*Proceedings of the 1988 Winter Simulation Conference*
M. Abrams, P. Haigh, and J. Comfort (eds.)

# Input modeling with the Johnson System of distributions

David J. DeBrota
Robert S. Dittus
Regenstrief Institute
for Health Care
Indiana University
School of Medicine
1001 West 10th Street
Indianapolis, IN 46202, U.S.A.

Stephen D. Roberts
James R. Wilson
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.

James J. Swain
School of Industrial and
Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, U.S.A.

Sekhar Venkatraman
Department of Industrial Engineering
Wichita State University
Wichita, KS 67208, U.S.A.

## ABSTRACT

This paper provides an introduction to the Johnson translation system of probability distributions, and it describes methods for using the Johnson system to model input processes in simulation experiments. The fitting methods based on available data are incorporated into the public-domain software package FITTR1. To handle situations in which little or no data is available, we present a visual interactive method for subjective distribution fitting that has been implemented in the public-domain software package VISIFIT. We present several examples illustrating the use of FITTR1 and VISIFIT for simulation input modeling.

## 1. INTRODUCTION

A common problem in modeling stochastic systems is the selection and estimation of probability distributions, and an important step in this process is the identification of a suitable family of distributions. When sample data are available, this is usually accomplished by hypothesizing standard parametric distributions and by performing diagnostic checks to assess the adequacy of the fit. When sample data are not available, we want the chosen family of distributions to possess a few desirable properties that are specified for the input process. To achieve an adequate representation of the process being modeled, it becomes necessary to use a family of distributions that is capable of yielding a wide variety of shapes. The range of shapes that is possible with a given family of distributions is a measure of its flexibility, and many of the standard parametric distributions are extremely limited in distributional shapes (Schmeiser 1977).

In this paper we discuss the use of the Johnson (1949) translation system to model univariate populations. (The term *method of translation* refers to the transformation of a continuous random variable to a standard normal variate.) The Johnson system is able to closely approximate many of the standard continuous distributions through one of four functional forms and is thus highly flexible. This paper describes the interactive software package FITTR1 which has been developed to model univariate populations with the Johnson system when sample data are available. We also describe a visual interactive method for subjective distribution fitting when little or no sample data are available, and we discuss the software package VISIFIT in which this visual approach has been implemented. Both FITTR1 and VISIFIT are in the public domain and are available from the authors upon request.

This paper is organized as follows. Section 2 is a discussion of the Johnson system of distributions. Section 3 summarizes the principal methods of distribution identification and parameter estimation for the Johnson system that have been implemented in FITTR1. The operation of FITTR1 is described in Section 4. The main issues arising in subjective estimation of probability distributions are discussed in Section 5. The operation of VISIFIT is detailed in Section 6. We summarize our conclusions about input modeling with Johnson distributions in Section 7.

## 2. THE JOHNSON TRANSLATION SYSTEM

Let $X$ be a continuous random variable with distribution function $F(x) = \Pr\{X \le x\}$ that is to be estimated using a flexible family of distributions. Johnson (1949) proposed three normalizing transformations having the general form

$$Z = \gamma + \delta \cdot f\left(\frac{X-\xi}{\lambda}\right). \tag{1}$$

Here $Z$ is a standard normal random variable, $\gamma$ and $\delta$ are shape parameters, $\lambda$ is a scale parameter and $\xi$ is a location parameter. As described below, $f$ is a simple function that is chosen to complete the specification of the transformation so that a probability distribution for $X$ is also specified. Without loss of generality or flexibility, we assume that $\delta > 0$ and $\lambda > 0$. The first transformation proposed by Johnson defines the lognormal system of distributions $S_L$:

$$Z = \gamma + \delta \cdot \log\left(\frac{X-\xi}{\lambda}\right), \quad X > \xi. \tag{2}$$

The unbounded system of distributions $S_U$ is defined by

$$Z = \gamma + \delta \cdot \log\left[\left(\frac{X-\xi}{\lambda}\right) + \left\{\left(\frac{X-\xi}{\lambda}\right)^2 + 1\right\}^{1/2}\right], \quad -\infty < X < +\infty, \tag{3}$$

and the bounded system $S_B$ is defined by

$$Z = \gamma + \delta \cdot \log\left(\frac{X-\xi}{\xi+\lambda-X}\right), \quad \xi < X < \xi+\lambda. \tag{4}$$

Finally, for the sake of completeness, Johnson (1949) defined the normal $S_N$ system

$$Z = \gamma + \delta \cdot X, \quad -\infty < X < +\infty. \tag{5}$$

Given the random variable $X$, we define the moments

$$\mu_1' \equiv E(X) \quad \text{and} \quad \mu_k \equiv E[(X-\mu_1')^k], \quad 2 \le k \le 4. \tag{6}$$

The skewness and kurtosis of $X$ are

$$\beta_1 \equiv \mu_3^2/\mu_2^3 \quad \text{and} \quad \beta_2 \equiv \mu_4/\mu_2^2 \tag{7}$$

respectively. Figure 1 shows that the Johnson translation system can accommodate all possible points on the $(\beta_1, \beta_2)$ plane; this means that there is a unique Johnson distribution corresponding to each feasible combination of $\beta_1$ and $\beta_2$. The $S_L$ system is represented as a line on this plane, and the region between the limiting line $\beta_2 - \beta_1 - 1 = 0$ and the lognormal line corresponds to the $S_B$ system of distributions. The remainder of the $(\beta_1, \beta_2)$ plane corresponds to the $S_U$ family.

