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

MODELING INPUT PROCESSES WITH JOHNSON 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
Trans World Airlines, Inc.
110/3 South Bedford Road
Mt. Kisco, NY 10549, 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. For situations in which little or no sample information is available, we have developed a visual interactive method to estimate bounded Johnson distributions subjectively; and we have implemented this technique in VISIFIT, a public-domain software package. For fitting all types of Johnson distributions based on sample data, we have implemented several new statistical-estimation methods as well as some standard techniques in FITTR1, another public-domain software package. We present several examples illustrating the use of VISIFIT and FITTR1 for simulation input modeling.

1. INTRODUCTION

In modeling and simulation of stochastic systems, a major problem is the selection of probability distributions that will adequately represent the input processes (populations) driving the simulation model. When it is feasible to collect sample data from a target population, simulation input modeling is usually accomplished by (a) hypothesizing a standard parametric distribution to describe that population, (b) estimating the associated parameters based on the sample information, and (c) performing diagnostic checks to assess the adequacy of the fit based on a comparison of the sample distribution with the fitted distribution. In the absence of sample information for parameter estimation and goodness-of-fit testing, practitioners usually try to elicit expert opinions about enough numerical characteristics of the target population (for example, the mode, the end points, or the mean) to specify uniquely a member of the hypothesized distribution family. Thus a fundamental consideration in simulation input modeling is the initial selection of a *flexible* family of distributions – that is, a family capable of yielding a wide variety of distributional shapes; unfortunately, many of the standard parametric distributions have an extremely limited range of possible shapes (Schmeiser 1977).

In this paper we discuss the use of the Johnson (1949) translation system of distributions to model continuous univariate populations. (The term *method of translation* refers to the transformation of a continuous random variable to a standard normal variate as explained below.) By incorporating four highly flexible families of distributions (specifically, the lognormal, unbounded, bounded, and normal families), the Johnson system can fit any distribution up to its first four moments; and in practice the Johnson system has been used successfully in a wide variety of disciplines. In this paper we describe a visual interactive method for subjective estimation of bounded Johnson distributions when little or no sample sample information is available, and we discuss the software package VISIFIT in which this visual approach has been implemented. For fitting all four families of Johnson distributions to sample data, we present the interactive software package FITTR1. Both VISIFIT and FITTR1 are in the public domain, run on microcomputers, and are available from the authors upon request.

This paper is organized as follows. Section 2 contains a brief introduction to the Johnson system of distributions. The main issues arising in subjective estimation of probability distributions are discussed in Section 3. The operation of VISIFIT is detailed in Section 4. In Section 5 we survey the

methods for distribution identification and parameter estimation that have been implemented in FITTR1. The operation of FITTR1 is described 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 cumulative distribution function (CDF) $F(x) = \Pr\{X \le x\}$ and probability density function (PDF) $p(x) = F'(x)$ that are to be estimated using a flexible family of distributions. Johnson (1949) proposed four *normalizing translations* with the general form

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

where Z is a standard normal random variable, γ and δ are shape parameters, λ is a scale parameter, ξ is a location parameter, and $f()$ is one of the following functions:

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

Without loss of generality, we assume that $\delta > 0$ and $\lambda > 0$. We also observe the following standard conventions: (a) for the S_N (normal) family, we always take $\lambda \equiv 1$ and $\xi \equiv 0$; and (b) for the S_L (lognormal) family, we always take $\lambda \equiv 1$. Note that once the functional form $f()$ has been identified and the parameters $\{\gamma, \delta, \lambda, \xi\}$ have been estimated by one of the fitting techniques described below, generating random variates from the fitted Johnson distribution is straightforward – given a random sample Z from the standard normal distribution, we compute the corresponding realization of the target variate X as follows:

$$X = \xi + \lambda \cdot f^{-1}\left(\frac{Z - \gamma}{\delta}\right), \tag{3}$$

where

$$f^{-1}(z) = \begin{cases} e^z, & \text{for the } S_L \text{ (lognormal) family,} \\ \frac{1}{2}\left(e^z - e^{-z}\right), & \text{for the } S_U \text{ (unbounded) family,} \\ 1/(1 + e^{-z}), & \text{for the } S_B \text{ (bounded) family,} \\ z, & \text{for the } S_N \text{ (normal) family.} \end{cases} \tag{4}$$

Thus the practitioner merely needs a standard normal random variate generator in order to sample variates from any of the Johnson distribution families. Both univariate and multivariate Johnson distributions are available in the INSIGHT simulation language (Roberts 1983).

