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

Comparison of methods for fitting data using Johnson translation distributions

Robert H. Storer
Department of Industrial Engineering
Lehigh University
Bethlehem, PA 18015

Sekhar Venkatraman
Department of Industrial Engineering
Wichita State University
Wichita, KS 67208

James J. Swain
School of Industrial and Systems Engineering
Georgia Tech
Atlanta GA 30345

James R. Wilson
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907

ABSTRACT

The Johnson translation family of distributions provides a variety of distributional shapes for the modelling of empirical data that are readily used in simulation models. We compare a number of methods for estimating the parameters of these distributions, including moment matching (MM), least squares (ordinary- (OLS), weighted- (WLS) and diagonally weighted- (DWLS) least squares), and maximum likelihood (MLE). A sampling study is made to determine the properties of the fitted parameters and estimates based on the fitted parameters, such as the quantiles of the distribution. We restrict attention to the case that the analyst knows the correct distribution when fitting the parameters.

1. INTRODUCTION

Several approaches can be taken to the problem of choosing and fitting distributions for use as simulation input models. These approaches include parametric modelling, empirical distributions, and use of flexible distributional families, of which the Johnson family is one example. The parametric approach involves identification, from theory, experience, or sample data, the parametric model likely to have generated the observed data. Distribution fitting is usually by maximum likelihood, using the assumed parametric distribution as the basis for estimation. At the other extreme, one can avoid the choice of an explicit model by using the data to form empirical distributions. Variants of this approach, such as the empirical distribution with exponential tails suggested by Bratley, Fox, and Schrage (1987), require little in the way of additional assumptions about the underlying distribution being estimated. Finally, using slightly stronger assumptions (e.g., smooth density function and restriction to uni- or bivariate distributions) one can construct general families of distributions, including the Johnson translation family, Pearson, and Schmeiser-Deutsch (1980), among others. In contrast to the parametric approach, members of these families are considered useful approximations, and not necessarily the "true" distribution that generated the data.

Our attention is focused on the Johnson translation family, which consists of three distributions whose variates can be transformed into normal variates. For completeness the normal distribution is treated as a fourth member of the family. The general form of the transformations is

$$\gamma + \delta\, f[\,(x-\xi)/\lambda\,]$$

where $f(\cdot)$ denotes the transformation, λ and ξ are scale-location parameters, and γ and δ are shape parameters. The two parameters λ and δ are taken by convention to be positive. Table 1 lists the transformation functions $f(\cdot)$ and their inverses $f^{-1}(\cdot)$, which are useful for variate generation. Note that if Y is normally distributed with mean $-\gamma/\delta$ and variance $1/\delta$, the variate X can be obtained via

$$X = \xi + \lambda\, f^{-1}(\,Y\,).$$

The four distributions are the lognormal (S_L), the bounded (S_B), the unbounded (S_U), and the normal (S_N). Bounded variates are supported on the range