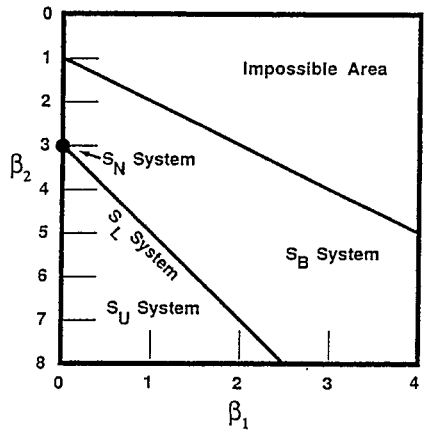



**Figure 1.** Chart for Johnson subsystem identification.

## 3. METHODS OF DISTRIBUTION FITTING WITH DATA

### 3.1. Moment Matching

Suppose we have a random sample $\{x_j : 1 \le j \le n\}$ from the target distribution $F$ that is to be approximated by a Johnson distribution. Then the sample analogs of equations (6) and (7) are

$$m_1' = n^{-1}\sum_{j=1}^{n} x_j; \quad m_k = n^{-1}\sum_{j=1}^{n}(x_j - m_1')^k, \quad 2 \le k \le 4; \tag{8}$$

and

$$\hat{\beta}_1 = m_3^2/m_2^3 \quad \text{and} \quad \hat{\beta}_2 = m_4/m_2^2. \tag{9}$$

The moment-matching technique for fitting a Johnson distribution to $F$ uses the location of the point $(\hat{\beta}_1, \hat{\beta}_2)$ in Figure 1 to identify the appropriate functional form among systems (2)–(5). The number $k$ of parameters to estimate depends on the selected system, and the principle of moment matching prescribes that the first $k$ sample moments should be equal to the corresponding population moments of the fitted theoretical distribution. The resulting system of $k$ nonlinear equations, which will be dependent on the $k$ parameters, is then solved to obtain the parameter estimates for the fitted distribution.

### 3.2. Percentile Matching

Percentile matching involves estimating $k$ required parameters by matching $k$ selected quantiles of the standard

normal distribution with corresponding quantile estimates of the target population. For given percentages $\{\alpha_j : 1 \le j \le k\}$, the corresponding quantiles $\{z_{\alpha_j}\}$ and $\{x_{\alpha_j}\}$ are given by

$$z_{\alpha_j} = \Phi^{-1}(\alpha_j) \tag{10}$$

and

$$x_{\alpha_j} = F^{-1}(\alpha_j), \tag{11}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Typical choices for $\{\alpha_j\}$ are $\{\alpha_1 = 0.25,\ \alpha_2 = 0.75\}$ when $k = 2$, $\{\alpha_1 = 0.07,\ \alpha_2 = 0.50,\ \alpha_3 = 0.93\}$ when $k = 3$, and $\{\alpha_1 = 0.07,\ \alpha_2 = 0.3118,\ \alpha_3 = 0.6882,\ \alpha_4 = 0.93\}$ when $k = 4$.

Once the functional form $f(\cdot)$ among systems (2)–(5) has been identified, the method of percentile matching attempts to solve the $k$ equations

$$z_{\alpha_j} = \gamma + \delta \cdot f[(\hat{x}_{\alpha_j} - \xi)/\lambda], \quad 1 \le j \le k, \tag{12}$$

where $\hat{x}_{\alpha_j}$ is an estimator of the quantile $x_{\alpha_j}$ based on sample data.

### 3.3. Least Squares

Least squares estimation for the Johnson system involves the minimization of the distance between a vector of "uniformized" order statistics and its corresponding expected value. (In this context, the distance between two $n \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined by a quadratic form $(\boldsymbol{a} - \boldsymbol{b})' \boldsymbol{W} (\boldsymbol{a} - \boldsymbol{b})$, where $\boldsymbol{W}$ is an appropriate $n \times n$ positive semidefinite matrix of "weights.") Given the ordered data set $x_{(1)} < \cdots < x_{(n)}$ obtained by sorting the random sample $\{x_j : 1 \le j \le n\}$ from the target distribution $F$, we can transform the $i$th sample order statistic $x_{(i)}$ into the uniformized order statistic

$$R_i(\underset{\sim}{\psi}) = \Phi\left\{\psi_1 + \psi_2 \cdot f\left[\frac{x_{(i)} - \psi_4}{\psi_3}\right]\right\}, \tag{13}$$

where $\underset{\sim}{\psi} = [\psi_1, \psi_2, \psi_3, \psi_4]' = [\gamma, \delta, \lambda, \xi]'$. Note that if the associated translation $Z = \gamma + \delta f[(X - \xi)/\lambda]$ yields a standard normal variate exactly, then $R_i(\underset{\sim}{\psi})$ has the distribution of $U_{(i)}$, the $i$th order statistic in a sample of $n$ random numbers from $U(0, 1)$, the uniform distribution on the unit interval (0, 1). Note moreover that $U_{(i)}$ has expected value $\rho_i = E[U_{(i)}] = i/(n+1)$. The difference $\epsilon_i(\underset{\sim}{\psi}) = R_i(\underset{\sim}{\psi}) - \rho_i$ represents the random deviation between the observed and expected values of the $i$th uniformized order statistic, so that $E[\epsilon_i(\underset{\sim}{\psi})] = 0$. The covariance between $\epsilon_i(\underset{\sim}{\psi})$ and $\epsilon_j(\underset{\sim}{\psi})$ is

$$\mathrm{Cov}[\epsilon_i(\underset{\sim}{\psi}), \epsilon_j(\underset{\sim}{\psi})] = \frac{i(n-j+1)}{(n+1)^2(n+2)}, \quad 1 \le i \le j \le n. \tag{14}$$

Let $\underset{\sim}{R}(\underset{\sim}{\psi}) = [R_1(\underset{\sim}{\psi}), \ldots, R_n(\underset{\sim}{\psi})]'$, $\underset{\sim}{\rho} = [\rho_1, \ldots, \rho_n]'$, and $\underset{\sim}{\varepsilon}(\underset{\sim}{\psi}) = \underset{\sim}{R}(\underset{\sim}{\psi}) - \underset{\sim}{\rho}$. In addition, let $\boldsymbol{V} = \|\mathrm{Cov}[\epsilon_i(\underset{\sim}{\psi}), \epsilon_j(\underset{\sim}{\psi})]\|$ denote the covariance matrix of the errors; and let $\boldsymbol{D}$ denote the diagonal matrix obtained by setting the off-diagonal elements of $\boldsymbol{V}$ to zero—that is, $\boldsymbol{D} = \mathrm{diag}\{\mathrm{Var}[\epsilon_1(\underset{\sim}{\psi})], \ldots, \mathrm{Var}[\epsilon_n(\underset{\sim}{\psi})]\}$.