For each distribution family in the Johnson system, the corresponding probability density function (PDF) is

$$p(x) = \frac{\delta}{\lambda\sqrt{2\pi}} f'\left(\frac{x - \xi}{\lambda}\right) \exp\left\{-\frac{1}{2}\left[\gamma + \delta \cdot f\left(\frac{x - \xi}{\lambda}\right)\right]^2\right\} \tag{5}$$

for all $x \in \mathbf{H}$, where

$$f'(y) = \begin{cases} 1/y, & \text{for the } S_L \text{ (lognormal) family,} \\ 1/\sqrt{y^2 + 1}, & \text{for the } S_U \text{ (unbounded) family,} \\ 1/[y(1-y)], & \text{for the } S_B \text{ (bounded) family,} \\ 1, & \text{for the } S_N \text{ (normal) family,} \end{cases} \tag{6}$$

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

$$\mathbf{H} = \begin{cases} [\xi, +\infty) & \text{for the } S_L \text{ (lognormal) family,} \\ (-\infty, +\infty) & \text{for the } S_U \text{ (unbounded) family,} \\ [\xi, \xi + \lambda] & \text{for the } S_B \text{ (bounded) family,} \\ (-\infty, +\infty) & \text{for the } S_N \text{ (normal) family.} \end{cases} \tag{7}$$

Thus we see that the terms *bounded* and *unbounded* describe the support of the density $p(x)$, which is also the space of the associated random variable X. As an aid in describing the location, scale, and shape of the distribution of X, we define the moments

$$\mu \equiv E[X] \quad \text{and} \quad \mu_c \equiv E[(X - \mu)^c], \quad c = 2, 3, 4. \tag{8}$$

Thus μ is the mean and μ_2 is the variance of X. The skewness and kurtosis of X are

$$\sqrt{\beta_1} \equiv \mu_3/\mu_2^{3/2} \quad \text{and} \quad \beta_2 \equiv \mu_4/\mu_2^2 \tag{9}$$

respectively.

Figures 1 and 2 show some typical densities in the S_U and S_B families with the corresponding values of the shape parameters as well as the skewness and the kurtosis. To conserve space, we have omitted density plots for the more familiar normal (S_N) and lognormal (S_L) densities. Figures 1 and 2 actually show the density of the standardized variate $Y = (X - \xi)/\lambda$, which has location parameter zero and scale

