# Introduction to Modeling and Generating Probabilistic Input Processes for Simulation

Michael E. Kuhl, RIT
Julie S. Ivy, NC State
Emily K. Lada, SAS Institute Inc.
Natalie M. Steiger, University of Maine
Mary Ann Wagner, SAIC
James R. Wilson, NC State

www.ise.ncsu.edu/jwilson

November 7, 2010

# OVERVIEW

# I. Introduction

- Stochastic simulations require valid input models—e.g., probability distributions that accurately mimic the random input processes driving the target system.

- Problems in using many conventional probability models:
  1. They cannot adequately represent real-world behavior, e.g. in the tails of the underlying distribution.
  2. Parameter estimation based on sample data or subjective information (expert opinion) is often troublesome.
  3. Fine-tuning the fitted model is difficult; e.g., many conventional probability distributions have the following drawbacks—
     (*a*) A limited number of parameters available to control the fitted distribution, and
     (*b*) No effective mechanism for directly manipulating the fitted distribution while simultaneously updating its parameter estimates.

- Conventional approach to identifying an input model uses sample data to select from a list of well-known alternatives based on
  1. informal graphical techniques such as probability plots, $Q$–$Q$ plots, histograms, empirical frequency distributions, or box-plots; and
  2. statistical goodness-of-fit tests such as the Kolmogorov-Smirnov, chi-squared, and Anderson-Darling tests.

- Drawbacks of conventional input modeling
  1. Visual comparison of a histogram to a fitted probability density function (p.d.f.) depends on the (arbitrary) layout of the histogram.
  2. Problems with statistical goodness-of-fit tests include:
     (*a*) In small samples, low power to detect lack of fit results in an inability to reject any alternatives.
     (*b*) In large samples, practically insignificant fit discrepancies result in rejection of all alternatives.

- Problems in estimating the parameters of the selected input model from sample data:
  - Matching the mean and standard deviation of the fitted distribution with that of the sample often fails to capture relevant shape characteristics.
  - Some estimation methods, such as maximum likelihood and percentile matching, may simply fail to estimate some parameters.
  - Users lack a comprehensive basis for selecting the “best-fitting” model.

- Problems with parameter estimation based on subjective information (expert opinion):
  - Subjective estimates of moments such as the mean and standard deviation can be unreliable and depend critically on the units of measurement.
  - Subjective estimates of extreme quantiles (e.g., lower and upper limits of the fitted distribution) are unreliable.
- Practitioners lack definitive procedures for identifying and estimating valid input models; thus, output analysis is often based on incorrect input processes.
- We focus on methods for input modeling that alleviate many of these problems.

# II. Univariate Input Models

Kuhl, M.E., J.S. Ivy, E.K. Lada, N.M. Steiger, M.A. Wagner, and J.R. Wilson. (2010) Univariate Input Models for Stochastic Simulation, *Journal of Simulation*, 4, 81-97.

## A. Generalized Beta Distribution Family

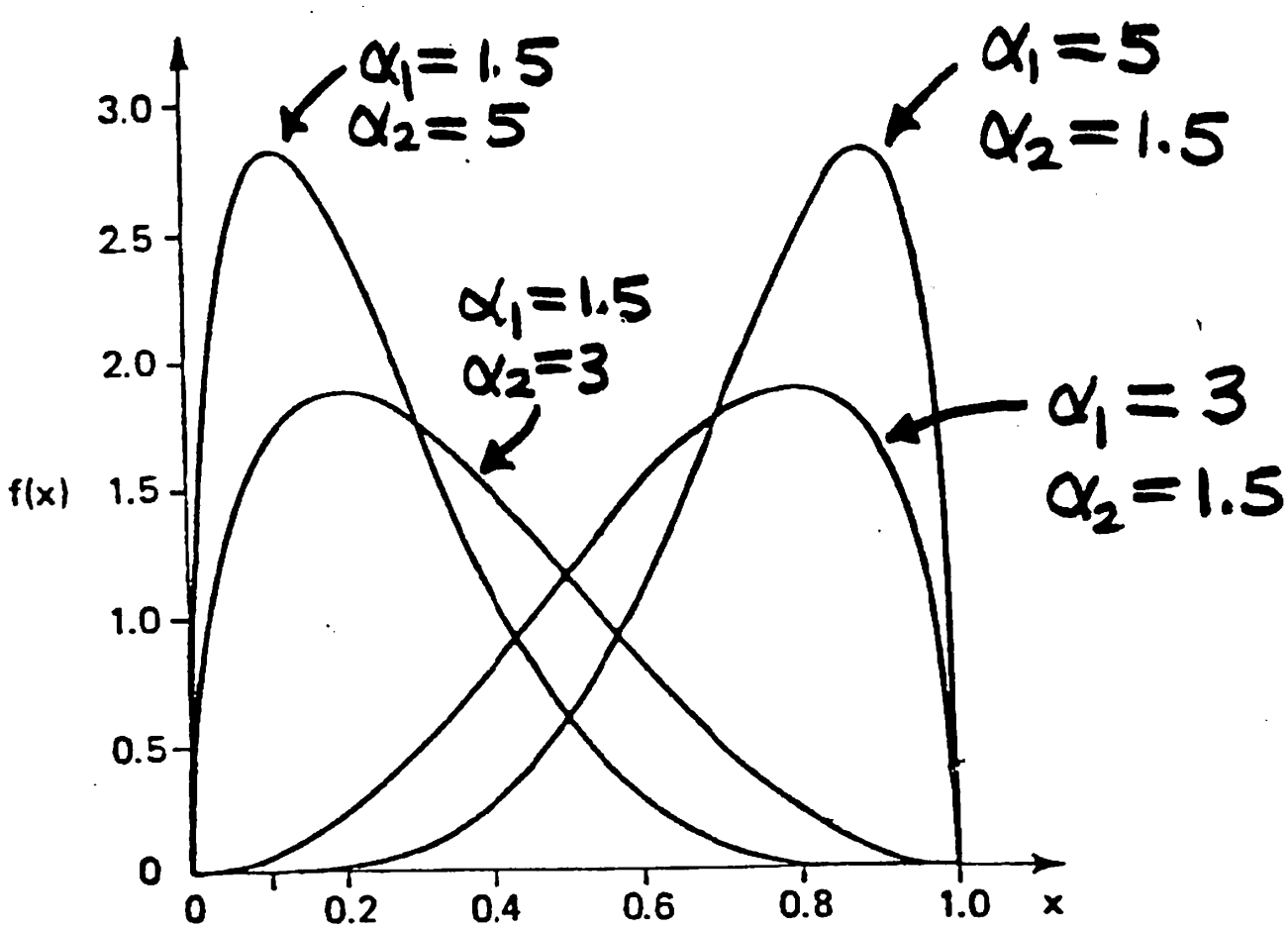
If $X$ is a continuous random variable with lower limit $a$ and upper limit $b$ whose distribution is to be approximated and subsequently sampled in a simulation, then often we can model the behavior of $X$ using a generalized beta distribution.

- Generalized beta p.d.f.

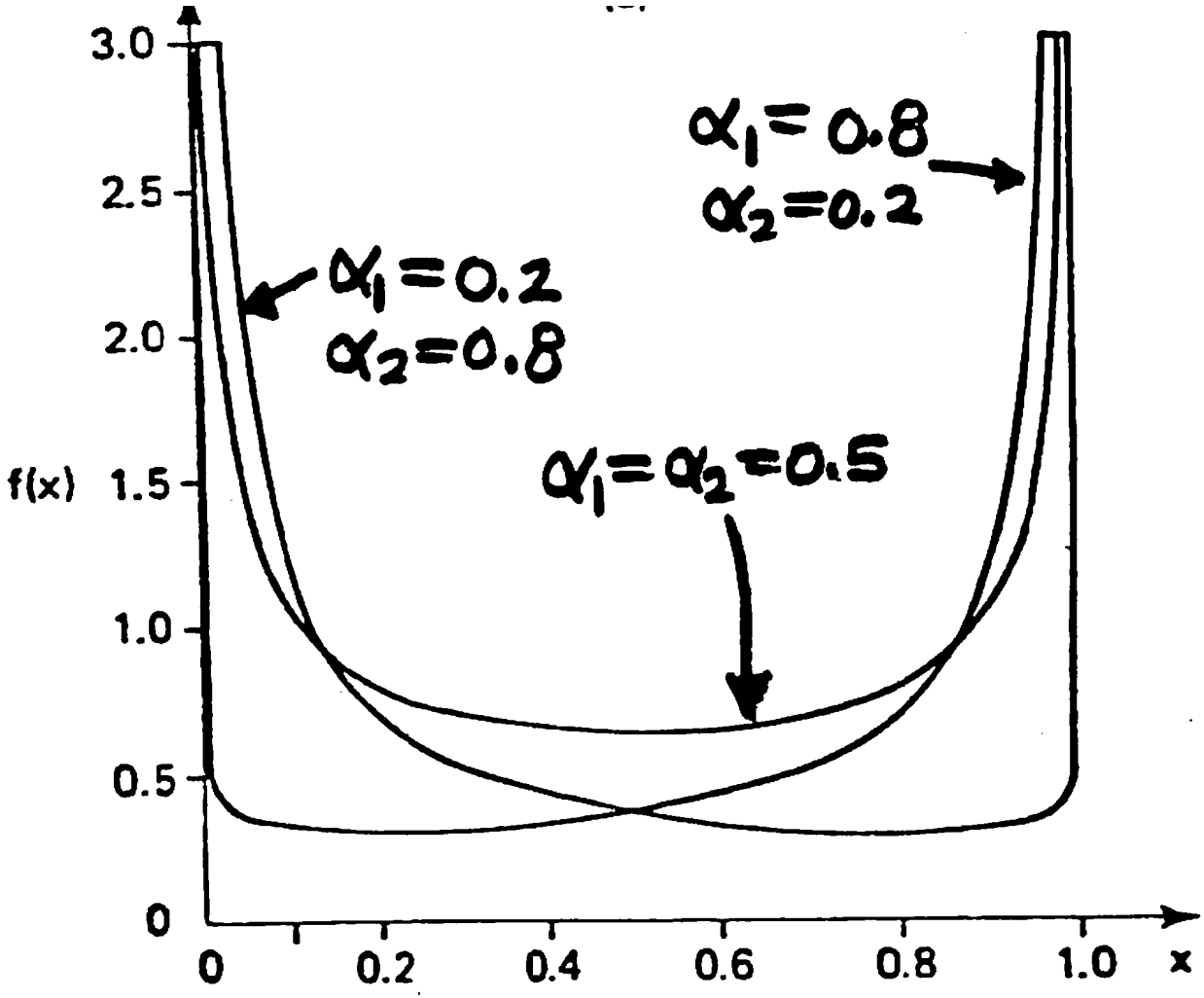
$$f_X(x) = \frac{\Gamma(\alpha_1 + \alpha_2)(x-a)^{\alpha_1 - 1}(b-x)^{\alpha_2 - 1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)(b-a)^{\alpha_1 + \alpha_2 - 1}} \quad \text{for} \quad a \le x \le b, \qquad (1)$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, \mathrm{d}t$ (for $z > 0$) denotes the gamma function.