The least squares approach to the parameter estimation problem can be stated as

$$\begin{aligned} &\underset{\underset{\sim}{\psi}}{\text{minimize}} \quad s(\underset{\sim}{\psi}) \equiv [\underset{\sim}{\varepsilon}(\underset{\sim}{\psi})]' \boldsymbol{W} [\underset{\sim}{\varepsilon}(\underset{\sim}{\psi})] \\ &\text{subject to:} \\ &\psi_2 \quad > 0, \\ &\psi_3 \begin{cases} > 0 & \text{for } S_U, \\ > x_{(n)} - \psi_4 & \text{for } S_B, \\ = 1 & \text{for } S_L \text{ and } S_N, \end{cases} \\ &\psi_4 \begin{cases} < x_{(1)} & \text{for } S_L \text{ and } S_B, \\ = 0 & \text{for } S_N. \end{cases} \end{aligned} \tag{15}$$

When the weight matrix $\boldsymbol{W} = \boldsymbol{I}$ in (15), we obtain the ordinary least squares (OLS) estimators for $\gamma$, $\delta$, $\lambda$ and $\xi$. Since the errors $\{\epsilon_i(\underset{\sim}{\psi})\}$ are neither independent nor homoscedastic, weighted least squares (WLS) parameter estimators are of interest. The standard approach in this situation is to take $\boldsymbol{W} = \boldsymbol{V}^{-1}$ in (15). However in small to medium samples, this approach can yield relatively large bias in the fitted CDF as well as in the WLS parameter estimators. As an alternative approach to WLS estimation, Swain, Venkatraman and Wilson (1988) took $\boldsymbol{W} = \boldsymbol{D}^{-1} = \mathrm{diag}\{1/\mathrm{Var}[\epsilon_1(\underset{\sim}{\psi})], \ldots, 1/\mathrm{Var}[\epsilon_n(\underset{\sim}{\psi})]\}$; and in a wide variety of data sets, they obtained WLS fits for Johnson distributions that are comparable (and often superior to) the fits obtained by the other methods described in this paper.

### 3.4. Minimum $L_p$ Norm Estimation

In this section we discuss the use of $L_1$ and $L_\infty$ norms in estimating the parameters of the Johnson distribution. The principle is to minimize some metric based on the distance between the empirical CDF $F_n$ and the fitted CDF $\hat{F}$. If $1 \le p < \infty$, then the $L_p$ norm for the distance between $F_n$ and $\hat{F}$ is defined as

$$\|F_n - \hat{F}\|_p = \left[\int_{-\infty}^{\infty} |F_n(x) - \hat{F}(x)|^p \, d\hat{F}(x)\right]^{1/p}. \tag{16}$$

To fit a Johnson distribution by minimum $L_p$ norm estimation, we take

$$\hat{F}(x; \underset{\sim}{\psi}, f) \equiv \Phi\left[\psi_1 + \psi_2 f\left(\frac{x-\psi_3}{\psi_4}\right)\right] \tag{17}$$

in (16) and minimize $\|F_n - \hat{F}\|_p$ over all values of the parameter vector $\underset{\sim}{\psi} = [\psi_1, \psi_2, \psi_3, \psi_4] = [\gamma, \delta, \lambda, \xi]$ that are feasible for the selected translation function $f(\cdot)$.

To fit a Johnson distribution based on the minimum sum of absolute errors, the $L_1$ norm

$$\|F_n - \hat{F}\|_1 \equiv \int_{-\infty}^{\infty} |F_n(x) - \hat{F}(x)| \, d\hat{F}(x) \tag{18}$$

can be evaluated as follows. For notational convenience, let $a_j = \hat{F}[x_{(j)}]$, for $j = 1, ..., n$, and let $b_j = (j-1)/n$, for $j = 2, ..., n$. Then we can evaluate the $L_1$ norm as

$$\|F_n - \hat{F}\|_1 = \tfrac{1}{2}a_1^2 + \sum_{j=2}^{n} Q_j + \tfrac{1}{2}(1-a_n)^2, \tag{19}$$

where

$$Q_j = \begin{cases} \frac{1}{2}(a_j^2 - a_{j-1}^2) - b_j(a_j - a_{j-1}) & \text{if } b_j \le a_{j-1}, \\ \frac{1}{2}(a_j^2 + a_{j-1}^2) - b_j(a_j + a_{j-1}) + b_j^2 & \text{if } a_{j-1} < b_j \le a_j, \\ b_j(a_j - a_{j-1}) - \frac{1}{2}(a_j^2 - a_{j-1}^2) & \text{if } b_j > a_j. \end{cases} \tag{20}$$

The $L_\infty$ norm is defined as

$$\|F_n - \hat{F}\|_\infty \equiv \sup_{-\infty < x < \infty} |F_n(x) - \hat{F}(x)|, \tag{21}$$

and is evaluated as $\max\{D_n^+, D_n^-\}$, where

$$D_n^+ = \max_{1 \le j \le n}\left\{\frac{j}{n} - \hat{F}[x_{(j)}]\right\} \text{ and} \tag{22}$$

$$D_n^- = \max_{1 \le j \le n}\left\{\hat{F}[x_{(j)}] - \frac{j-1}{n}\right\}. \tag{23}$$

Note that (21) is simply the Kolmogorov-Smirnov statistic corresponding to the fitted distribution $\hat{F}$.