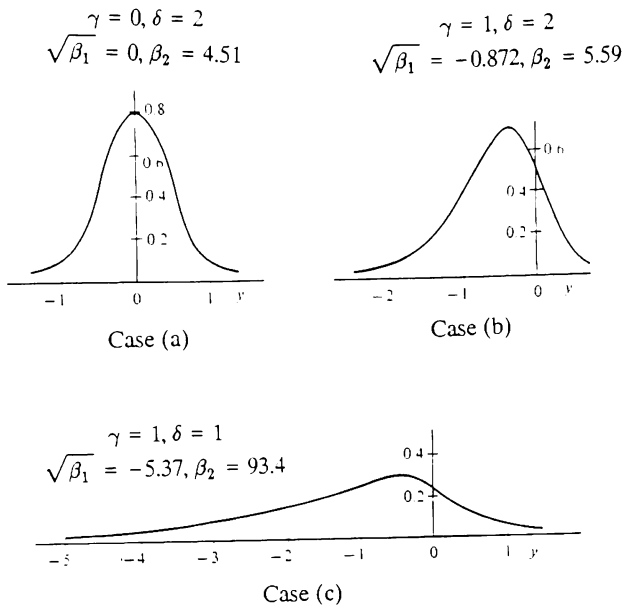

Figure 1: Some Typical Densities in the S_U Family

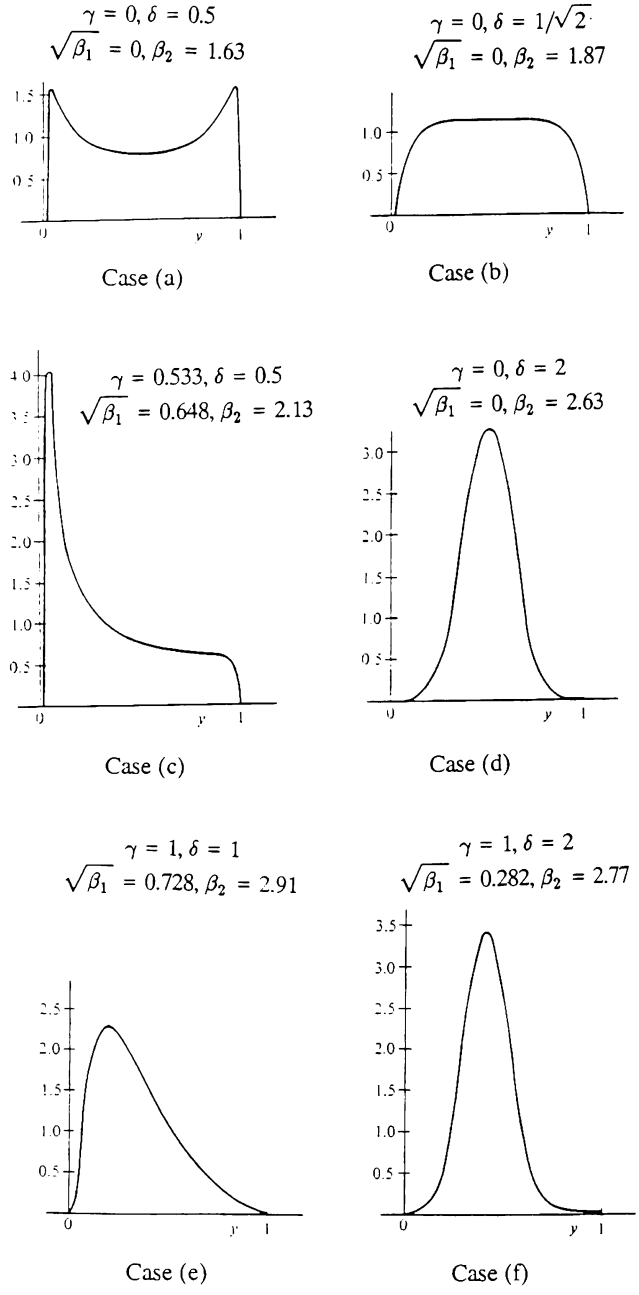

Figure 2: Some Typical Densities in the S_B Family

parameter one, in order to clearly reveal the role of the shape parameters γ and δ in determining the layout of a distribution in each of these families. Note that an S_U or S_B distribution is symmetric about its mean μ if and only if $\gamma = 0$. If γ is fixed, then as δ increases we observe that the corresponding S_U or S_B density becomes more sharply peaked. As suggested by Figure 1, every S_U distribution has a unique mode. Figure 2 demonstrates that S_B curves may be either unimodal or bimodal with an antimode between the two modes. Figure 2(c) is a deformation of Figure 2(a) in which the right-hand mode and the antimode have coalesced into a point of inflection. Figure 2b shows that the S_B family includes distributions that are nearly uniform over the interval $[\xi, \xi + \lambda]$. Finally we note that for both the S_U and S_B families, the density $p(x)$ and all of its derivatives tend to zero as x tends to extreme values in the support $\mathbf{H}$; this means that the density is a "perfectly smooth" (infinitely differentiable) function of x for all real values of x. Other well-known families of distributions do not possess this smoothness property.

3. SUBJECTIVE DISTRIBUTION FITTING

3.1. Johnson S_B Alternatives to Common Input Distributions

In developing a visual interactive approach to finding input distributions when little or no sample information is available, we confined ourselves to the S_B family because it matched well our notions of the general characteristics of many potential envisioned target distributions. The S_B distributions are bounded, thin-tailed, and capable of matching the skewness and kurtosis of many practical distributions. Real-world measurements are always bounded, even if only by the limits of technology, and typically extremal values near the end points of

a distribution are unlikely.