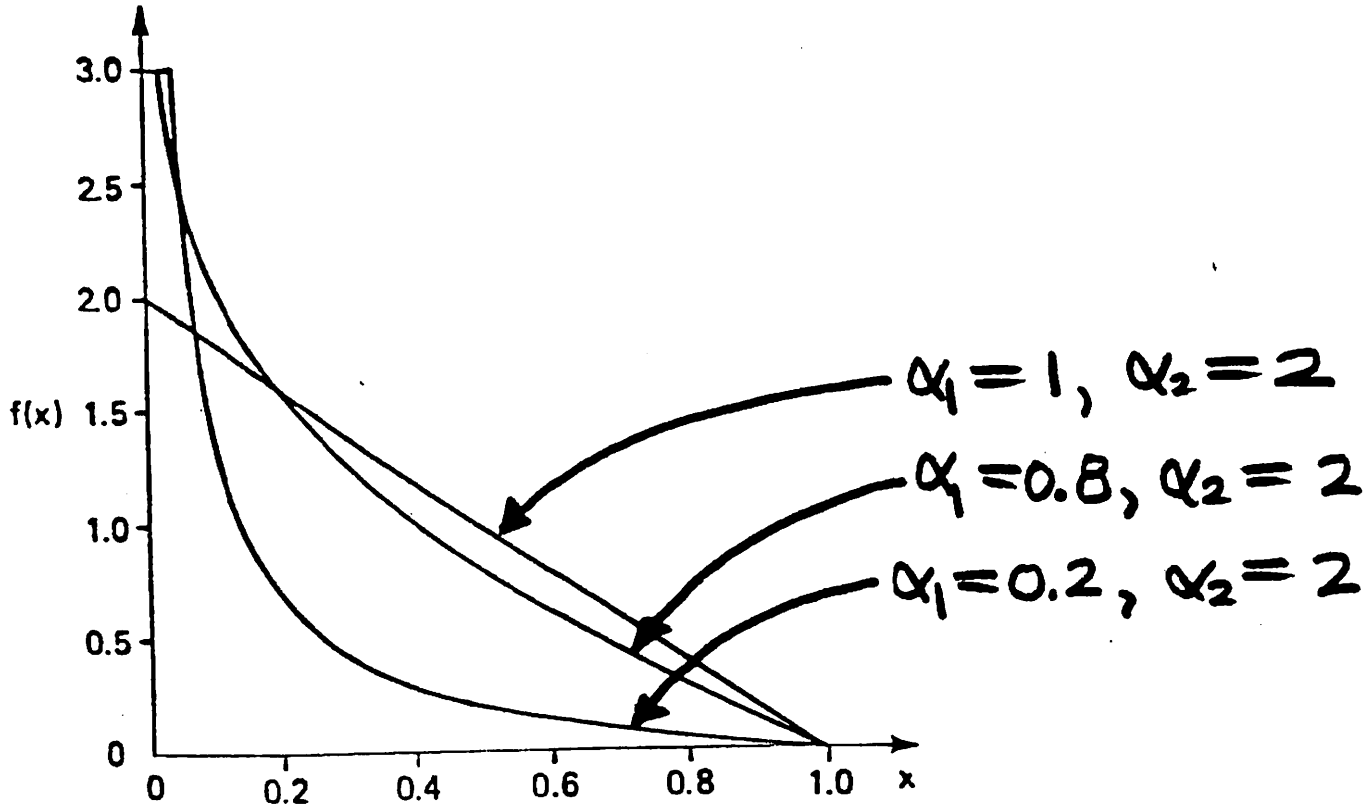
- The beta p.d.f. can accommodate a wide variety of shapes, including
  - symmetric and positively or negatively skewed unimodal p.d.f.'s;
  - $J$- and $U$-shaped p.d.f.'s;
  - left- and right-triangular p.d.f.'s; and
  - uniform p.d.f.'s.
- Some examples illustrating the range of distributional shapes achievable with the beta p.d.f. follow.



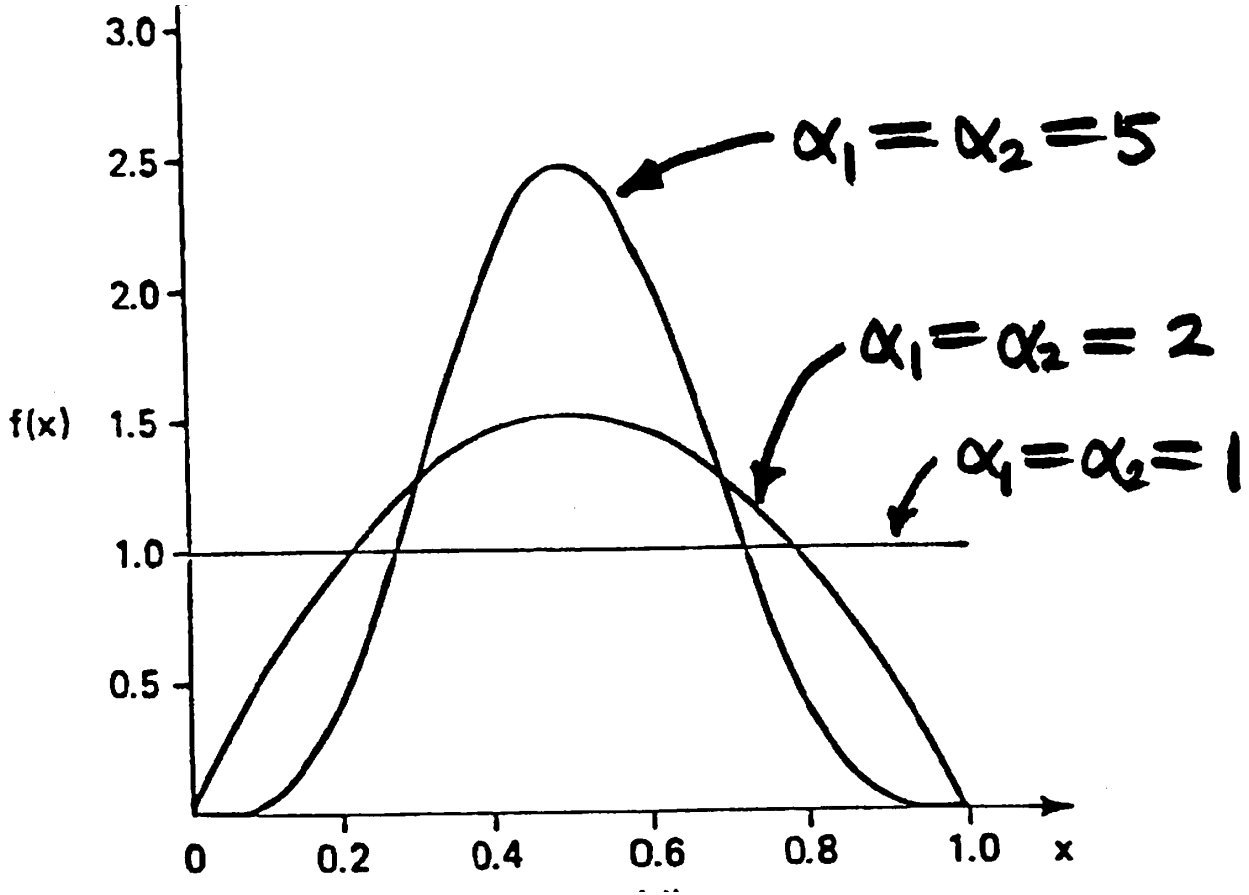

Positively and Negatively Skewed Unimodal Beta Densities




$U$-shaped Beta Densities




$J$-shaped and Left-triangular Beta Densities




Symmetric and Uniform Beta Densities

- Cumulative distribution function (c.d.f.) of beta variate $X$,

$$F_X(x) = \Pr\{X \leq x\} = \int_{-\infty}^{x} f_X(w)\,\mathrm{d}w \quad \text{for all real } x,$$

has no convenient analytical expression.

- Mean and variance of $X$ are given by

$$\left.\begin{aligned} \mu_X &= \mathrm{E}[X] = \frac{\alpha_1 b + \alpha_2 a}{\alpha_1 + \alpha_2}, \\ \sigma_X^2 &= \mathrm{E}\left[\left(X - \mu_X\right)^2\right] = \frac{(b-a)^2 \alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)}. \end{aligned}\right\} \tag{2}$$

- Provided $\alpha_1, \alpha_2 > 1$ so that the p.d.f. (1) is unimodal, the mode is given by

$$m = \frac{(\alpha_1 - 1)b + (\alpha_2 - 1)a}{\alpha_1 + \alpha_2 - 2}. \tag{3}$$

- The key distributional characteristics (2) and (3) are simple functions of $a$, $b$, $\alpha_1$, and $\alpha_2$; and this facilitates rapid input modeling.

## Fitting Beta Distributions to Data or Subjective Information

Given the data set $\{X_i : i = 1, \ldots, n\}$ of size $n$, we let

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

denote the order statistics; and we compute the sample statistics

$$\left.\begin{aligned} \widehat{a} = 2X_{(1)} - X_{(2)}, \quad \widehat{b} = 2X_{(n)} - X_{(n-1)}, \\ \bar{X} = n^{-1} \sum_{i=1}^{n} X_i, \quad S^2 = (n-1)^{-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2. \end{aligned}\right\}$$

- Moment-matching estimates of $\alpha_1, \alpha_2$ are computed from

$$\hat{\alpha}_1 = \frac{d_1^2(1-d_1)}{d_2^2} - d_1, \quad \hat{\alpha}_2 = \frac{d_1(1-d_1)^2}{d_2^2} - (1-d_1),$$

where

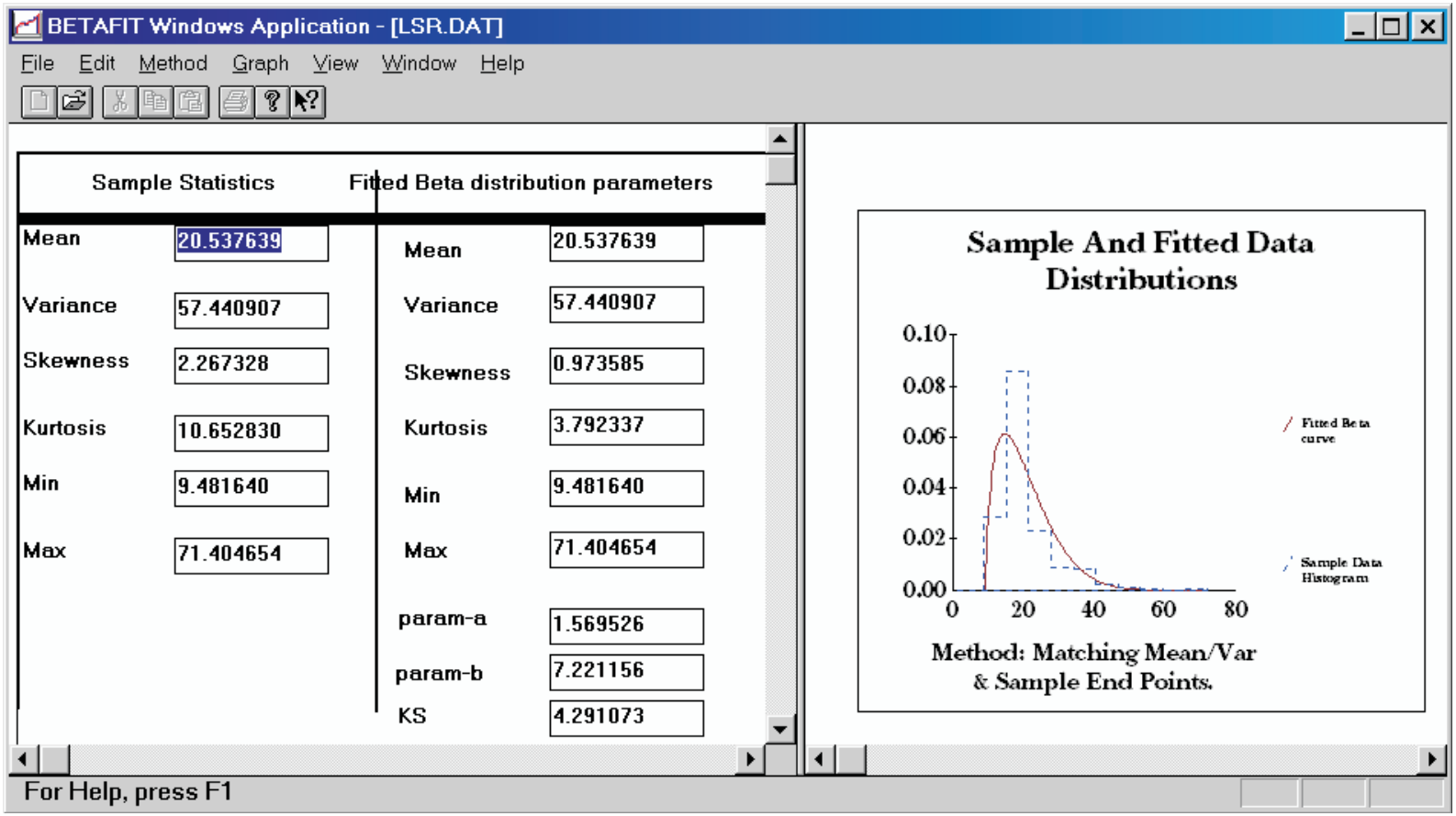
$$d_1 = \frac{\bar{X} - \hat{a}}{\hat{b} - \hat{a}} \quad \text{and} \quad d_2 = \frac{S}{\hat{b} - \hat{a}}.$$