## 4. USING THE FITTR1 SOFTWARE

The program starts execution by prompting the user for: (a) the name of a "script" file that will maintain a record of the entire interactive session, and (b) the name of the input data file. (At a later time, the first file can be printed out to provide a hard copy of the results of the interactive session.) After the user has responded to these prompts, the data set is read in, some basic data checks are performed, and a variety of sample statistics are calculated and displayed. FITTR1 then computes and displays moment matching estimates for the parameters, and then waits for a command from the user. The available commands are explained below. A sample interactive session with FITTR1 is shown in Appendix A. See Venkatraman and Wilson (1988) for a more complete discussion of FITTR1.

`stat` **Command.** This command displays the computed sample statistics—namely, the mean, standard deviation, skewness $\hat{\beta}_1$, kurtosis $\hat{\beta}_2$, range, minimum and maximum of the data set being analyzed. (See equations (8) and (9) for precise definitions of $\hat{\beta}_1$ and $\hat{\beta}_2$.)

`fit ijkm` **Command.** This command fits a new distribution to the sample data as specified by the fitting code `ijkm`. The four digit code `ijkm` is parsed to obtain the values of the variables `i`, `j`, `k` and `m`. The table below describes the values of these variables.

| | |
|---|---|
| `i` | 0 = automatic distribution selection<br>1 = $S_L$ distribution<br>2 = $S_U$ distribution<br>3 = $S_B$ distribution<br>4 = $S_N$ distribution |
| `j` | 0 = no end point known<br>1 = lower end point known<br>2 = upper end point known<br>3 = both end points known |
| `k` | 0 = compute starting parameter values<br>1 = use previous parameter values |
| `m` | 0 = moment matching<br>1 = percentile matching<br>2 = ordinary least squares estimation of the CDF<br>3 = weighted least squares estimation of the CDF<br>4 = $L_1$ estimation of the CDF (minimize sum of absolute errors)<br>5 = $L_\infty$ estimation of the CDF (minimize maximum absolute error) |

The command `fit 0` has special meaning: it identifies the type of distribution to fit based on the location of the point $(\hat{\beta}_1, \hat{\beta}_2)$ and also performs parameter estimation by moment matching. All valid `fit` commands except the `fit 0` command query the user for additional fitting information when the program is in the verbose mode (see below).

`par` **Command.** This command displays the parameters of the fitted distribution—namely the type of distribution that

has been fitted, the fitting method, and the current values of the parameters $\gamma$, $\delta$, $\lambda$ and $\xi$.

`gof` **Command.** This command invokes goodness of fit testing using the latest set of estimated parameters. The $\chi^2$ goodness of fit test and the Kolmogorov-Smirnov test are performed and the results displayed on the screen.

`cdf` **Command.** This command creates files of fitted and empirical CDFs that can be used as input to plotting packages for display on high-resolution output devices—usually color monitors and/or laser printers. The plot-files generated by this command are text files—that is, they contain free-format numbers specifying the appropriate abscissa (X) and/or ordinate (Y) values for the points to be plotted. To create the desired graphs, the user may pass these files to any available plotting package. The specified points should be connected with lines to obtain the desired graph.

`pdf` **Command.** This command creates files of histogram values and fitted density function values that can be used as input to plotting packages for display on high-resolution output devices—usually color monitors and/or laser printers. The plot-files generated by this command are text files—that is, they contain free-format numbers specifying the appropriate abscissa (X) and/or ordinate (Y) values for the points to be plotted. To create the desired graphs, the user may pass these files to any available plotting package. The specified points should be connected with lines to obtain the desired graph.

`mode` **Command.** This command will toggle the program from the verbose mode to the terse mode and vice versa. In verbose mode, the user has an opportunity to change any of the parameters of the available fitting procedure (for example, error tolerances and iteration limits). In terse mode, the user is not prompted for these parameters; instead, standard default values are taken.

`spec` **Command.** This command allows the user to change the specifications of the available fitting procedures (for example, error tolerances and iteration limits). It is different from the mode command in that the program will be in terse mode after execution of this command.

`next` **Command.** When there is more than one set of data to be analyed during one interactive session, this command is used to proceed to the next data set within the input file.

`comm` **Command.** This command allows the user to insert comments into the script file. Such comments may be useful for future reference when reviewing a printed copy of the script file.

`what` **Command.** This command displays the name of the data set being currently analyzed.

`help` **Command.** This command displays the available list of commands.

`stop` **Command.** This command provides for a quick and graceful exit from the interactive program.

## 5. SUBJECTIVE DISTRIBUTION FITTING

### 5.1. The Johnson $S_B$ System Revisited

In developing a visual interactive approach to fitting Johnson distributions when little or no data are available, we confined outselves to the $S_B$ subsystem of the Johnson translation system because it matches well our notions of the general characteristics of many potential envisioned target distributions. $S_B$ distributions are bounded and they are capable of matching the skewness and kurtosis of most practical distributions. Real-world measurements are always bounded, even if only by the limits of technology. The $S_B$ is capable of assuming U-shaped forms, but because we seek to model fundamental input processes, we consider only unimodal Johnson $S_B$ distributions to be appropriate.

Now if $X$ has an $S_B$ distribution with parameters $\gamma$, $\delta$, $\lambda$, and $\xi$, then the "standardized" variate

$$Y = \left(\frac{X-\xi}{\lambda}\right) \tag{24}$$

lies between zero and one with the same shape parameters as $X$ but with location parameter zero and scale parameter one. The density of $Y$ is then given by (Johnson 1949):

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \frac{1}{y(1-y)} \exp\left\{-\frac{1}{2}\left[\gamma + \delta \cdot \log\left(\frac{y}{1-y}\right)\right]^2\right\},$$

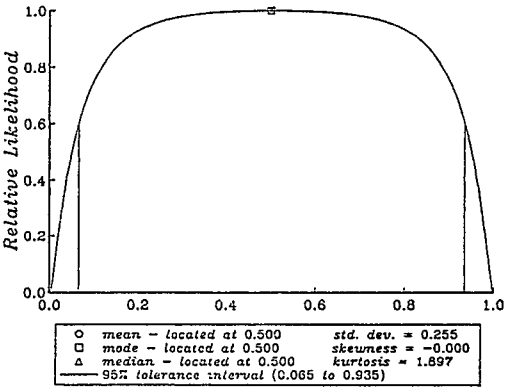
$$\delta > 0, \quad -\infty < \gamma < \infty. \tag{25}$$

The density $p(y)$ can take a surprising variety of shapes. Figure 2 presents some examples of $S_B$ distributions. Figure 2a has "broad shoulders" but is symmetric. In Figure 2b, the distribution is negatively skewed and broadly dispersed. Figure 2c presents a symmetric distribution with nearly normal shoulders that is a common shape for many practical applications. Figure 2d and Figure 2e show some asymmetric distributions that are typical of the activity-time distributions used in many simulation studies. Figure 2f displays a density shape within the scope of the Johnson $S_B$ function, but perhaps unlikely to describe a real-world process.