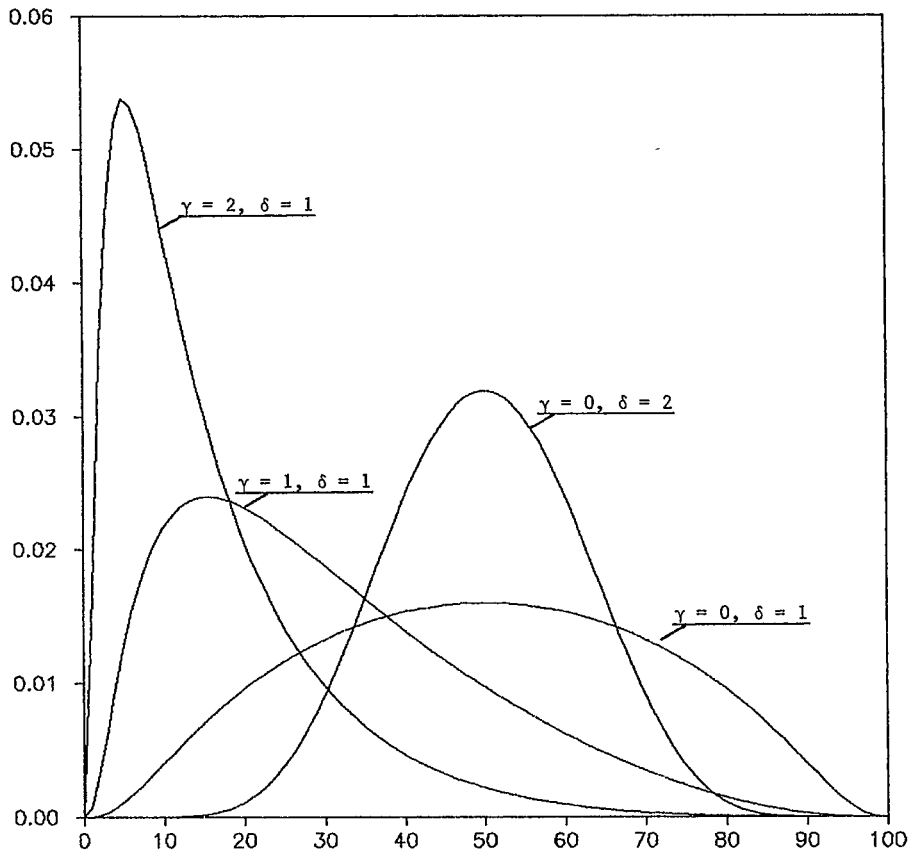

Figure 1. Four Examples of the Johnson S_B Family.
Parameters γ and δ are specified above; for all cases $\xi = 0$ and $\lambda = 100$.

$(\xi, \xi+\lambda)$. Several examples of the S_B and the S_U are illustrated in Figures 1 and 2.

TABLE 1: Transformations for Johnson distributions

Family	f(x)	$f^{-1}(x)$
S_L	e^x	$\log(x)$
S_U	$\log(x+(1+x^2)^{1/2})$	$\frac{(e^x - e^{-x})}{2}$
S_B	$\log(x/(1-x))$	$(1+e^{-x})^{-1}$
S_N	x	x

Figures 1 and 2 suggest the variety of distributional shapes that the Johnson S_U and S_B distributions can assume. It can also be shown that there is a unique member of the Johnson family for each permissible combination of third and fourth moments and any first and second moments. The moment matching method uses this fact to choose the parameters on the basis of the sample moments.

Moment matching and other parameter estimation methods are discussed in the next section. In section 3 a sampling experiment is conducted to examine the properties of the estimated distributions as a function of the fitting method. The section concludes with a brief discussion of the results and plans for further work.

2. METHODS OF FITTING

The moment matching algorithm consists of two parts: in the first part, the proper distribution is determined using the standardized third and fourth moments, β_1 and β_2. The actual parameters of the chosen distribution are obtained by equating the sample moments to the moments of the distribution taken as a function of the parameters, from which the parameters may be solved. Hill, Hill, and Holder (1976) provide a FORTRAN implementation of the algorithm.