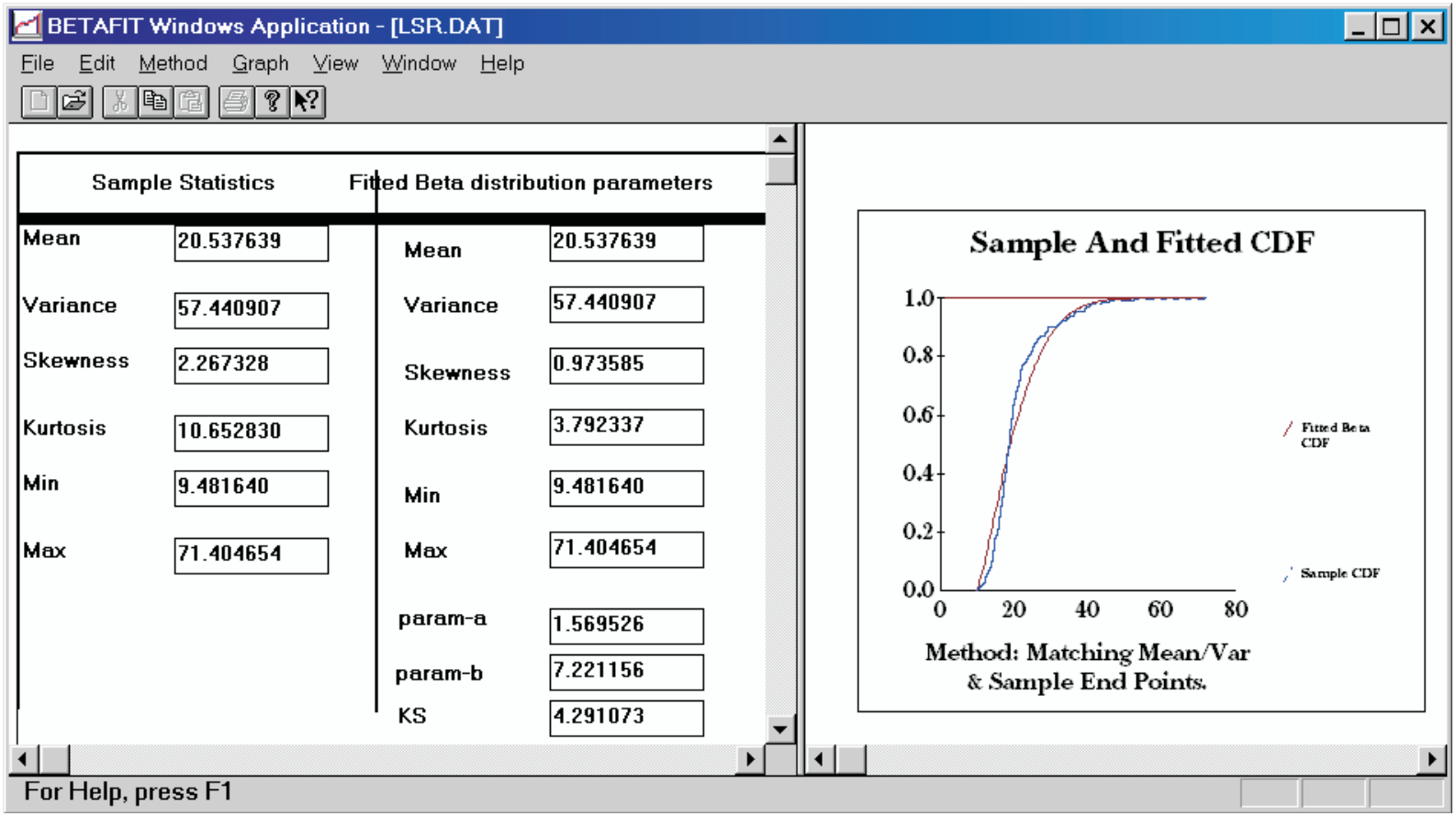
- BetaFit (AbouRizk, Halpin, and Wilson 1994) is a Windows-based package for fitting the beta distribution to sample data by computing $\hat{a}$, $\hat{b}$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$ using the following estimation methods:
  - moment matching;
  - feasibility-constrained moment matching (so that the feasibility conditions $\hat{a} < X_{(1)}$ and $X_{(n)} < \hat{b}$ are always satisfied);
  - maximum likelihood (assuming $a$ and $b$ are known and thus are not estimated); and
  - ordinary least squares (OLS) and diagonally weighted least squares (DWLS) estimation of the c.d.f.

- BetaFit is in the public domain and is available on the Web via

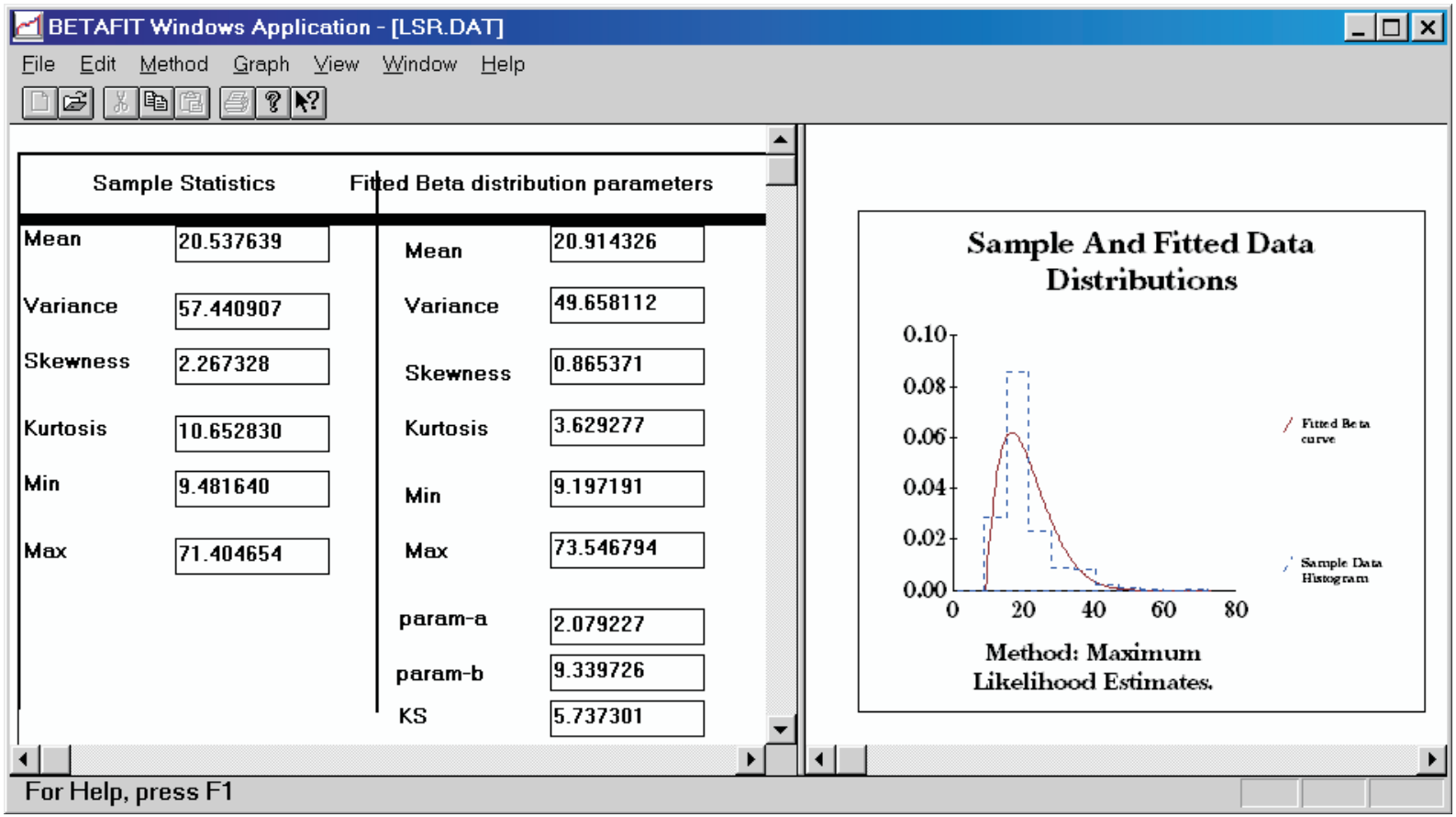
  <www.ise.ncsu.edu/jwilson/page3>.

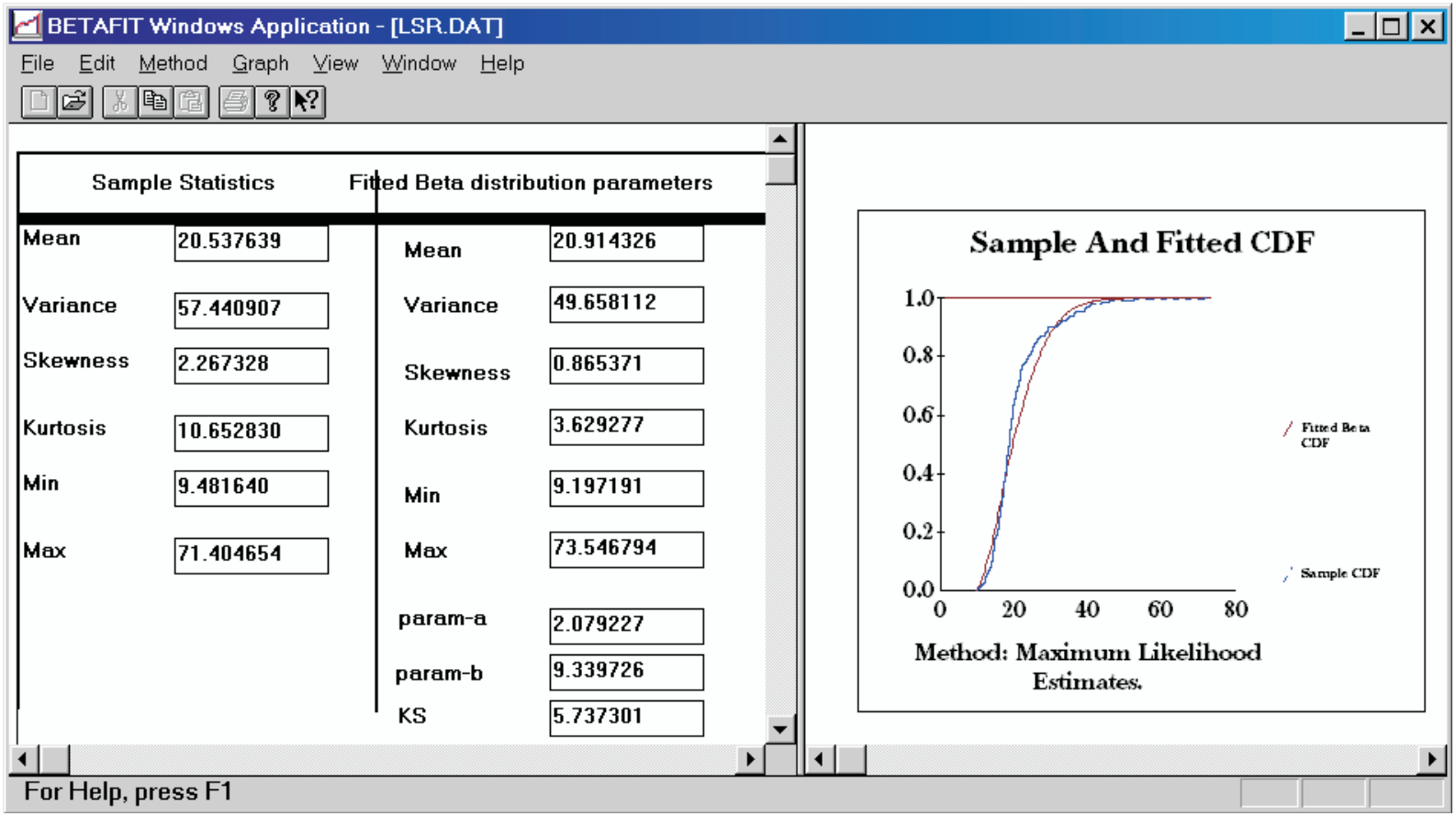
# Application of BetaFit to a Sample of $n = 9{,}980$ Observations of End-to-End Chain Lengths (in Angströms) of Nafion, an Ionic Polymer Used As a "Smart Material," Based on the Method of Moment Matching