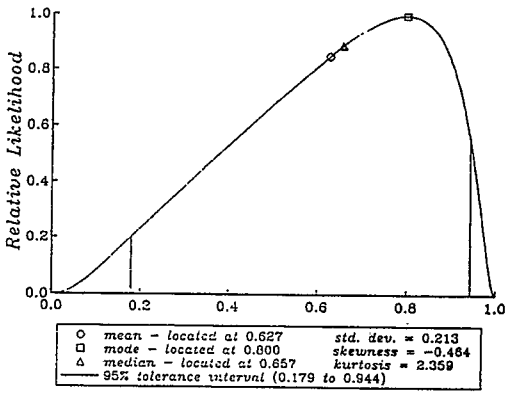
Historically the $S_B$ distribution has been difficult to work with because of the mathematically complex relationship of its shape to the parameters $\gamma$ and $\delta$. There are no convenient explicit equations relating the mode or any of the moments of an $S_B$ distribution to its parameter values. Therefore, for the distribution to be useful, the shape parameters must be recast into familiar terms that correspond to the envisioned characteristics of a target distribution.

### 5.2. Subjective Specification of $S_B$ Distributions

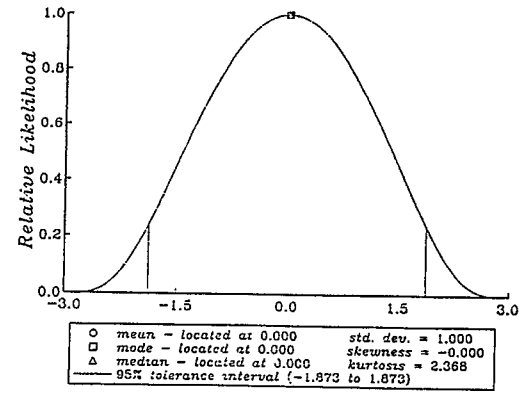
Describing a distribution in sufficient detail to permit its




(a)

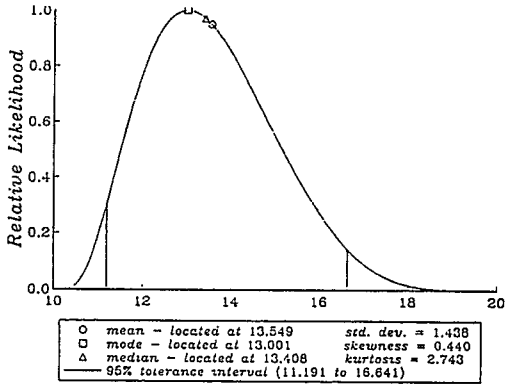



(b)

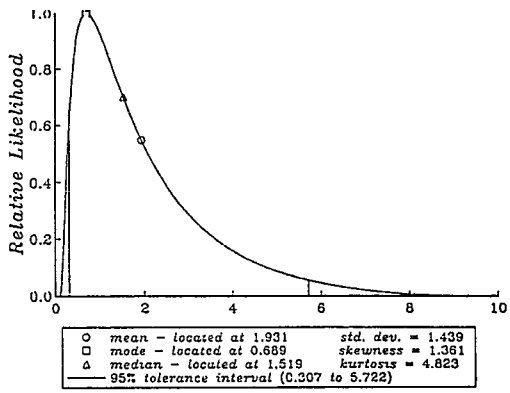



(c)

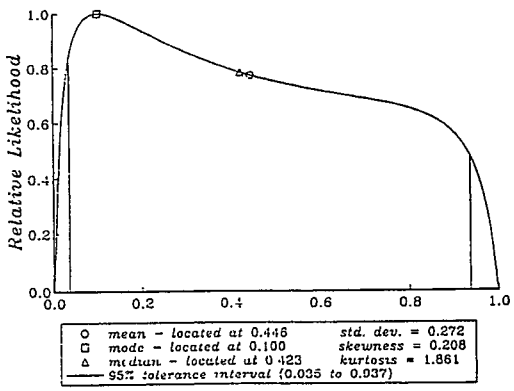
**Figure 2.** Examples of $S_B$ Distributions.




(d)




(c)




(f)

**Figure 2 (continued).** Examples of $S_B$ Distributions.

approximation by a parameterized functional form is a nontrivial task, even when restricting consideration to smooth, thin-tailed, unimodal densities. Typically, numerical measures of central tendency, variability, and other complex nuances of a density's "shape" are employed. Familiar examples include the mean, standard deviation, skewness, and kurtosis. While these statistical descriptors are easy to obtain from raw data, they are difficult to estimate for an envisioned distribution. The mean of an asymmetric, bounded distribution rarely coincides with other common measures of central tendency such as the mode, median, and midrange; and inexperienced estimators are frequently unable to make the proper distinctions among these measures (Spencer 1963). Subjective estimates of means are influenced by distributional variance and skewness, and may be biased (Beach and Swenson 1966). Intuitive variability estimates are inappropriately correlated with the magnitude of the mean (Lathrop 1967). Descriptors defined in terms of a distribution's higher moments are for practical purposes unavailable except by calculation from data.