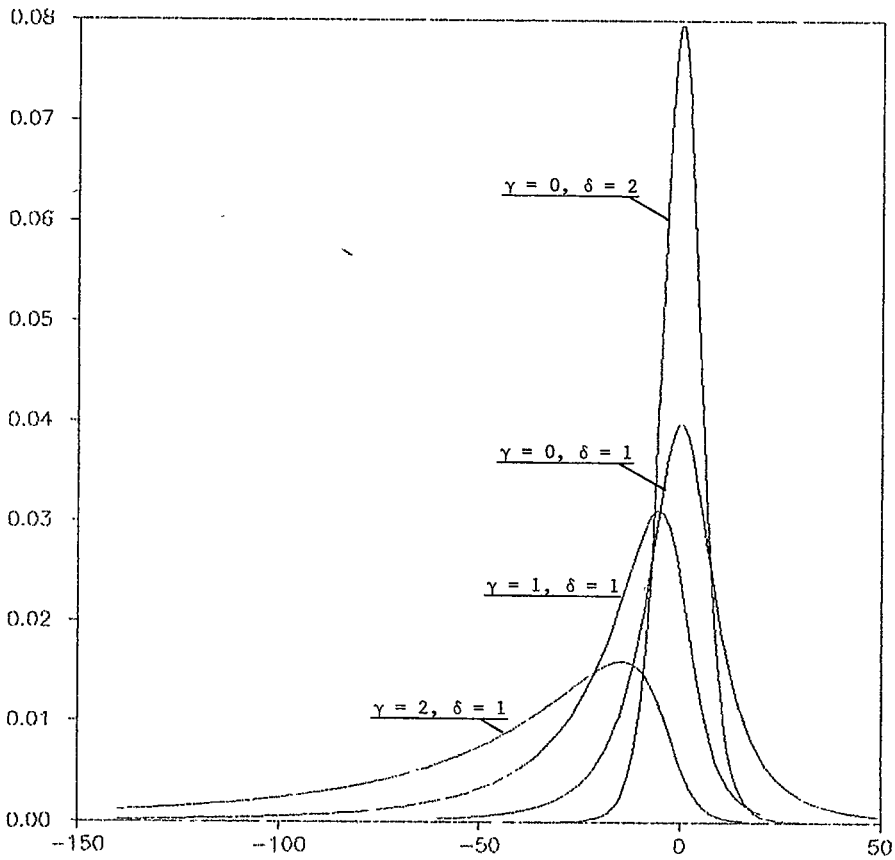

Figure 2. Four Examples of the Johnson S_U Family.
Parameters γ and δ are specified above; for all cases $\xi = 0$ and $\lambda = 10$.

More recently, in an effort to find a fitting method that could be easily automated and to avoid the feasibility problems sometimes encountered by MM, Wilson (1983) proposed a least squares criterion which matches uniformized order statistics to their expected values under the correct Johnson normalizing transformation. That is, letting $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ denote order statistics from a sample of size n, the variates

$$R_i(\underline{\psi},f) = \Phi\{ \psi_1 + \psi_2 f[(X_{(i)} - \psi_4)/\psi_3]\}$$

for $\underline{\psi} = (\gamma,\delta,\lambda,\xi)^T$, will have the same distribution as the order statistics from the uniform distribution under the correct choice of the transformation $f(\cdot)$ and the parameters $\underline{\psi}$. Let $\rho_i = i/(n+1) = E[U_{(i)}]$ denote the expected value of uniform order statistics, $U_{(i)}$. The least squares algorithm fits parameters by minimizing the squared distance between the $R_i(\underline{\psi},f)$ and the ρ_i. Further details are provided in Swain and Wilson (1985), and Swain, Venkatraman, and Wilson (1988). Venkatraman and Wilson (1987) provide a FORTRAN program FITTR1 which performs the fitting for a variety of estimators including least squares.

The errors $\varepsilon_i = R_i - \rho_i$, i=1,...,n are dependent and nonidentically Beta distributed. For most i, and moderate samples sizes (n>30), the Beta distributions will be fairly normal and the correlations could be neglected, so that ordinary least squares (OLS) is suitable for estimation. A weighting scheme based upon the known variance and covariance function of the uniform order statistics leads to a weighted least squares (WLS) estimator. The WLS estimator is often biased in practice. An illustration of this bias and an explanation is provided by Swain, Venkatraman, and Wilson (1988). Better results can be obtained by basing the weights exclusively on the inverse of the marginal variances, for a diagonally weighted least squares (DWLS) estimator.

The distribution functions of the members of the Johnson family of distributions are known, so maximum likelihood estimation can be used to estimate parameters. Storer (1987) examines the properties of maximum likelihood estimators (MLE's) and develops algorithms for their solution. Solution of the likelihood equations leads to the elimination of the parameters δ and γ as functions of the two remaining parameters, λ and ξ. That is, define the transformed random variates $A_i = f[(X_i-\xi)/\lambda]$. Let V_A be the MLE for the variance of A, and let $\bar{A}$ be its sample mean. Then the MLE for δ is given by $\hat{\delta}^{-2} = V_A/n$, and the MLE for γ is $\hat{\gamma} = -\bar{A}/\sqrt{V_A} = -\hat{\delta}\,\bar{A}$. Finally, defining $G_i = dA_i/dx$, the log-likelihood (less constants) can be compactly expressed

$$L = -\frac{n}{2}\log[\,V_A\,] + \sum_{i=1}^{n}\log[\,G_i\,].$$

While the log-likelihood is a function of only the remaining parameters λ and ξ, the surface is not simple. Storer (1987) examined the properties of the log-likelihood function and details a strategy for obtaining solutions.

3. EXPERIMENTAL RESULTS

A statistical sampling experiment is run to examine the properties of the parameters for each of the different fitting methods. To simplify the presentation of the results, we focus our attention primarily upon the estimation of the quantiles of the parent distributions. The quantiles provide an indication of how well the fitted distributions would reproduce variates from the parent distribution. In addition, since the behavior of the quantiles and distribution probabilities are related, the quantiles can be used to infer how well the distribution function can be estimated.