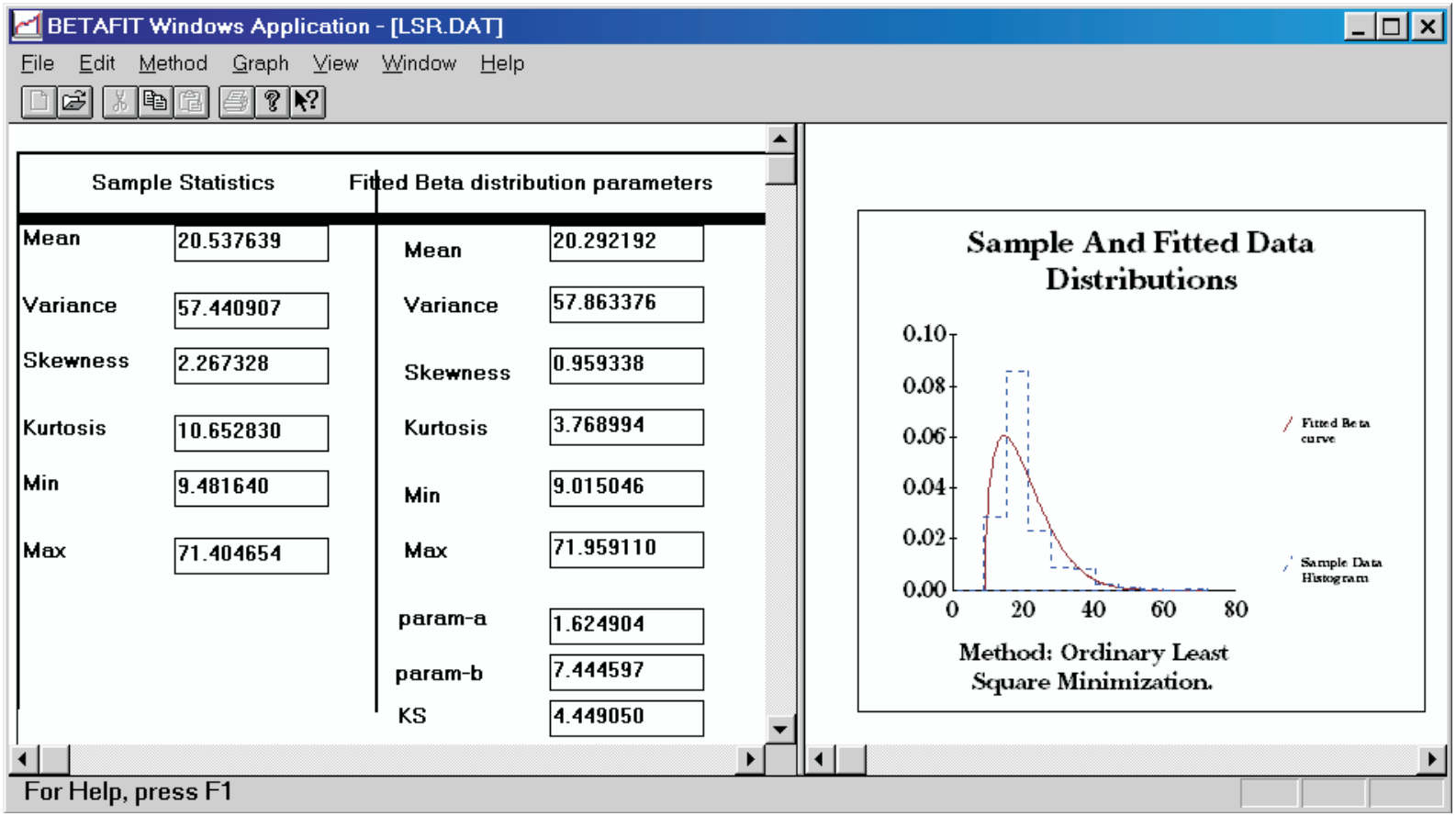

BETAFIT Windows Application - [LSR.DAT]
File Edit Method Graph View Window Help
Sample Statistics
Fitted Beta distribution parameters
Mean 20.537639
Variance 57.440907
Skewness 2.267328
Kurtosis 10.652830
Min 9.481640
Max 71.404654
Mean 20.537639
Variance 57.440907
Skewness 0.973585
Kurtosis 3.792337
Min 9.481640
Max 71.404654
param-a 1.569526
param-b 7.221156
KS 4.291073
Sample And Fitted CDF
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
Fitted Beta CDF
Sample CDF
Method: Matching Mean/Var & Sample End Points.
For Help, press F1

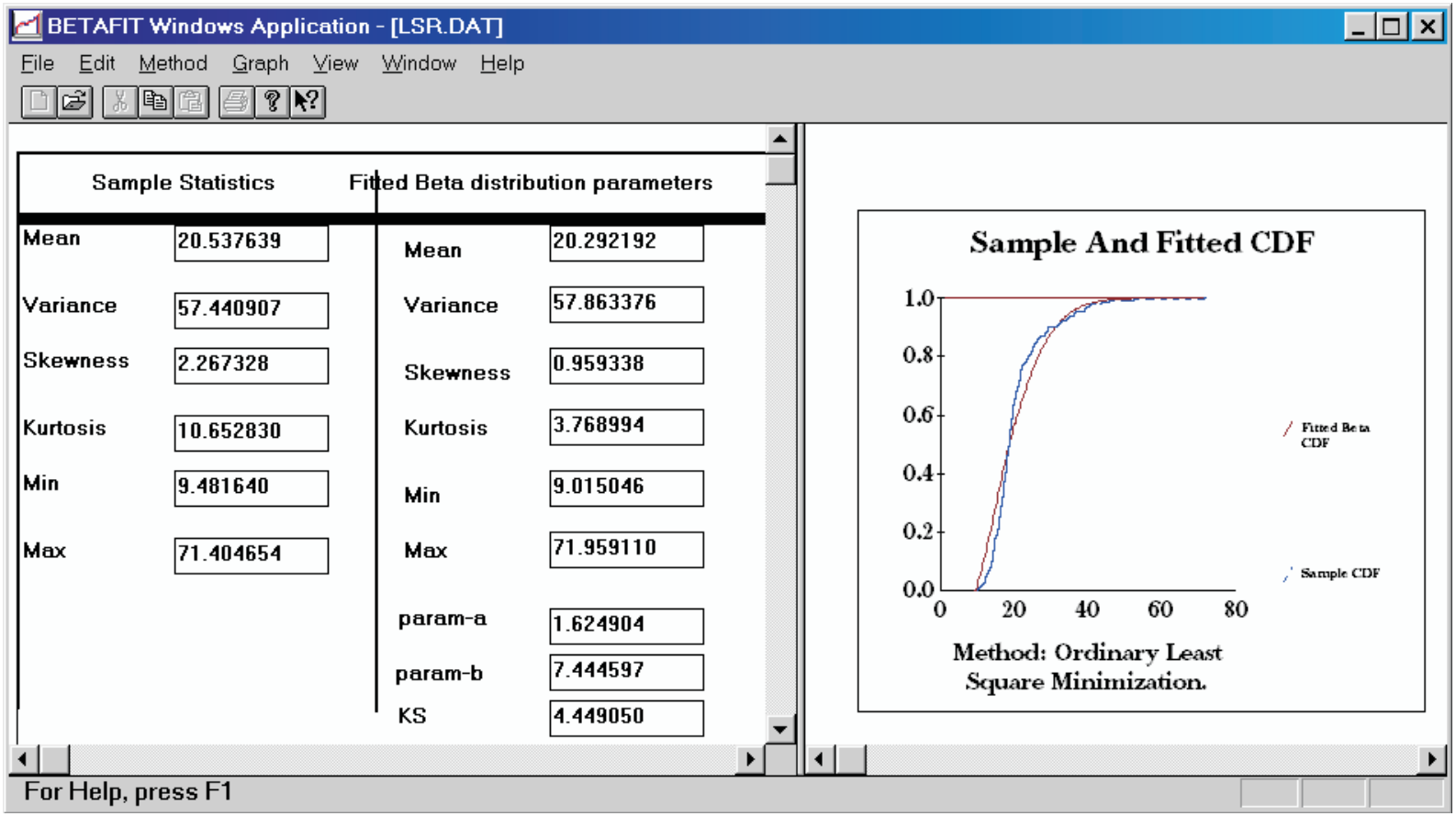
# Result of Applying BetaFit to Nafion Data Set Using Maximum Likelihood Estimation




BETAFIT Windows Application - [LSR.DAT]
File Edit Method Graph View Window Help
Sample Statistics
Fitted Beta distribution parameters
Mean 20.537639
Variance 57.440907
Skewness 2.267328
Kurtosis 10.652830
Min 9.481640
Max 71.404654
Mean 20.914326
Variance 49.658112
Skewness 0.865371
Kurtosis 3.629277
Min 9.197191
Max 73.546794
param-a 2.079227
param-b 9.339726
KS 5.737301
Sample And Fitted CDF
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
Fitted Beta CDF
Sample CDF
Method: Maximum Likelihood Estimates.
For Help, press F1

# Result of Applying BetaFit to Nafion Data Set Using Ordinary Least Squares Estimation of the C.d.f.




BETAFIT Windows Application - [LSR.DAT]
File Edit Method Graph View Window Help
Sample Statistics
Fitted Beta distribution parameters
Mean 20.537639
Variance 57.440907
Skewness 2.267328
Kurtosis 10.652830
Min 9.481640
Max 71.404654
Mean 20.292192
Variance 57.863376
Skewness 0.959338
Kurtosis 3.768994
Min 9.015046
Max 71.959110
param-a 1.624904
param-b 7.444597
KS 4.449050
Sample And Fitted CDF
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
Fitted Beta CDF
Sample CDF
Method: Ordinary Least Square Minimization.
For Help, press F1

- Rapid input modeling with subjective estimates $\hat{a}$, $\hat{m}$, and $\hat{b}$ of the minimum, mode, and maximum, respectively, of the target distribution:

$$\hat{\alpha}_1 = \frac{d^2 + 3d + 4}{d^2 + 1} \quad \text{and} \quad \hat{\alpha}_2 = \frac{4d^2 + 3d + 1}{d^2 + 1}, \tag{4}$$

where

$$d = \frac{\hat{b} - \hat{m}}{\hat{m} - \hat{a}}.$$

The mode of the fitted beta distribution will differ from $\hat{m}$ by at most 4.4%; in practice the error is usually at most 1%.

- VIBES (AbouRizk, Halpin, and Wilson 1991) is a Windows-based package for fitting the beta distribution to subjective estimates of:
  1. the endpoints $a$ and $b$; and
  2. any of the following combinations of distributional characteristics—
     - the mean $\mu_X$ and the variance $\sigma_X^2$,
     - the mean $\mu_X$ and the mode $m$,
     - the mode $m$ and the variance $\sigma_X^2$,
     - the mode $m$ and an arbitrary quantile $x_p = F_X^{-1}(p)$ for $p \in (0, 1)$, or
     - two quantiles $x_p$ and $x_q$ for $p, q \in (0, 1)$.

- Advantages of the beta distribution as an input-modeling tool:
  - sufficient flexibility to represent with reasonable accuracy a wide diversity of distributional shapes; and
  - convenient estimation of parameters from sample data or subjective information.
- Disadvantages of the beta distribution as an input-modeling tool:
  - difficult to explain; and
  - difficult to sample—some popular beta variate generators break down when $\alpha_1 > 30$ or $\alpha_2 > 30$.

## Generating Beta Variates

To generate a generalized beta variate $X$ with minimum $a$, maximum $b$, and shape parameters $\alpha_1$ and $\alpha_2$:

**[1]** Generate $Y(\alpha_1, \alpha_2)$, a standard beta variate with minimum 0, maximum 1, and shape parameters $\alpha_1$ and $\alpha_2$, using `Gammadev` of Press *et al* (2007); and

**[2]** Deliver

$$X = a + (b - a)Y(\alpha_1, \alpha_2)\,.$$

## Generating Beta Variates by Inversion

- The standard beta variate $Y(\alpha_1, \alpha_2)$ has c.d.f.

$$\Pr\{Y(\alpha_1, \alpha_2) \le x\} = I_x(\alpha_1, \alpha_2)$$
$$= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^x t^{\alpha_1 - 1}(1-t)^{\alpha_2 - 1}\, \mathrm{d}t \quad \text{for} \quad 0 \le x \le 1,$$

  which is the *incomplete beta function*.

- To generate the generalized beta variate $X$ with minimum $a$, maximum $b$, shape parameters $\alpha_1$ and $\alpha_2$, and c.d.f. $F_X(\cdot)$:

  **[1]** Generate a random number $U \sim \text{Uniform}[0, 1]$ ; and

  **[2]** Deliver
  $$X = F_X^{-1}(U) = a + (b-a) I_U^{-1}(\alpha_1, \alpha_2)\,,$$
  where $I_x^{-1}(\alpha_1, \alpha_2)$ is approximated by procedure `invbetai` of Press *et al* (2007).

## Application of Beta Distributions to Pharmaceutical Manufacturing

Pearlswig (1995) developed a simulation of a proposed facility for manufacturing effervescent tablets.

- For each operation, he obtained three time estimates $\left(\hat{a}, \hat{m}, \text{ and } \hat{b}\right)$ from the process engineers.
- Extremely conservative estimates given for upper limits (so that $\hat{b} \gg \hat{m}$).
- With triangular distributions to model processing times, bottlenecks resulted in excessively low simulation estimates of annual production.
- Using (4), Pearlswig fitted beta distributions to all operation times; and then the simulation results conformed to production levels of similar plants elsewhere.