Figure 3 depicts the capability of the S_B family to mimic the properties of three distributions that are commonly used in large-scale discrete simulation studies – the triangular, normal and beta distributions. Although S_B densities will not closely fit all triangular distributions, they can serve as appealing alternatives. For example, the paired S_B and triangular densities shown in Figure 3(a) have the same minimum, maximum, and mode; and the standard deviation of each S_B distribution is taken as one-sixth of the range. Figure 3(b) shows that an S_B density can yield an excellent approximation to a normal density over a bounded interval. Figure 3(c) shows by setting the parameters $\{\gamma, \delta, \lambda, \xi\}$ of an S_B distribution to yield the same end points, mean, and variance as a unimodal beta distribution, we obtain a fitted S_B density which closely approximates the beta density.

Historically the S_B distribution has been difficult to work with because of the mathematically complex relationship of its shape to the parameters γ and δ. 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.

3.2. Subjective Specification of S_B Distributions

Describing an envisioned distribution in sufficient detail to permit its approximation by a parameterized functional form is a nontrivial task, even when restricting consideration to smooth, thin-tailed, unimodal densities such as the S_B 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.

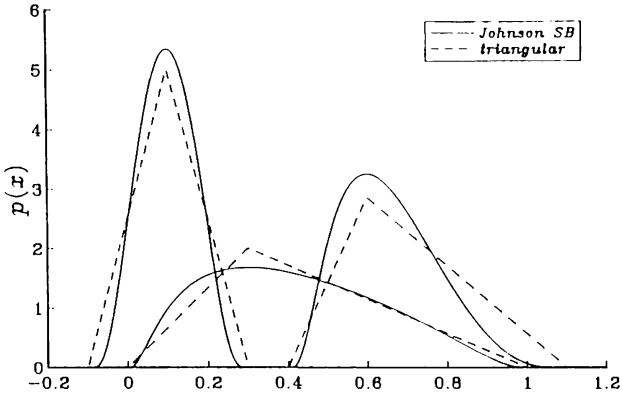

Figure 3(a): Matching Triangular and S_B Densities

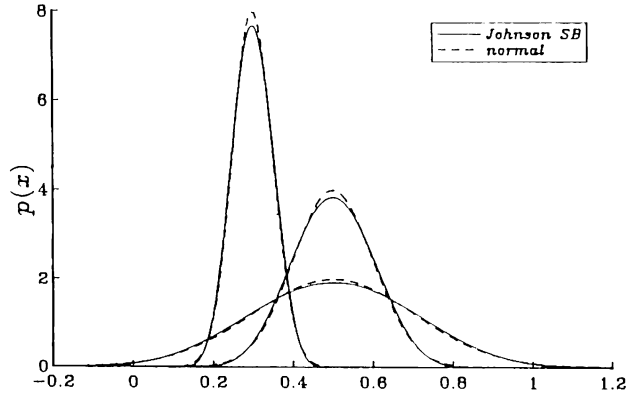

Figure 3(b): Matching Normal and S_B Densities

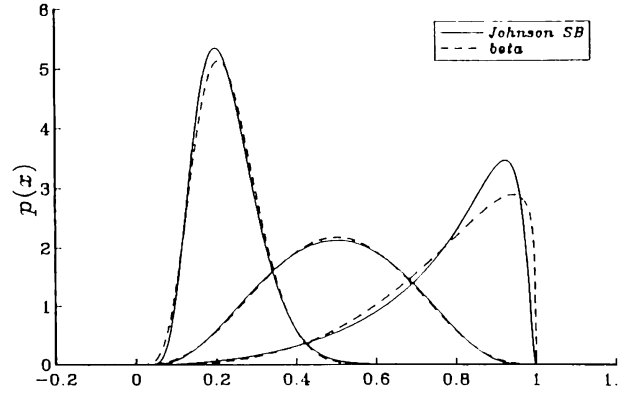

Figure 3(c): Matching Beta and S_B Densities

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.

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 of 0.45 and a mode of 0.1. 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. The VISIFIT software package is designed to permit visual display and interactive editing of the density shape.