The comparisons considered here are limited to a single sample size and the initial restriction that the analyst "knows" the correct distribution for fitting. Sampling is performed for both a symmetric and skewed member of the S_B and S_U families. To simplify the comparison among estimators, the simulations employ common random numbers. Normal random variates are generated using the IMSL (1985) subroutine GGNPM, and the sampling was performed on Control Data Cyber 830 computers. Each sample consists of 200 replicates of n = 50 observations.

The two S_B cases are considered first. The first case is a symmetric S_B with parameters $\gamma = 0$, $\delta = 1$, $\lambda = 100$, and $\xi = 0$, while the second S_B case has $\gamma = 1$; both distributions are illustrated in Figure 1. Results for the quantile estimates for $q_{.1}$, $q_{.25}$, $q_{.5}$, $q_{.75}$, and $q_{.9}$ are provided in Tables 2 and 3. Note that only partial samples are included for the MM estimators: these are the cases for which the parameter estimates are consistent with the data observed. All five estimators perform well for the S_B's, though the MM and LS estimators do not do quite as well for the extreme quantiles, $q_{.1}$ and $q_{.9}$, particularly in the skewed case.

Table 2: Sample Means and Variances of Estimated Quantiles for S_B, Case 1. (Variances in (·))

Est.	Quantile				
	$q_{.10}$ 21.7	$q_{.25}$ 33.7	$q_{.50}$ 50.0	$q_{.75}$ 66.3	$q_{.90}$ 78.3
MM [n=81]	22.0 (8.88)	33.1 (9.12)	49.5 (9.88)	66.2 (9.53)	78.0 (9.04)
OLS	22.8 (15.7)	33.4 (17.2)	50.1 (15.2)	67.0 (14.6)	78.1 (14.3)
WLS	22.1 (24.6)	33.0 (28.7)	50.8 (26.2)	68.6 (25.7)	79.5 (22.3)
DWLS	22.4 (14.5)	33.1 (16.4)	50.0 (14.8)	67.2 (13.7)	78.4 (13.3)
MLE	21.5 (13.8)	32.9 (17.5)	50.1 (15.2)	67.2 (14.1)	78.6 (12.2)

Table 3: Sample Means and Variances of Estimated Quantiles for S_B, Case 2. (Variances in (·))

Est.	Quantile				
	$q_{.10}$ 3.62	$q_{.25}$ 6.45	$q_{.50}$ 11.9	$q_{.75}$ 21.0	$q_{.90}$ 32.8
MM [n=52]	3.91 (.764)	6.64 (1.37)	12.3 (2.67)	21.7 (5.44)	33.0 (12.3)
OLS	3.84 (.730)	6.39 (1.32)	12.1 (2.81)	21.8 (8.21)	32.5 (21.0)
WLS	3.78 (1.16)	6.34 (2.18)	12.3 (4.84)	22.6 (15.1)	33.7 (33.3)
DWLS	3.76 (.662)	6.31 (1.24)	12.1 (2.74)	22.0 (7.91)	32.9 (20.2)
MLE	3.58 (.552)	6.26 (1.23)	12.1 (2.65)	22.2 (7.20)	34.0 (18.7)

The two S_U distributions are also represented by a symmetric and a skewed distribution with common parameters $\delta = 1$, $\lambda = 10$, and $\xi = 0$. The first case is symmetric, $\gamma = 0$, and the second case is skewed, $\gamma = 1$. Both cases are illustrated in Figure 2. Here the case for the MLE appears to be stronger. The MM and LS quantile estimates were often biased, particularly for the second case, while the MLE's exhibited much less bias.

One of the hazards of using the MM estimator with the S_U is illustrated here. The MM estimator switches to an S_B fit when indicated by the sample moments. Though the S_U distributions used in this experiment are not close to the region for the S_B, variability in the sample moments is sufficient to require an S_B fit. In the second S_U case, for instance, 193 of the 200 replications were fit to an S_B instead of the S_U.

Table 4: Sample Means and Variances of Estimated Quantiles for S_U, Case 1. (Variances in (·))

Est.	Quantile				
	$q_{.10}$ -16.6	$q_{.25}$ -7.27	$q_{.50}$ 0.00	$q_{.75}$ 7.27	$q_{.90}$ 16.6
MM [n=151]	-18.5 (20.2	-9.24 (7.73)	.0185 (5.05)	9.404 (5.67)	19.0 (17.1)
OLS	-14.7 (13.0)	-7.73 (5.14)	-.0769 (3.21)	8.14 (4.61)	16.6 (14.9)
WLS	-16.8 (43.2)	-8.50 (13.6)	.274 (6.60)	9.66 (11.7)	19.6 (38.2)
DWLS	-15.2 (13.7)	-7.99 (5.24)	-.144 (3.44)	8.28 (4.83)	17.0 (14.4)
MLE	-16.5 (17.5)	-7.24 (5.47)	.0915 (3.00)	7.38 (4.62)	16.5 (13.9)

While the case for the MLE appears fairly strong, further sampling is necessary. In addition, since the MLE owes some of its superior performance to the knowledge of the parent distribution, it is of interest to compare the estimators when the parent distribution is either unknown or is a general distribution, not necessarily a member of the Johnson family.

