, IN, 1982.

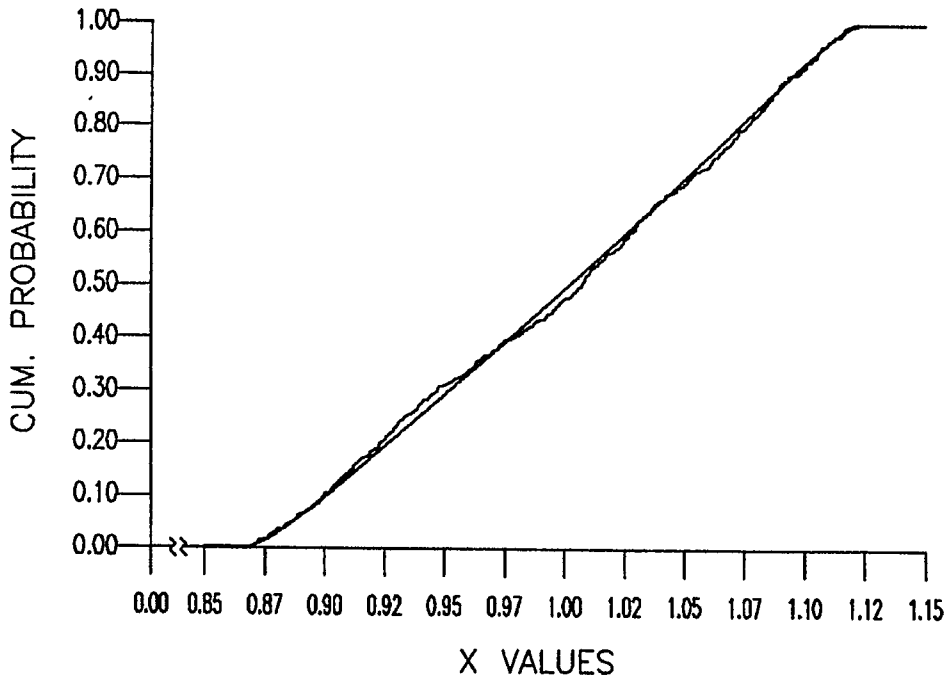


Figure 7: Weighted Least Squares Fit for Nonlinear Model with Uniform Errors (Smooth line is fitted curve).

JAMES J. SWAIN is an assistant professor in the School of Industrial and Systems Engineering at the Georgia Institute of Technology. From 1977 to 1979 he was a systems analyst in the Management Information Department of Air Products and Chemicals, Allentown, PA. He received a BA in Liberal Studies in 1974, a BS in Engineering Science in 1975, and an MS in Mechanical Engineering in 1977 from the University of Notre Dame. He received his PhD in Industrial Engineering from Purdue University in 1982. His current research interests include the analysis of nonlinear regression models, Monte Carlo variance reduction methods in statistical problems, and numerical methods. He is a member of ASA, IIE, ORSA, and SCS.

James J. Swain
School of ISYE
Georgia Tech
Atlanta, GA 30332
(404) 894-3025

JAMES R. WILSON is an associate professor in the School of Industrial Engineering at Purdue University. From 1979 to 1984, he was an assistant professor in the Operations Research Group of the Mechanical Engineering Department in The University of Texas at Austin. He received a BA in mathematics from Rice University in 1970, an MS in idustrial engineering from Purdue University in 1977, and a PhD in industrial engineering from Purdue University. His current research interest include simulation output analysis, ranking and selection procedures, and variance reduction techniques. He is a member of ACM, AIDS, IIE, ORSA, SCS, and TIMS.

James R. Wilson
School of Industrial Engineering
Purdue University
(317) 494-5408