4. 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. VISIFIT should run under most versions of MS-DOS or PC-DOS (we recommend version 3.0 or higher) on all IBM-compatible microcomputers. A numeric coprocessor is utilized if present, but it is not required. Because VISIFIT performs extensive floating-point computations, we strongly suggest running it on a fast AT-class microcomputer (that is, a machine running at or above 6 MHz) with a numeric coprocessor. VISIFIT requires at least 112K bytes of memory to execute. Currently, VISIFIT supports two types of video-display graphics: (a) color graphics for EGA- and VGA-compatible display adapters; and (b) Hercules monochrome graphics.

4.1. 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:

1. Mode
2. Mean
3. Median
4. Arbitrary percentile point(s)
5. Width of the central 95% of the distribution
6. 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 specified numerical characteristics (namely, two end points and two percentile points) must correspond to equidistant normal deviates. When the user gives no indication of the desired spread, VISIFIT suggests up to three choices for the standard deviation: (a) one-sixth of the range; (b) the standard deviation of the corresponding triangular distribution with the user's specified measure of centrality; and (c) the standard deviation that yields the closest fit to a normal distribution within the specified interval $[\xi, \xi + \lambda]$. VISIFIT also allows the user to specify the parameters of a beta distribution, to which it fits an S_B with the same end points, mean, and standard deviation.

By accepting a variety of different descriptions, VISIFIT minimizes the need for prior processing of information. The modeler is free to use whatever information is convenient, familiar, or easily understood. After the user has entered the desired numerical characteristics of the target population, VISIFIT computes the parameters for the S_B distribution that most closely matches those characteristics. Several numerical and approximative techniques are employed in this calculation, and all are detailed in DeBrota et al. (1989).

4.2. 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 4). 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:

1. Move the mode towards the upper bound
2. Move the mode towards the lower bound
3. Increase the width of the curve
4. Decrease the width of the curve
5. Move the 2.5th percentile point to the right
6. Move the 2.5th percentile point to the left
7. Move the 97.5th percentile point to the right
8. 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 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 user.

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.

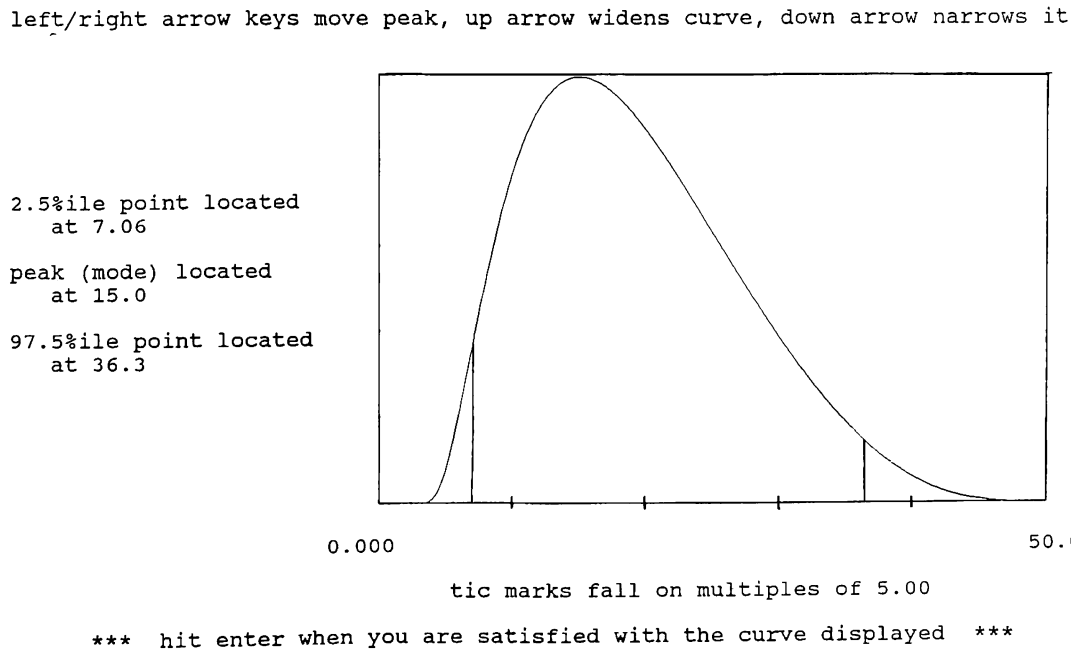

Figure 4: VISIFIT's Graphical Display

5. METHODS OF DISTRIBUTION FITTING WITH DATA

In contrast to the case involving little or no sample information, we now consider the problem of fitting Johnson distributions to sample data. We assume that the target population can be adequately described by a distribution in the Johnson system; and we want to use sample data from this population for two purposes: (a) to identify the Johnson distribution family that provides the most appropriate model for the population, and (b) to compute the parameter values for the selected family that yield the "best" fit. Thus in this section, our basic approach to simulation input modeling is to use sample data to determine the form of the input model as well as the parameter values for that model.