## B. Johnson Translation System of Distributions

To fit a distribution to the continuous random variable $X$, Johnson (1949a) proposed finding a "translation" of $X$ to a standard normal random variable $Z$ with mean 0 and variance 1 so that $Z \sim N(0, 1)$.

For a detailed discussion of the Johnson translation system, see

DeBrota, D. J., R. S. Dittus, S. D. Roberts, J. R. Wilson, J. J. Swain, and S. Venkatraman. 1989a. Modeling input processes with Johnson distributions. In *Proceedings of the 1989 Winter Simulation Conference*, pp. 308–318. Available online via <www.ise.ncsu.edu/jwilson/files/wsc89jnsn.pdf>.

The proposed normalizing translations have the general form

$$Z = \gamma + \delta \cdot g\left(\frac{X - \xi}{\lambda}\right), \tag{5}$$

where $\gamma$ and $\delta$ are shape parameters, $\lambda$ is a scale parameter, $\xi$ is a location parameter, and the function $g(\cdot)$ defines the four distribution families in the Johnson translation system,

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

- Johnson c.d.f.

  If (5) is an exact normalizing translation of $X$ to a standard normal random variable, then the c.d.f. of $X$ is given by

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

  where: $\Phi(z) = (2\pi)^{-1/2} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}w^2\right) \mathrm{d}w$ is the standard normal c.d.f.; and the space of $X$ is

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

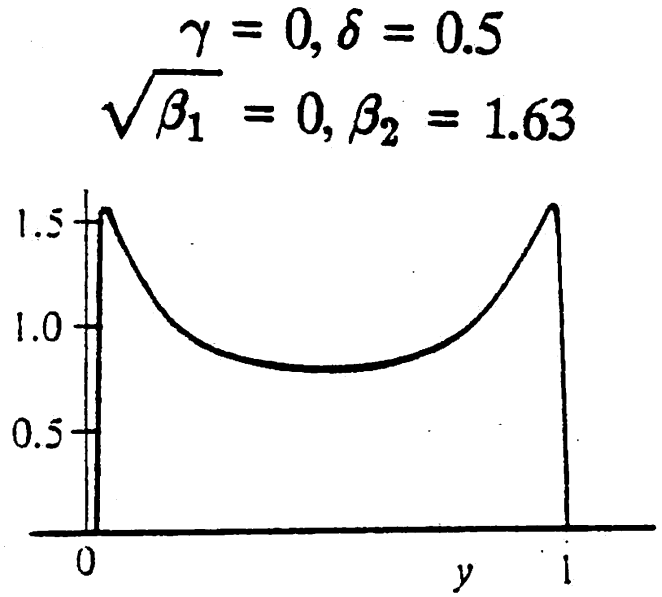
- Johnson p.d.f. is

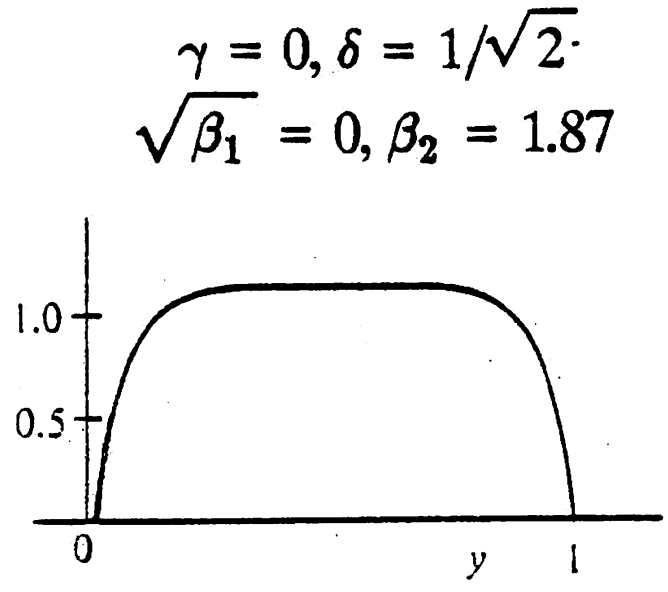
$$f_X(x) = \frac{\delta}{\lambda(2\pi)^{1/2}} g'\left(\frac{x-\xi}{\lambda}\right) \exp\left\{-\frac{1}{2}\left[\gamma + \delta \cdot g\left(\frac{x-\xi}{\lambda}\right)\right]^2\right\}$$

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

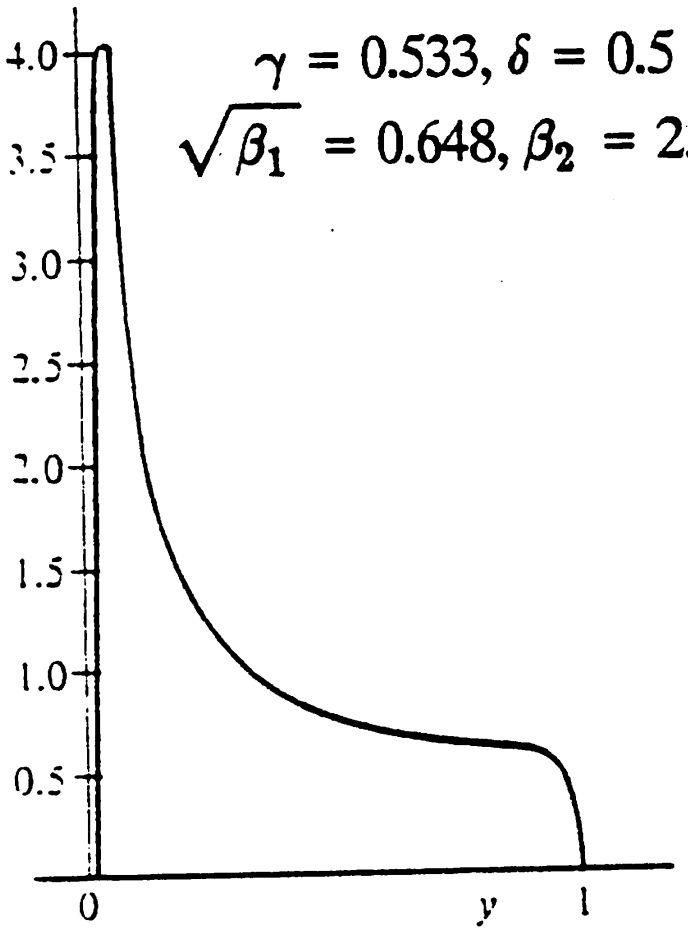
$$g'(y) = \begin{cases} 1/y, & \text{for } S_L \text{ (lognormal) family,} \\ 1/\sqrt{y^2+1}, & \text{for } S_U \text{ (unbounded) family,} \\ 1/[y/(1-y)], & \text{for } S_B \text{ (bounded) family,} \\ 1, & \text{for } S_N \text{ (normal) family.} \end{cases}$$

Following are examples illustrating all the distributional shapes in the Johnson system.

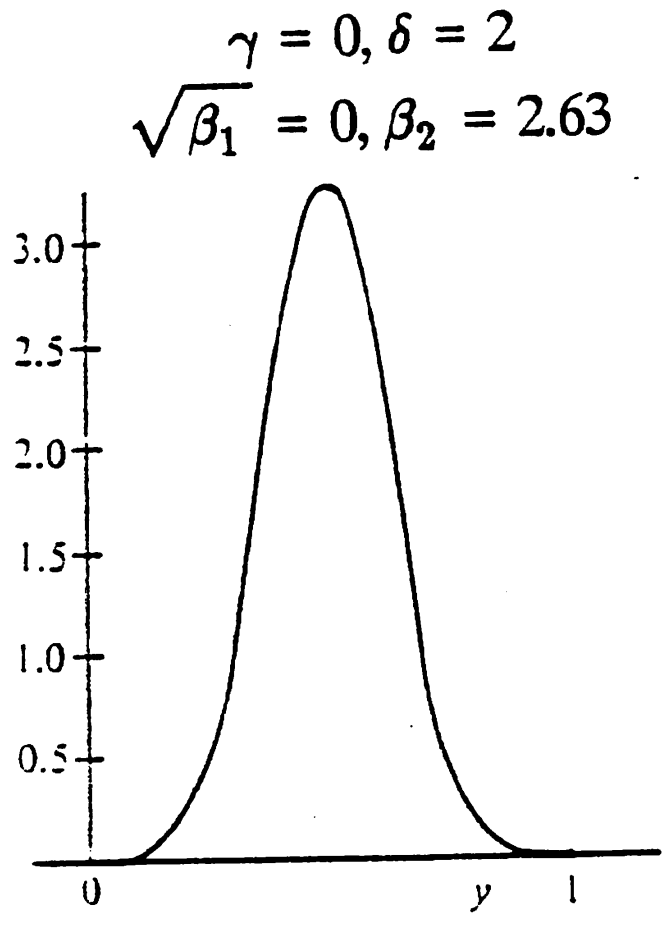



Symmetric Bimodal and Nearly Uniform Johnson $S_B$ Densities




Case (c)

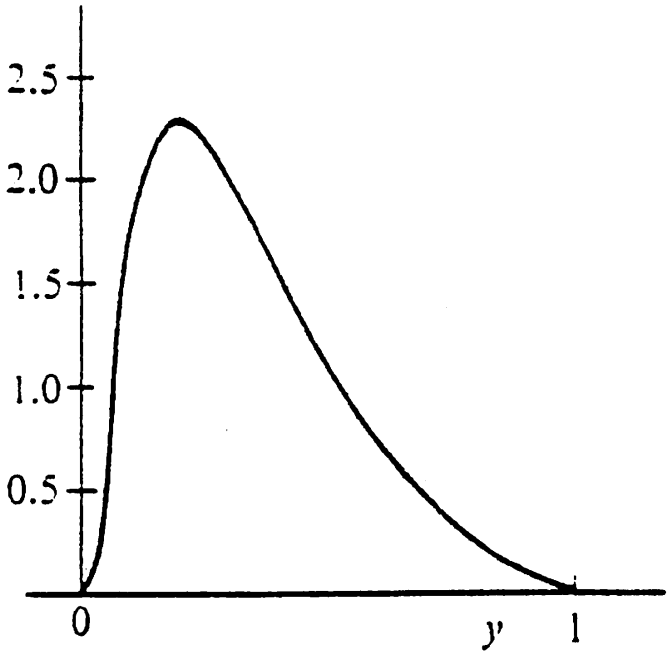



Case (d)

Nearly $J$-shaped and Symmetric Unimodal Johnson $S_B$ Densities

$\gamma = 1, \delta = 1$

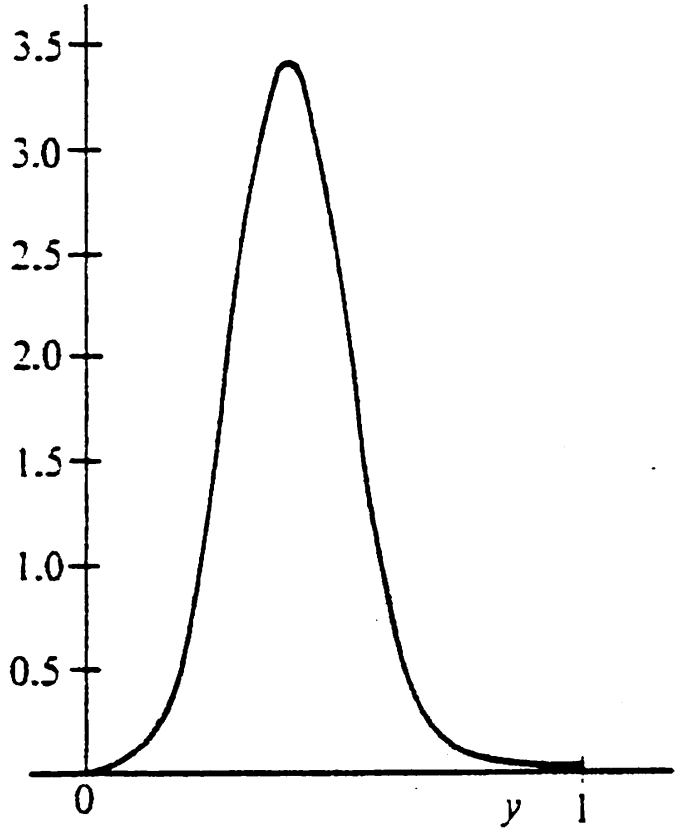
$\sqrt{\beta_1} = 0.728, \beta_2 = 2.91$




Case (e)