Table 5: Sample Means and Variances of Estimated Quantiles for S_U, Case 2. (Variances in (·))

Est.	Quantile				
	$q_{.10}$ -48.4	$q_{.25}$ -25.7	$q_{.50}$ -11.8	$q_{.75}$ -3.31	$q_{.90}$ 2.85
MM [n=7]	-42.8 (130)	-27.1 (57.3)	-13.0 (19.6)	-0.724 (6.62)	10.3 (34.2)
OLS	-31.1 (34.6)	-23.8 (21.4)	-14.1 (9.34)	-2.18 (3.83)	10.8 (15.8)
WLS	-28.8 (108)	-22.1 (81.2)	-13.3 (49.1)	-2.69 (17.3)	9.08 (24.5)
DWLS	-30.1 (30.9)	-23.3 (19.5)	-14.5 (9.98)	-3.94 (4.35)	7.33 (7.71)
MLE	-48.0 (115)	-25.2 (24.3)	-11.6 (6.49)	-3.32 (3.35)	2.81 (4.59)

4. REFERENCES

Bratley, P., B.L. Fox, and L.E. Schrage (1987). *A Guide to Simulation*, Second Edition, Springer-Verlag, New York.

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

International Mathematical and Statistical Library (1985). *IMSL Library Reference Manual*, IMSL, Inc. Houston, Texas.

Schmeiser, B.W., and S.J. Deutsch (1980). "A Versatile four Parameter family of Probability Distributions, Suitable for Simulation," *AIIE Transactions 9*, pp. 176-182.

Storer, R.H. (1987). *Adaptive Estimation by Maximum Likelihood Fitting of Johnson Distributions*, unpublished Ph.D. thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

Swain, J.J., and J.R. Wilson (1985). "Fitting Johnson Distributions using Least Squares: Simulation Applications," *Proceedings of the 1985 Winter Simulation Conference*, (D.T. Gantz, S.L. Solomon, and G.C. Blais, eds.), Institute of Electrical and Electronic Engineers, Piscataway, New Jersey, pp. 150-157.

Swain, J.J., S. Venkatraman, and J.R. Wilson (1988). "Least-Squares Estimation of Distribution Functions in Johnson's Translation System," *Journal of Statistcal Computation and Simulation, 29*, pp. 271-297.

Venkatraman, S., and J.R. Wilson (1987). "Modeling Univariate Populations with Johnson's Translation System -- Description of the FITTR1 Software," Research Memorandum 87-21, School of Industrial Engineering, Purdue University.

Wilson, J.R. (1983). "Fitting Johnson Curves to Univariate and Multivariate Data," *Proceedings of the 1983 Winter Simulation Conference*, (S.D. Roberts, J. Banks, and B.W. Schmeiser, eds.) Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, p. 115.

ROBERT H. STORER is an Assistant Professor of Industrial Engineering at Lehigh University. He received a B.S. in Industrial and Operations Engineering from the University of Michigan (1979), and an M.S. in Operations Research (1982) and Ph.D. in Industrial and Systems Engineering (1987) from the Georgia Institute of Technology. He spent two years as a Research Engineer with General Dynamics (1980-81) working on various simulation projects. His current research interests include adaptive and robust estimation methods and transformation techniques in applied statistics, control aspects of quality control, and probabilistic search heuristics in combinatorial optimization. He is a member of ORSA, ASA, ASQC, and IIE.

Robert H. Storer
Dept. of Industrial Engineering
Mohler Lab 200
Lehigh University
Bethlehem PA 18015, U.S.A.
(215) 758-4436
rhs2%lehigh.bitnet

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 B.A. in Liberal Studies in 1974, a B.S. in Enginering Science in 1975, and an M.S. 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 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, SCS, and TIMS.

James J. Swain
School of ISYE
Georgia Tech
Atlanta, GA 30332-0205, U.S.A.
(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