Fitting Johnson distributions to sample data generally requires the user to select a fitting method as well as a fitting criterion. In this section we discuss four basic distribution-fitting methods – moment matching, percentile matching, least squares, and minimum L_p norm estimation. For the last two fitting methods, we must also specify a fitting criterion that measures the "distance" between the sample distribution and the fitted Johnson distribution. It has been our experience that no single fitting method or fitting criterion is uniformly superior in all cases; different applications may require different statistical tools to yield appropriate input models.

5.1. Moment Matching

Suppose we have a random sample $\{x_j : j = 1, ..., n\}$ from the target distribution $F()$ that is to be approximated by a Johnson distribution. Then the sample analogs of equations (8) and (9) are

$$\hat{\mu} = n^{-1} \sum_{j=1}^{n} x_j; \quad \hat{\mu}_c = n^{-1} \sum_{j=1}^{n} (x_j - \hat{\mu})^c, \quad c = 2, 3, 4; \tag{10}$$

$$\widehat{\sqrt{\beta_1}} = \hat{\mu}_3/\hat{\mu}_2^{3/2}, \quad \text{and} \quad \hat{\beta}_2 = \hat{\mu}_4/\hat{\mu}_2^2. \tag{11}$$

For every possible pair of values of the sample skewness and sample kurtosis, there is exactly one Johnson distribution whose theoretical skewness and kurtosis match these values. The moment-matching technique for fitting a Johnson distribution to sample data uses the values of the sample statistics (11) to identify the appropriate distribution type from the four families defined by equations (5) through (7). As discussed just after equation (2), the number k of parameters to estimate depends on the selected distribution type; and the principle of moment matching prescribes that the first k sample moments should be equal to the corresponding theoretical moments of the fitted Johnson distribution. The resulting system of k nonlinear equations in k unknowns is then solved to obtain the parameter estimates for the fitted distribution.

5.2. Percentile Matching

Percentile matching involves estimating k parameters of a Johnson distribution by matching k selected percentiles of the standard normal distribution with the corresponding sample percentile estimators for the target population after those sample percentiles have been "normalized" via equation (1). For given percentages $\{\alpha_j : 1 \le j \le k\}$, the corresponding percentiles $\{z_{\alpha_j}\}$ and $\{x_{\alpha_j}\}$ of the standard normal distribution $\Phi()$ and the unknown target distribution $F()$ are respectively defined by

$$z_{\alpha_j} = \Phi^{-1}(\alpha_j) \quad \text{and} \quad x_{\alpha_j} = F^{-1}(\alpha_j), \quad j = 1, ..., k. \tag{12}$$

For example when $k = 2$, it is common to select the percentages $\{\alpha_1 = 0.25, \alpha_2 = 0.75\}$; and this implies that the lower and upper standard normal quartiles $z_{0.25} \cong -0.674$ and $z_{0.75} \cong 0.674$ are to be matched. Once the functional form $f()$ in equation (2) has been identified by some means, the method of percentile matching attempts to solve the k nonlinear equations

$$z_{\alpha_j} = \gamma + \delta \cdot f[(\hat{x}_{\alpha_j} - \xi)/\lambda], \quad j = 1, ..., k \tag{13}$$

in the k unknowns among the parameters $\{\gamma, \delta, \lambda, \xi\}$, where $\hat{x}_{\alpha_j}$ is a standard sample estimator of the percentile x_{α_j} of the target population.