We believe that a target distribution's mode is more easily specified than any other measure of central tendency. It is a natural, easily understood "best guess" of what one is most likely to see on any single realization of the target random variable. Unlike the mean, the mode is not necessarily tied to the behavior of the distribution in its tails; and unlike the median, it is not necessarily tied to the degree of asymmetry in the distribution. For skewed distributions, estimates of the mode and median are demonstrably better than estimates of the mean (Peterson and Miller 1964).

In addition to the end points and mode, which suffice for the triangular distribution, at least one other descriptor is necessary to uniquely specify the more complex functional form of the Johnson $S_B$ distribution. Fortunately, percentile points for envisioned distributions can be subjectively estimated with accuracy (Kahneman, Slovic, and Tversky 1982). An $S_B$ distribution can be uniquely determined by its

end points together with (a) two percentile points, or (b) the mode and one percentile point. Doubilet et al. (1985) have developed a method for the estimation of a logistic-normal distribution (which is a Johnson $S_B$ distribution with $\xi = 0.0$ and $\lambda = 1.0$) from the mean and either the 5th or 95th percentile point. In an extreme case, if one provides multiple percentile points, an approximation to the distribution's cumulative distribution function or probability density function can be generated directly, but such information demands are in most cases unrealistic.

Some combinations of desired distributional characteristics describe "impossible" distributions, which cannot be approximated by the Johnson $S_B$, or any other smooth *unimodal* density. One example is a distribution bounded between 0.0 and 1.0, with a mean of 0.2 and a mode of 0.4. Even when an $S_B$ distribution can be found with the desired characteristics, the corresponding density may have a shape quite unlike what the modeler imagines, as with a distribution bounded between 0.0 and 1.0 with a mean 0.45 and a mode of 0.1 (see Figure 2f). If a modeler fails to describe the target distribution accurately (that is, if he specifies characteristics inappropriate for the envisioned distribution), then the only way that this can be detected in the absence of data is by visual inspection of the resulting density's shape.

## 6. USING THE VISIFIT SOFTWARE

VISIFIT combines flexible numerical description with interactive visual curve modification to capture and refine available subjective information into a parameterized Johnson $S_B$ density. Primary design goals were ease of use, high speed on inexpensive microcomputers, and the requirements of minimal information and information processing from the user.

### 6.1. Specifying the Desired Characteristics

At the outset of the interaction with VISIFIT, the user must specify the upper (maximum) and lower (minimum) end points of the distribution. These are subject to later modification, if desired. Next the modeler is prompted for values of any *two* of the following characteristics:

(a) Mode

(b) Mean

(c) Median

(d) Arbitrary percentile point(s)

(e) Standard deviation

Significantly, the user is free to provide two arbitrary, asymmetric percentile points, such as the 10th and 25th percentile points. Unlike other algorithms (Mage 1980), there is no requirement that the four input points (two end points and two percentile points) must correspond to equidistant normal deviates. As a default, motivated by the three-parameter specification of a PERT-type estimate (Wilson et al. 1982), the standard deviation can be optionally chosen to be one-sixth of the range. By accepting a variety of different descriptions, we minimize the need for processing information prior to its input. Thee modeler is free to use whatever is convenient, familiar, or easily understood.

When the desired characteristics are entered, VISIFIT computes the parameters of the $S_B$ distribution that most closely match those characteristics. Several miscellaneous numerical techniques are employed in this calculation, and all are detailed in DeBrota et al. (1988).

### 6.2. Interactive Curve Modification

Once the parameterization of the fitted $S_B$ density is complete, the user is immediately presented with the distribution's actual shape on a graphical display screen. Such visual feedback will sometimes suggest to the user different values for the characteristics of the target random variable $X$ than were originally chosen. From these revised specifications a new set of parameter values is generated, and

then a new fitted density is presented to the user (see Figure 3). Cyclic interaction permits the user to experiment with different curve shapes until a satisfactory one is obtained.

VISIFIT also provides a still simpler scheme of interactive curve shape modification that frees the user from having to deal with numerical input by providing single-keystroke commands that directly manipulate the shape of the displayed curve. The modeler can adjust the shape of a displayed Johnson $S_B$ curve by trial-and-error until he is satisfied with the way it looks. Motivated by our belief in the universal ease of specifying the mode, width, and percentile points of a distribution, we implemented various single-keystroke commands producing the following immediate effects:

(a) Move the mode towards the upper bound

(b) Move the mode towards the lower bound

(c) Increase the width of the curve

(d) Decrease the width of the curve

(e) Move the 2.5th percentile point to the right

(f) Move the 2.5th percentile point to the left

(g) Move the 97.5th percentile point to the left

(h) Move the 97.5th percentile point to the left

The magnitude of the change (in the direction indicated by the choice of control key pressed) is determined by an adaptive seeking strategy. The modeler need only indicate the direction of desired change from each displayed curve to the next. The curve can be updated approximately twice each second on an IBM PC/AT class machine with a 80287 numeric coprocessor, and thus the overall process of changing a curve, even drastically from an initial shape, takes at most a few seconds in the hands of an experienced useer.

Modification of the end points may be accomplished in two ways. The scale of the $X$ axis may be changed, preserving the shape of the distribution while altering the absolute values of the end points. This rescaling also changes the absolute values of the mode and width. Alternatively, the absolute values of the mode and width may be preserved during a change in the end points, in which case a new curve with a visually different shape is obtained.

## 7. CONCLUSIONS

As a general tool for simulation input modeling, the main advantage of the Johnson translation system of probability distributions is its flexibility in approximating the target distributions that arise in a diversity of applications. The main disadvantage of the Johnson system is its analytical intractability. The software packages FITTR1 and VISIFIT have been specifically designed to alleviate this latter limitation.

Another attractive feature of the Johnson system is that it can be extended easily to provide systems of multivariate distributions, and this property should enable us to conveniently model dependencies among the inputs to a simulation. Multivariate extensions of FITTR1 and VISIFIT are currently being developed (Venkatraman 1988).