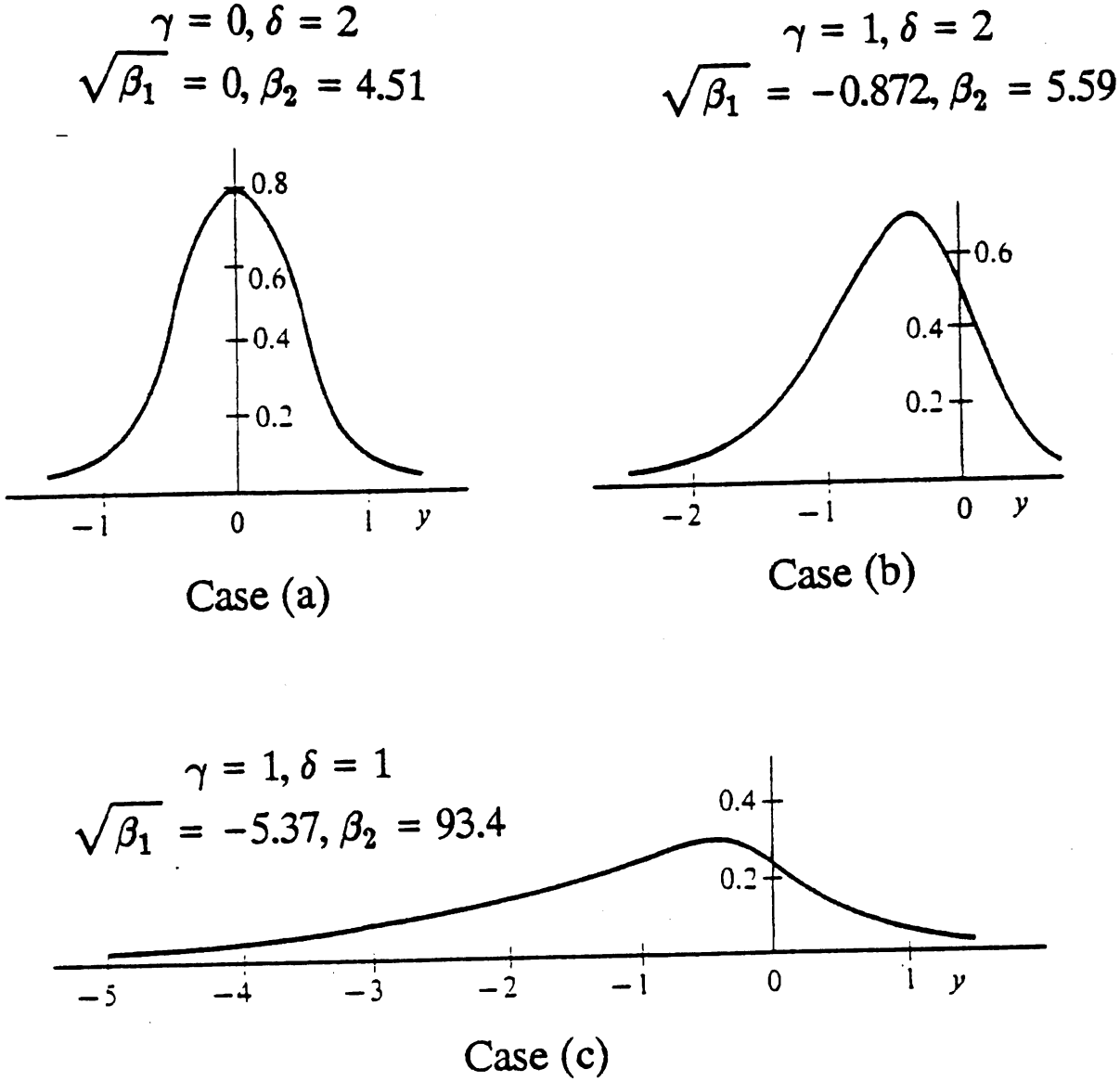
$\gamma = 1, \delta = 2$

$\sqrt{\beta_1} = 0.282, \beta_2 = 2.77$




Case (f)

Positively Skewed and Symmetric Unimodal Johnson $S_B$ Densities




Symmetric and Negatively Skewed Johnson $S_U$ Densities

## Fitting Johnson Distributions to Sample Data

We select an estimation method and the desired translation function $g(\cdot)$ and then obtain estimates of $\gamma$, $\delta$, $\lambda$, and $\xi$.

The Johnson system has the flexibility to match—

(*a*) any feasible combination of values for the mean $\mu_X$, variance $\sigma_X^2$, skewness

$$\text{Sk}_X = \text{E}\left[\left(X - \mu_X\right)^3 / \sigma_X^3\right] \quad \left(\text{often denoted by } \sqrt{\beta_1}\right),$$

and kurtosis

$$\text{Ku}_X = \text{E}\left[\left(X - \mu_X\right)^4 / \sigma_X^4\right] \quad (\text{often denoted by } \beta_2);$$

or

(*b*) sample estimates of the moments $\mu_X$, $\sigma_X^2$, $\text{Sk}_X$, and $\text{Ku}_X$.

- FITTR1 (Swain, Venkatraman, and Wilson 1988) is a software package for fitting Johnson distributions to sample data using the following estimation methods:
  - OLS and DWLS estimation of the c.d.f.;
  - minimum $L_1$ and $L_\infty$ norm estimation of the c.d.f.;
  - moment matching; and
  - percentile matching.

- VISIFIT (DeBrota et al. 1989b) is a Windows-based software package for fitting Johnson $S_B$ distributions to subjective information, possibly combined with sample data. The user must provide estimates of $a$, $b$, and any two of the following characteristics:
  - the mode $m$;
  - the mean $\mu_X$;
  - the median $x_{0.5}$;
  - arbitrary quantile(s) $x_p$ or $x_q$ for $p, q \in (0, 1)$;
  - the width of the central 95% of the distribution; or
  - the standard deviation $\sigma_X$.

  Venkatraman, Swain and Wilson (1988), DeBrota et al. (1989b), FITTR1, and VISIFIT are available on the Web via

  <www.ise.ncsu.edu/jwilson/more_info>.

## Generating Johnson Variates by Inversion

**[1]** Generate $Z \sim N(0,1)$.

**[2]** Apply to $Z$ the inverse translation

$$X = \xi + \lambda \cdot g^{-1}\left(\frac{Z-\gamma}{\delta}\right), \tag{6}$$

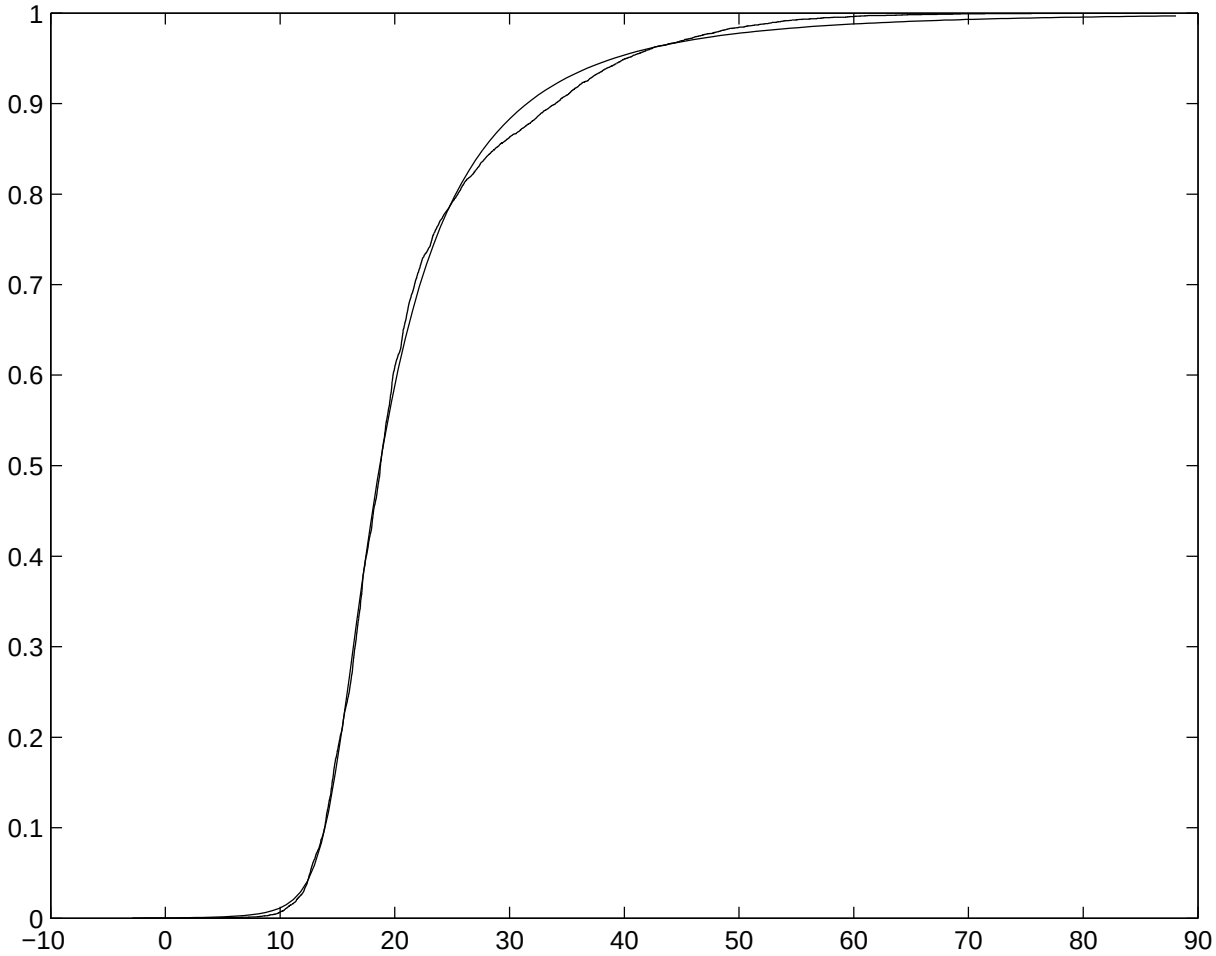
where for all real $z$ we define the inverse translation function

$$g^{-1}(z) = \begin{cases} e^z, & \text{for } S_L \text{ (lognormal) family,} \\ \left(e^z - e^{-z}\right)/2, & \text{for } S_U \text{ (unbounded) family,} \\ 1/\left(1+e^{-z}\right), & \text{for } S_B \text{ (bounded) family,} \\ z, & \text{for } S_N \text{ (normal) family.} \end{cases} \tag{7}$$

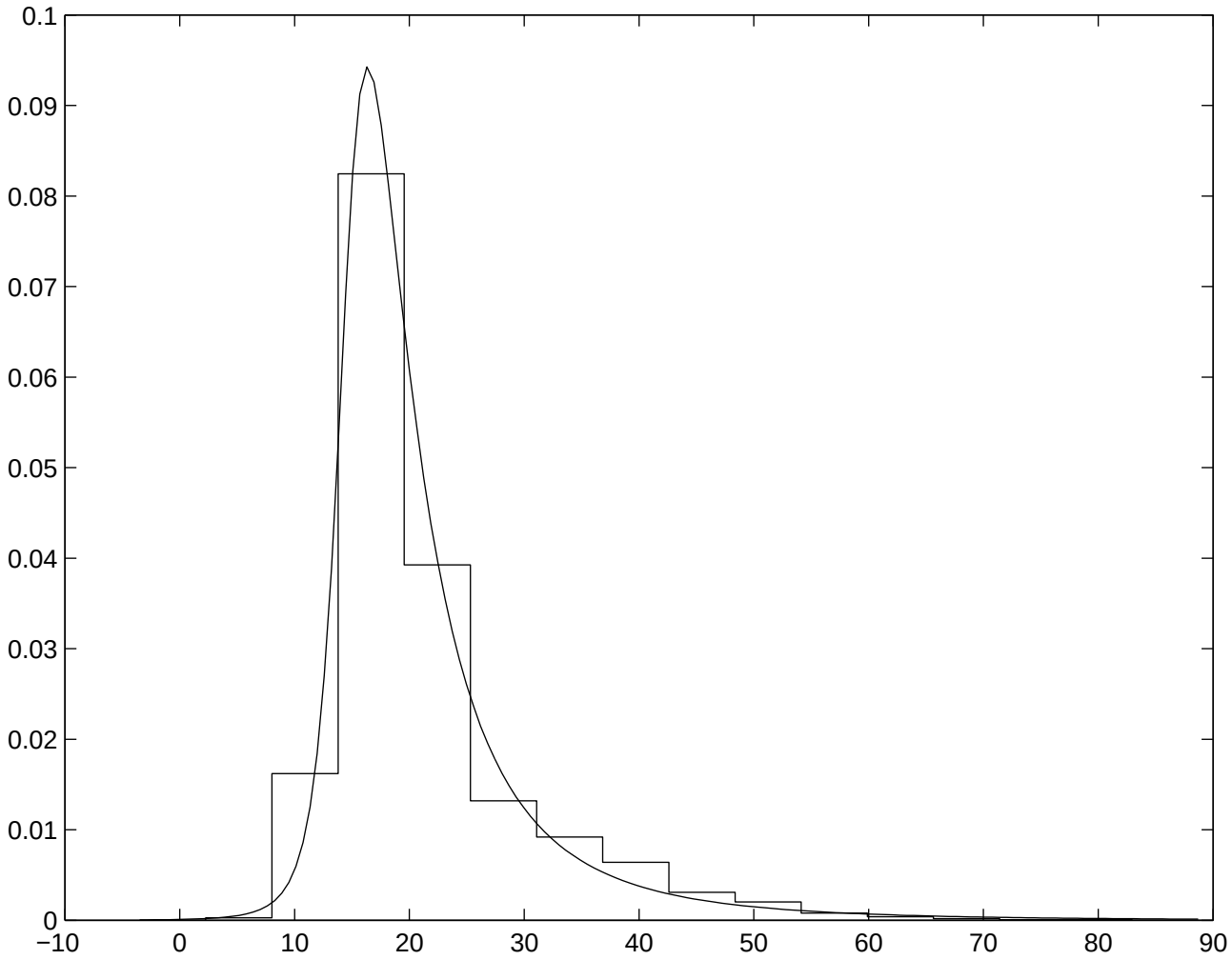
## Application of Johnson Distributions to Smart Materials Research