5.3. Least Squares

Least squares estimation for the Johnson system involves minimization of the distance between a vector of "uniformized" order statistics and its corresponding expected value. Given the order statistics $x_{(1)} < \cdots < x_{(n)}$ obtained by sorting the random sample $\{x_i : 1 \le i \le n\}$ in ascending order, we can transform the ith order statistic $x_{(i)}$ into the uniformized order statistic

$$U_{(i)} = \Phi\left\{\gamma + \delta \cdot f\left[\frac{x_{(i)} - \xi}{\lambda}\right]\right\}, \quad i = 1, ..., n. \tag{14}$$

If the translation (1) yields a standard normal variate exactly, then $U_{(i)}$ has the distribution of the ith smallest observation in a sample of n random numbers from the uniform distribution on the unit interval (0, 1). In this case $U_{(i)}$ has expected value $\rho_i = E[U_{(i)}] = i/(n + 1)$. The "error" $\varepsilon_i = U_{(i)} - \rho_i$ represents the

random deviation between the observed and expected values of the ith uniformized order statistic so that

$$E[\varepsilon_i] = 0 \text{ and } \mathrm{Var}[\varepsilon_i] = \frac{i(n-i+1)}{(n+1)^2(n+2)}, \quad i = 1, \ldots, n. \tag{15}$$

Assigning the weight w_i to the error ε_i, we can formulate the least squares approach to parameter estimation for the Johnson system as follows:

$$\underset{\gamma, \delta, \lambda, \xi}{\text{minimize}} \sum_{i=1}^{n} w_i \cdot \varepsilon_i^2 \tag{16}$$

subject to

$$\left.\begin{array}{l} \delta > 0, \\ \lambda \begin{cases} > 0 & \text{for } S_U, \\ > x_{(n)} - \xi & \text{for } S_B, \\ = 1 & \text{for } S_L \text{ and } S_N, \end{cases} \\ \xi \begin{cases} < x_{(1)} & \text{for } S_L \text{ and } S_B, \\ = 0 & \text{for } S_N. \end{cases} \end{array}\right\} \tag{17}$$

When the weights $\{w_i\}$ in (16) are all equal to one, the objective function is equal to $|\boldsymbol{U} - \boldsymbol{\rho}|^2$, the squared length of the distance between the vector $\boldsymbol{U} = [U_{(1)}, \ldots, U_{(n)}]$ of uniformized order statistics and its expected value $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_n]$; and in this case the minimization of (16) yields the ordinary least squares (OLS) estimators for γ, δ, λ, and ξ.

Since the errors $\{\varepsilon_i\}$ in (16) do not have a constant variance, it is reasonable to take $w_i = 1/\mathrm{Var}[\varepsilon_i]$ for $i = 1, \ldots, n$. With this setup, we obtain the weighted least squares (WLS) estimators of the Johnson parameters. In a wide variety of applications, Swain, Venkatraman, and Wilson (1988) obtained WLS fits for Johnson distributions that were comparable and often superior to the fits obtained by the other methods described in this paper.

5.4. Minimum L_p Norm Estimation

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

$$\|F_n - \hat{F}\|_p \equiv \left[\int_{-\infty}^{\infty} |F_n(x) - \hat{F}(x)|^p \, d\hat{F}(x) \right]^{1/p}. \tag{18}$$

When we take $p = 1$ in (18), the L_1 norm $\|F_n - \hat{F}\|_1$ is the "area" between the plot of the empirical distribution $F_n(x)$ and the plot of the fitted distribution $\hat{F}(x)$ for all real x. The L_∞ norm is

$$\|F_n - \hat{F}\|_\infty = \max_{-\infty < x < \infty} |F_n(x) - \hat{F}(x)|, \tag{19}$$

the Kolmogorov-Smirnov goodness-of-fit statistic corresponding to the fitted distribution $\hat{F}$. In each case, the minimization of (18) is carried out subject to the constraints given in display (17). Since L_1 and L_∞ norm estimation directly seek to eliminate the gap between the empirical distribution and the fitted distribution, the resulting fits are very appealing visually.

6. USING THE FITTR1 SOFTWARE

All of the fitting methods described in Section 5 have been implemented for the Johnson translation system in the interactive software package FITTR1. In this section we briefly describe the operation of FITTR1. For a complete description of this software and the numerical methods on which it is based, see Venkatraman and Wilson (1987). As for VISIFIT, FITTR1 should run on all IBM-compatible microcomputers using versions of MS-DOS or PC-DOS numbered 3.0 or higher. FITTR1 will use a numeric coprocessor if it is present, and this is strongly recommended for fitting distributions to large data sets. The current version of FITTR1, which is configured to handle data sets of up to 500 observations, requires 212K bytes of memory to execute. Because all of the output of FITTR1 is plain text, virtually any video display can be used to run FITTR1.