In many simulation studies the analyst has both sample data and subjective information about the input process to be modeled, and he would like to use both sources of information in an integrated procedure for building a simulation input model. We are currently developing methodology and software that effectively synthesizes FITTR1 and VISIFIT to provide a unified approach to input modeling.

## ACKNOWLEDGMENTS

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

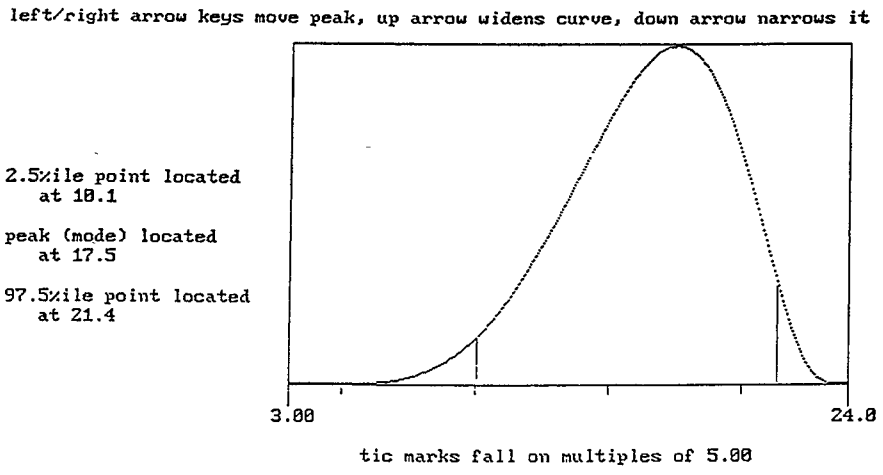



**Figure 3.** VISIFIT's graphical display.

## APPENDIX A: SAMPLE INTERACTIVE SESSION

```
                    **** FITTR1 Version 1.1 ****
                            May 1988

 Enter file name for printable copy of session
 >> example

 Enter file name for data set to be fitted
 >> gluc.dat

          Fitting population Glucose Data Set

                     *** Sample statistics ***
     Sample size =   80
     Mean        =    226.2        Std. Dev. =    122.7
     Skewness    =    1.340        Kurtosis  =    4.205
     Minimum     =    34.00        Maximum   =    666.0
     Range       =    632.0

               *** Infeasible moment matching estimates ***

                     *** Parameter estimates ***
     Distribution: SB
     Method: Moment matching
     Gamma  =     1.653        Delta =     1.059
     Lambda =     855.7        Xi    =     47.24

 >> fit 3001

     Requested fit characteristics
        Distribution: SB
        Method: Percentile matching
        Number of parameters: 4
        Will compute starting parameter values
        Neither end point known
```

```
                    *** Goodness of fit tests ***
    Kolmogorov-Smirnov statistic    0.0942
    Significance probability        0.4762

    Chi-squared statistic           7.3000
    Significance probability        0.1209

>> mode

                    *** Verbose mode on ***

>> fit 3002

     Requested fit characteristics
        Distribution: SB
        Method: Ordinary least squares
        Number of parameters: 4
        Will compute starting parameter values
        Neither end point known

 Enter no. of accurate digits in l.s. methods
 Current values are     4
 For no change, press RETURN

 >>

 Enter no. of function evals. in l.s. methods
 Current values are   500
 For no change, press RETURN

 >>

 Enter rel. function tol. for l.s. methods
 Current values are   0.1000E-03
 For no change, press RETURN

 >>

               *** Switching to Nelder-Mead algorithm ***

               *** Least squares iteration limit ***

     Least squares minimum SSE   0.2905E-01

                    *** Goodness of fit tests ***
     Kolmogorov-Smirnov statistic    0.0612
     Significance probability        0.9257

     Chi-squared statistic           4.8250
     Significance probability        0.3057

 >> mode

                    *** Terse mode on ***

 >> cdf

 Do you want a table of CDF values?

 >> yes

 For empirical vs. fitted CDF over the range
 low (incr) high, enter values for low, incr, high

 >> 30  32  670

CDF values for Glucose Data Set
Fitting method: Ordinary least squares
Distribution: SB
Gamma  =     3.676        Delta =     1.344
Lambda =     2623.        Xi    =     34.00
```

```
     -----X-----      --Empirical--      ---Fitted---

        30.00            0.0000E+00         0.0000E+00
        62.00            0.2500E-01         0.7945E-02
        94.00            0.6250E-01         0.8528E-01
        126.0            0.2000             0.2180
        158.0            0.3500             0.3592
        190.0            0.5125             0.4861
        222.0            0.6125             0.5924
        254.0            0.6750             0.6782
        286.0            0.7250             0.7464
        318.0            0.7500             0.8002
        350.0            0.8500             0.8424
        382.0            0.8875             0.8755
        414.0            0.9125             0.9015
        446.0            0.9250             0.9219
        478.0            0.9625             0.9380
        510.0            0.9750             0.9507
        542.0            0.9750             0.9607
        574.0            0.9875             0.9687
        606.0            0.9875             0.9750
        638.0            0.9875             0.9800
        670.0             1.000             0.9840

Enter file name if CDF values are to be saved
Press RETURN otherwise

>> gluc.tbl

         *** CDF file named gluc.tbl        has been created ***

Do you want a plot of CDF values?

>> yes

For the empirical and fitted CDFs each considered
separately, you can request:
(a) 1 plot-file with all (X, Y) pairs for the
    CDF to be plotted; or
(b) 2 plot-files of X- and Y-values for the
    CDF to be plotted.

Enter a or b for the desired option

>> a

Enter file name for (X, Y) values of fitted CDF
>> gluc.fdf

         *** CDF file named gluc.fdf        has been created ***

Enter file name for (X, Y) values of empirical CDF

>> gluc.edf

         *** CDF file named gluc.edf        has been created ***

>> next

Is the next data set in a different file?

>> yes

Enter file name for data set to be fitted
>> lsedf1

         Fitting population Ozturk and Dale Data Set

                  *** Sample statistics ***
   Sample size =   75
   Mean        =    5.109        Std. Dev. =    1.006
   Skewness    =    1.019        Kurtosis  =    3.344
   Minimum     =    3.836        Maximum   =    7.863
   Range       =    4.027
```

```
                *** Parameter estimates ***
Distribution: SB
Method: Moment matching
Gamma  =    1.088        Delta =    0.7871
Lambda =    5.008        Xi    =     3.824

                *** Goodness of fit tests ***
Kolmogorov-Smirnov statistic    0.0462
Significance probability        0.9971

Chi-squared statistic           7.2000
Significance probability        0.1257

>> comm This is a demo of the FITTR1 comm command

>> stop
```

## REFERENCES


Beach, L. R. and Swenson, R. G. (1966). Intuitive estimation of means. *Psychon. Sci.* **5**, 161-162.