- Matthews et al. (2006) and Weiland et al. (2005) use a multiscale modeling approach to predict material stiffness of a certain class of smart materials called ionic polymers.
- Material stiffness depends on effective length of the polymer chains comprising the material.
- In a case study of the ionic polymer Nafion, Matthews et al. (2006) develop a simulation of polymer-chain conformation on a nanoscopic level so as to generate a large number of end-to-end chain lengths.
- The chain-length p.d.f. is estimated and used as input to a macroscopic-level mathematical model to predict material stiffness.

# Johnson $S_U$ C.d.f. Fitted to $n = 9{,}980$ Nafion Chain Lengths Using DWLS Estimation



# Johnson $S_U$ P.d.f Fitted to $n = 9{,}980$ Nafion Chain Lengths Using DWLS Estimation



- Matthews et al. (2006) and Weiland et al. (2005) obtain more accurate and intuitively appealing fits to Nafion chain-length data with Johnson p.d.f.'s than with other distributions.
  - Material stiffness is computed from the second derivative $f_X''(x)$ of the fitted p.d.f.
  - There is a relatively simple relationship between the Johnson parameters and material stiffness.

## Application of Johnson Distributions to Healthcare

- To model arrival patterns of patients who have scheduled appointments at community healthcare clinics in San Diego, Alexopoulos et al. (2008) estimate the distribution of patient tardiness—that is, deviation from the scheduled appointment time.
- Alexopoulos et al. (2008) perform an exhaustive analysis of 18 continuous distributions, concluding that the $S_U$ distribution provided superior fits to the available data.

## C. Bézier Distribution Family

### Definition of Bézier Curves

- A Bézier curve is often used to approximate a smooth function on a bounded interval by forcing the Bézier curve to pass in the vicinity of selected *control points*

$$\{\boldsymbol{p}_i \equiv (x_i, z_i)^{\mathrm{T}} : i = 0, 1, \ldots, n\}$$

in two-dimensional Euclidean space.

- A Bézier curve of degree $n$ with control points $\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\}$ is given parametrically by

$$\mathbf{P}(t) \quad = \quad \sum_{i=0}^{n} B_{n,i}(t)\, \boldsymbol{p}_i \quad \text{for} \quad t \in [0,\ 1], \tag{8}$$

where the *blending function*,

$$B_{n,i}(t) \equiv \frac{n!}{i!\,(n-i)!}\, t^i (1-t)^{n-i} \quad \text{for} \quad t \in [0, 1], \tag{9}$$

is the $i$th Bernstein polynomial for $i = 0, 1, \ldots, n$.

## Bézier Distribution and Density Functions

- If $X$ is a continuous random variable on $[a, b]$ with c.d.f. $F_X(\cdot)$ and p.d.f. $f_X(\cdot)$, then we can approximate $F_X(\cdot)$ arbitrarily closely using a Bézier curve of the form (8) by taking a sufficient number $(n+1)$ of control points with appropriate coordinates

$$\boldsymbol{p}_i = (x_i, z_i)^{\mathrm{T}}$$

for the $i$th control point, where $i = 0, \ldots, n$.

- If $X$ is Bézier, then the c.d.f. of $X$ is given parametrically by

$$\mathbf{P}(t) = \{x(t), F_X[x(t)]\}^{\mathrm{T}} \text{ for } t \in [0, 1], \tag{10}$$

where

$$\left.\begin{aligned} x(t) &= \sum_{i=0}^{n} B_{n,i}(t) x_i, \\ F_X[x(t)] &= \sum_{i=0}^{n} B_{n,i}(t) z_i \end{aligned}\right\} \text{ for } t \in [0, 1]. \tag{11}$$

For a detailed discussion of Bézier distributions, see

Wagner, M. A. F., and J. R. Wilson. 1996a. Using univariate Bézier distributions to model simulation input processes. *IIE Transactions* 28 (9): 699–711. Available online via

<www.ise.ncsu.edu/jwilson/files/wagner96iie.pdf>

- If $X$ is Bézier with c.d.f. $F_X(\cdot)$ given by (10), then the p.d.f. $f_X(x)$ is

$$\mathbf{P}^*(t) \quad = \quad \{x(t),\, f_X[x(t)]\}^{\mathrm{T}} \text{ for } t \in [0,\, 1],$$

where $x(t)$ is given by (11) and

$$f_X[x(t)] \quad = \quad \frac{\sum_{i=0}^{n-1} B_{n-1,i}(t)\Delta z_i}{\sum_{i=0}^{n-1} B_{n-1,i}(t)\Delta x_i},$$

where

$$\Delta x_i \equiv x_{i+1} - x_i \text{ and } \Delta z_i \equiv z_{i+1} - z_i \text{ for } i = 0, 1, \ldots, n-1.$$

## Generating Bézier Variates by Inversion

**[1]** Generate a random number $U \sim \text{Uniform}[0,\ 1]$.

**[2]** Find $t_U \in [0,\ 1]$ such that

$$F_X[x(t_U)] = \sum_{i=0}^{n} B_{n,i}(t_U) z_i = U. \tag{12}$$

**[3]** Deliver the variate

$$X = x(t_U) = \sum_{i=0}^{n} B_{n,i}(t_U) x_i \ .$$

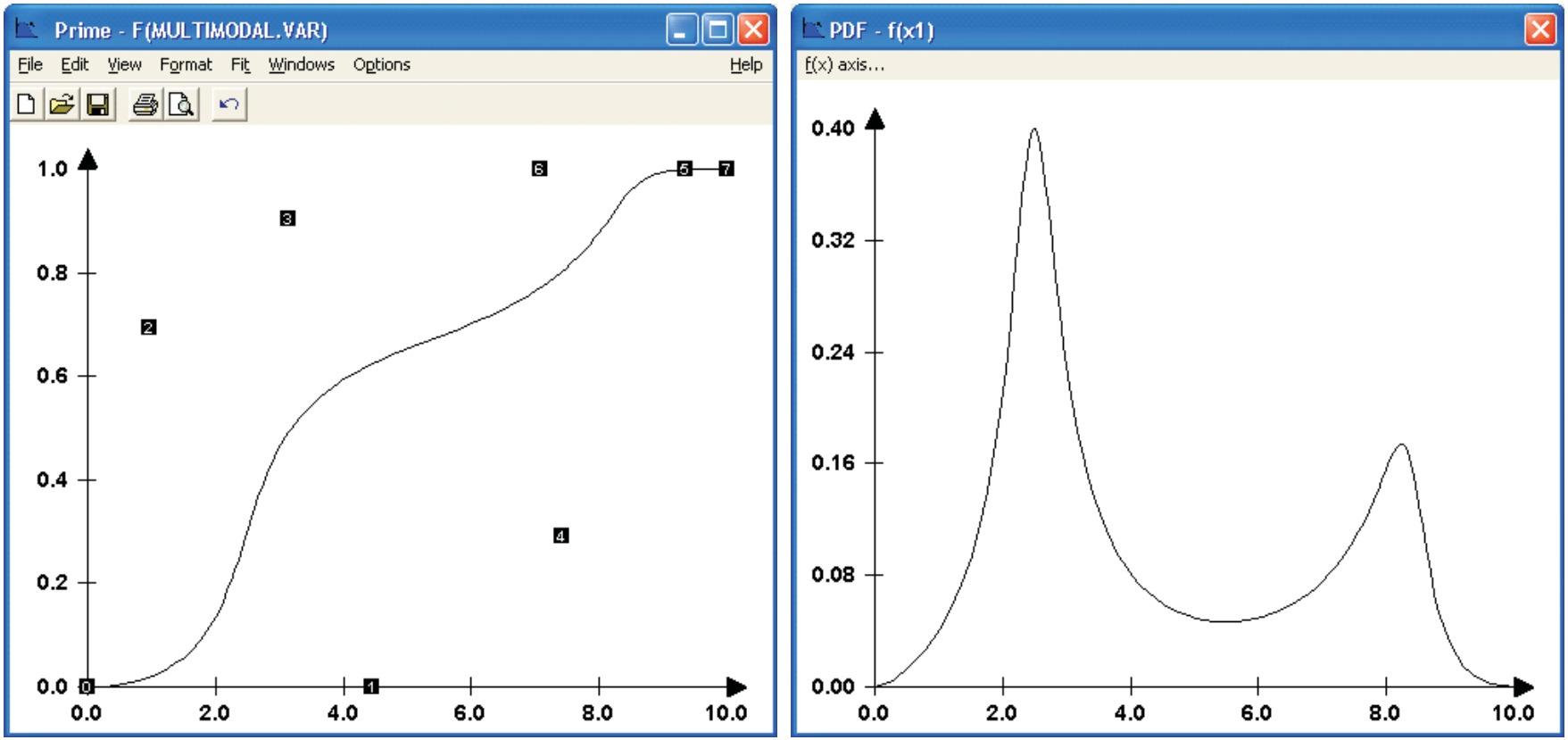
Codes to implement this approach are available on the Web via

<www.ise.ncsu.edu/jwilson/page3>.

## Using PRIME to Model Bézier Distributions

- PRIME (Wagner and Wilson 1996a) is a Windows-based system for fitting Bézier distributions to data or subjective information.
- PRIME is available on the previously mentioned Web site.
- Control points appear as indexed black squares that can be manipulated with the mouse and keyboard.
  - Each control point exerts on the c.d.f. a “magnetic” attraction whose strength is given by the associated Bernstein polynomial (9).
  - Moving a control point causes the displayed c.d.f. to be updated (nearly) instantaneously.