The program begins 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 filename for the data set to be fitted. (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 standard descriptive statistics are calculated and displayed. FITTR1 also automatically computes and displays the results of fitting the data set by moment matching. Beyond this point in the interactive session, FITTR1 is command-driven. The available commands provide for fitting Johnson distributions, generating tables and plots of fitted and empirical distributions, and

manipulating input data files. Some of these commands are explained below.

`stat` **Command**: Displays the computed sample statistics – namely, the mean, standard deviation, skewness, and kurtosis as defined in (10) and (11).

`fit ijkm` **Command**: Fits a new distribution to the sample data set 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 that can be assigned to these variables.

`i`	0 = automatic distribution selection 1 = S_L distribution 2 = S_U distribution 3 = S_B distribution 4 = S_N distribution
`j`	0 = no end point known 1 = lower end point known 2 = upper end point known 3 = both end points known
`k`	0 = compute starting parameter values 1 = use previous parameter values
`m`	0 = moment matching 1 = percentile matching 2 = ordinary least squares estimation of the CDF 3 = weighted least squares estimation of the CDF 4 = L_1 estimation of the CDF (minimize sum of absolute errors) 5 = L_∞ 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 value of the pair $(\hat{\beta}_1, \hat{\beta}_2)$ and also performs parameter estimation by moment matching.

`par` **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 γ, δ, λ, and ξ.

`gof` **Command**: Computes chi-square and Kolmogorov-Smirnov goodness-of-fit statistics for the latest estimated Johnson distribution.

`cdf` **and** `pdf` **Commands**: Create files of fitted and empirical CDFs (respectively, PDFs) 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 ASCII (plain 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 straight lines to obtain the desired graph. Note that the `cdf` command can also be used to display tables of fitted versus empirical CDFs directly on the terminal.

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 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 pursuing methodology and software that effectively synthesizes VISIFIT and FITTR1 to provide a unified approach to input modeling.

ACKNOWLEDGMENTS

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

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. *Simulation* **52**, 199-205.

Doubilet, P., Begg, C. B., Weinstein, 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. (1988). Modeling multivariate populations with translation systems. Unpublished Ph.D. dissertation, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

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 presently a General Internal Medicine Fellow at the Indiana University School of Medicine. He received a B.S. in chemistry and physics from Butler University and an M.D. from Indiana University. He became a diplomate of the American Board of Internal Medicine in 1987, and he continues to practice both outpatient and inpatient medicine. His research interests are in decision support systems, medical decision making, and artificial 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 Associate Professor at the Indiana University School of Medicine. He received a B.S.I.E. from Purdue University, an M.P.H. from the University of North Carolina, and an M.D. from Indiana University. He is Director of the Fellowship Training Program in General Internal Medicine and the Clinical Practice Analysis Section of the Regenstrief Institute. He serves on the Editorial Board of *Medical Decision Making*. His current research interests are in medical decision making and clinical epidemiology.

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. 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 B.S.I.E., M.S.I.E., and Ph.D. degrees 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 in addition to making presentations and chairing sessions at conferences. Presently he

is a member of the Board of Directors of WSC '89, Chairman of SIGSIM, and Area Editor of *Simulation*.

Stephen D. Roberts
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.
(317) 494-5425
steverob@gb.ecn.purdue.edu

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 B.A. in liberal studies in 1974, a B.S. in engineering science in 1975, and an M.S. in mechanical engineering in 1977 from the University of Notre Dame. He received a Ph.D. 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
jswain%gtri01.bitnet

SEKHAR VENKATRAMAN is an Operations Research Analyst at Trans World Airlines, Inc. He received a B.S. in mechanical engineering from Annamalai University (India), an M.S.E. in operations research from The University of Texas at Austin, and a Ph.D. in industrial engineering from Purdue University.

Sekhar Venkatraman
Trans World Airlines, Inc.
110/3 S. Bedford Road
Mt. Kisco, NY 10549, U.S.A.
(914) 242-3150

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