DeBrota, D., Roberts, S. D., Dittus, R. S., and Wilson, J. R. (1988). Visual interactive fitting of probability distributions. Research Memorandum No. 88-3, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

Doubilet, P., Begg, C. B., Weinstien, M. C., Braun, P., and McNeil, B. J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation, a practical approach. *Medical Decision Making* **5**, 157-177.

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

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases.* Cambridge University Press.

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology* **73**, 498-502.

Mage, D. T. (1980). An explicit solution for $S_B$ parameters using four percentile points. *Technometrics* **22**, 247-251.

Peterson, C. R. and Miller, A. (1964). Mode, median, and mean as optimal strategies. *Journal of Experimental Psychology* **68**, 363-367.

Roberts, S. D. (1983). *Simulation Modeling and Analysis with INSIGHT.* Regenstrief Institute for Health Care, Indianapolis, Indiana.

Schmeiser, B. W. (1977). Methods for modelling and generating probabilistic components in digital computer simulation when the standard distributions are not adequate: A survey. *Proceedings of the 1977 Winter Simulation Conference*, Highland, Sargent and Schmidt (eds.), 51-55. The Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Spencer, J. (1963). A further study of estimating averages. *Ergonomics* **6**, 255-265.

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.

Wilson, J. R., Vaughan, D. K., Naylor, E. and Voss, R. G. (1982). Analysis of Space Shuttle ground operations. *Simulation* **38**, 187-203.

## AUTHORS' BIOGRAPHIES

DAVID J. DEBROTA is a Lecturer in Medicine at the Indiana University School of Medicine and a Fellow in General Internal Medicine. He is Board Certified in Internal Medicine. He received his B.S. in Chemistry and Physics from Butler University and his M.D. from Indiana University. He has over twelve years experience in microcomputing and his research interests are in decision support systems, medical decision making, and artifical intelligence.

David J. DeBrota
Regenstrief Institute for Health Care, 6th Floor
1001 West 10th Street
Indianapolis, IN 46202, U.S.A.
(317) 630-8245

ROBERT S. DITTUS is an Assistant Professor at the Indiana University School of Medicine. He is Director of the Fellowship in General Internal Medicine Training program and the Clinical Practice Analysis Section of the Regenstrief Institute. He serves on the Editorial Board of *Medical Decision Making* and is the author of several articles in medical decision making and clinical epidemiology. He received his B.S.I.E. from Purdue University, M.P.H. from North Carolina, and M.D. from Indiana University.

Robert S. Dittus
Regenstrief Institute for Health Care, 5th Floor
1001 West 10th Street
Indianapolis, IN 46202, U.S.A.
(317) 630-7447
dittus@gb.ecn.purdue.edu

STEPHEN D. ROBERTS is Professor of Industrial Engineering at Purdue University and Professor of Internal Medicine at the Indiana University School of Medicine. His academic and teaching responsibilities are in simulation modeling. He is Director of Health Systems Research at the Regenstrief Institute for Health Care, focusing on simulation of medical decisions. His methodological research is in simulation language design and includes INSIGHT (INS), a general purpose, discrete event language, and SLN for the Simulation of Logical Networks. He is also a principal in SysTech, Inc. which distributes the simulation languages and consults on their application.

He received his BSIE, MSIE, and PhD in Industrial Engineering from Purdue University and has held research and faculty positions at the University of Florida. He is active in several professional societies and in addition to making presentations and chairing sessions at conferences, he was *Proceedings* Editor for WSC '83, Associate Program Chairman for WSC '85, and Program Chairman for WSC '86. Presently he is a member of the Board of Directors of WSC '88, Chairman of SIGSIM, and Area Editor of *Simulation.*

Stephen D. Roberts
SysTech, Inc.
P.O. Box 509203
Indianapolis, IN 46250, U.S.A.
(317) 842-6586

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 has 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 Enginering Science in 1975, and an MS in Mechanical Engineering in 1977 from the University of Notre Dame. He received his Ph.D. in

Industrial Engineering from Purdue University in 1982. His currrent 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
jswain%gtri01.bitnet

SEKHAR VENKATRAMAN is an Assistant Professor of Industrial Engineering at Wichita State University. He received a B.S. in mechanical engineering from Annamalai University (India) in 1980, an M.S.E. in operations research from The University of Texas at Austin in 1983, and a Ph.D. in industrial engineering from Purdue University in 1988. His current research interests include variance reduction techniques and multivariate input modeling. He is a member of ACM, ASA, ORSA and SCS.

Sekhar Venkatraman
Department of Industrial Engineering
Wichita State University
Wichita, KS 67208, U.S.A.
(317) 494-6403
venkatra%twsuvm.bitnet

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 in 1970, and M.S. and Ph.D. degrees in industrial engineering from Purdue University in 1977 and 1979 respectively. He has been involved in various simulation studies while working as a research analyst for the Houston Lighting & Power Company (1970–72) and while serving as a U.S. Army officer (1972–75). From 1979 to 1984, he was an Assistant Professor in the Mechanical Engineering Department of The University of Texas at Austin. His current research interests include simulation output analysis, variance reduction techniques, ranking-and-selection procedures, and stopping rules. He is a member of ASA, IIE, ORSA, SCS, and TIMS.

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