# PRIME Windows Showing the Bézier C.d.f. (Left Panel) with Its Control Points and the P.d.f. (Right Panel)



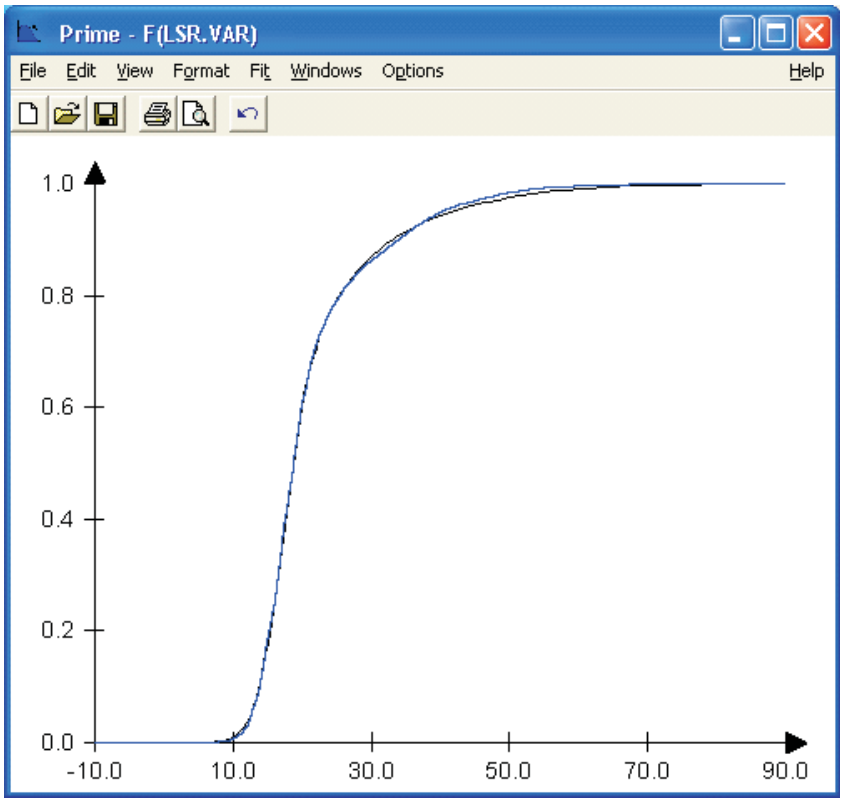
PRIME includes the following methods for fitting Bézier distributions to sample data:

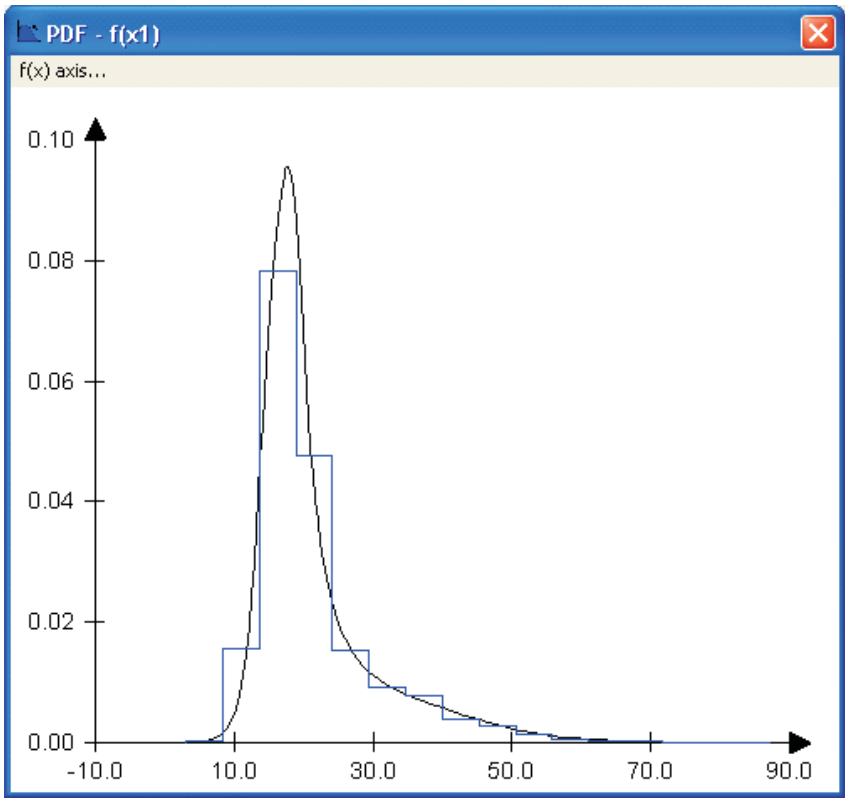
- OLS estimation of the c.d.f.;
- minimum $L_1$ and $L_\infty$ norm estimation of the c.d.f.;
- maximum likelihood estimation (assuming $a$ and $b$ are known);
- moment matching; and
- percentile matching.

For automatic estimation of the number of control points, see

Wagner, M. A. F., and J. R. Wilson. 1996b. Recent developments in input modeling with Bézier distributions. In *Proceedings of the 1996 Winter Simulation Conference*, pp. 1448–1456. Available online as <www.ise.ncsu.edu/jwilson/files/bezwsc96.pdf>.

# Bézier Distribution Fitted to $n = 9{,}980$ Nafion Chain Lengths Using OLS Estimation of the C.d.f.




Stat F(x1) - LSR.DAT

| | CDF | Data |
|---|---|---|
| Mean | 21.465 | 21.317 |
| Std Dev | 9.542 | 8.837 |
| Skewness | 2.349 | 1.975 |
| Kurtosis | 10.130 | 7.828 |
| Min | 3.037 | 3.056 |
| Max | 83.902 | 82.149 |
| Num | 14 | 9980 |

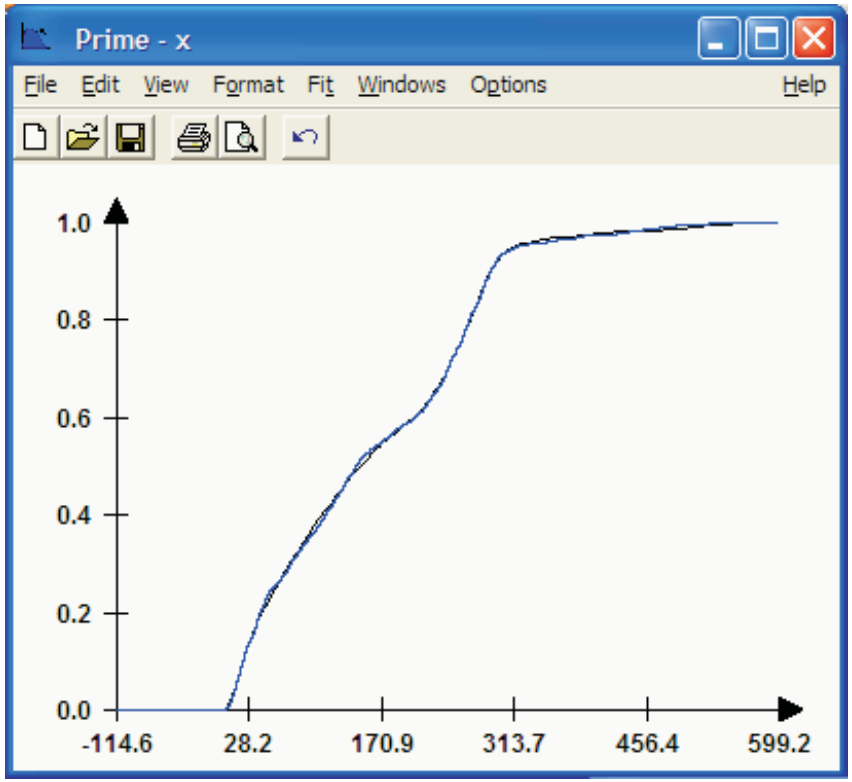
Fit Measures - F(x1)

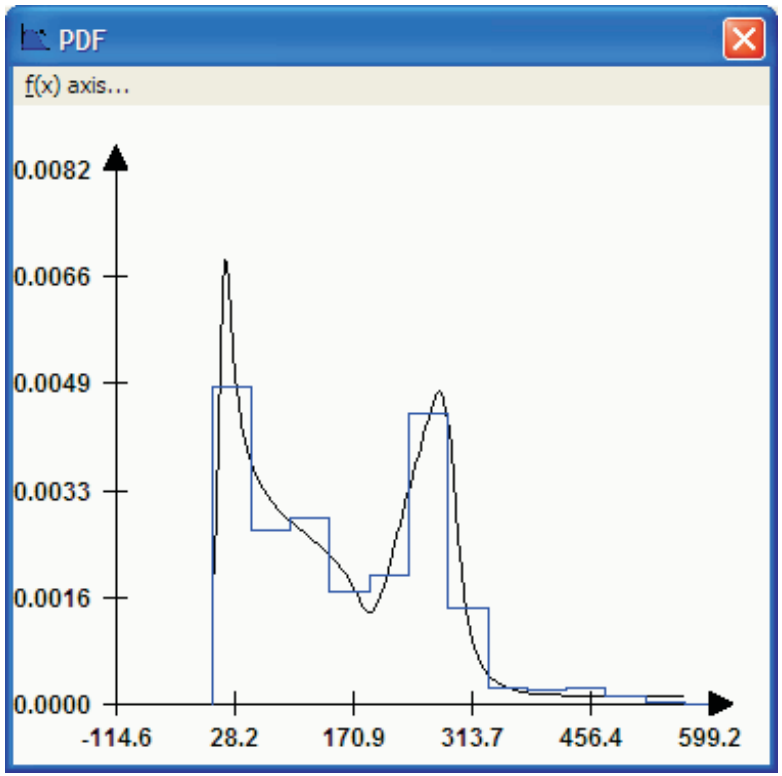
| | |
|---|---|
| Least Squares | 2.907e-001 |
| L-One Norm | 4.964e-003 |
| L-Infinity Norm | 1.162e-002 |
| Moment Matching | 5.958e+000 |
| Log-Likelihood | -1.414e+004 |
| Percentile Matching | 1.214e+000 |

Percentile Information - F(x1)

| Percentile | CDF | Data |
|---|---|---|
| 10 | 13.777 | 13.880 |
| 25 | 15.933 | 16.020 |
| 50 | 18.678 | 18.748 |
| 75 | 23.298 | 23.236 |
| 90 | 33.121 | 34.013 |

# Bézier Distribution Fitted to $n = 2{,}083$ Capacitor Lot Sizes Using OLS Estimation of the C.d.f. and Automatic Determination of the Number of Control Points




BIMODS.DAT

| | CDF | Data |
|---|---|---|
| Mean | 160.091 | 160.303 |
| Std Dev | 112.737 | 112.551 |
| Skewness | 0.514 | 0.473 |
| Kurtosis | 2.863 | 2.650 |
| Min | -0.002 | 0.000 |
| Max | 569.906 | 569.383 |
| Num | 13 | 2083 |

BIMODS.DAT

| | CDF | Data |
|---|---|---|
| Mean | 160.091 | 160.303 |
| Std Dev | 112.737 | 112.551 |
| Skewness | 0.514 | 0.473 |
| Kurtosis | 2.863 | 2.650 |
| Min | -0.002 | 0.000 |
| Max | 569.906 | 569.383 |
| Num | 13 | 2083 |

Percentile Information

| Percentile | CDF | Data |
|---|---|---|
| 10 | 20.941 | 21.615 |
| 25 | 56.197 | 53.028 |
| 50 | 147.521 | 142.961 |
| 75 | 255.096 | 255.189 |
| 90 | 288.191 | 287.868 |

Advantages of the Bézier distribution family:

- It is extremely flexible and can represent a wide diversity of distributional shapes, including multiple modes and mixed distributions.
- If data are available, then the likelihood ratio test of Wagner and Wilson (1996b) can be used with any of the available estimation methods to find automatically both the number and location of the control points.
- In the absence of data, PRIME can be used to determine the conceptualized distribution based on known quantitative or qualitative information.
- As the number $(n + 1)$ of control points increases, so does the flexibility in fitting Bézier distributions.

# III. Time-Dependent Arrival Processes

- Many simulation applications require high-fidelity input models of arrival processes with arrival rates that depend strongly on time.
- Nonhomogeneous Poisson processes (NHPPs) have been used successfully to model complex time-dependent arrival processes in many applications.
- An NHPP $\{N(t) : t \geq 0\}$ is a counting process such that
  - $N(t)$ is the number of arrivals in the time interval $(0, t]$;
  - $\lambda(t)$ is the instantaneous arrival rate at time $t$, and $\lambda(t)$ satisfies the Poisson postulates; and
  - the (cumulative) mean-value function is given by

$$\mu(t) \equiv \mathrm{E}[N(t)] = \int_0^t \lambda(z)\,\mathrm{d}z \quad \text{for all} \quad t \geq 0. \tag{13}$$

- We discuss the nonparametric approach of Leemis (1991, 2000, 2004) for modeling and simulation of NHPPs; see

Leemis, L. M. 1991. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process. *Management Science* 37 (7): 886–900.

Leemis, L. M. 2000. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process from overlapping realizations. *Management Science* 46 (7): 989–998.

Leemis, L. M. 2004. Nonparametric estimation and variate generation for a nonhomogeneous Poisson process from event count data. *IIE Transactions* 36:1155–1160.

- The context is a recent application to modeling and simulating unscheduled patient arrivals to a community healthcare clinic (Alexopoulos et al. 2008)

- Suppose we have a time interval $(0, S]$ over which we observe several independent replications (realizations) of a stream of unscheduled patient arrivals constituting an NHPP with arrival rate $\lambda(t)$ for $t \in (0, S]$.

  For example, $(0, S]$ might represent the time period on each weekday during which unscheduled patients may walk into a clinic—say, between 9 A.M. and 5 P.M. so that $S = 480$ minutes.

- Suppose $k$ realizations of the arrival stream over $(0, S]$ have been recorded so that we have
  - $n_i$ patient arrivals in the $i$th realization for $i = 1, 2, \ldots, k$; and
  - $n = \sum_{i=1}^{k} n_i$ patient arrivals accumulated over all realizations.

- Let $\{t_{(i)} : i = 1, \ldots, n\}$ denote the overall set of arrival times for all unscheduled patients expressed as an offset from the beginning of $(0, S]$ and then sorted in increasing order.

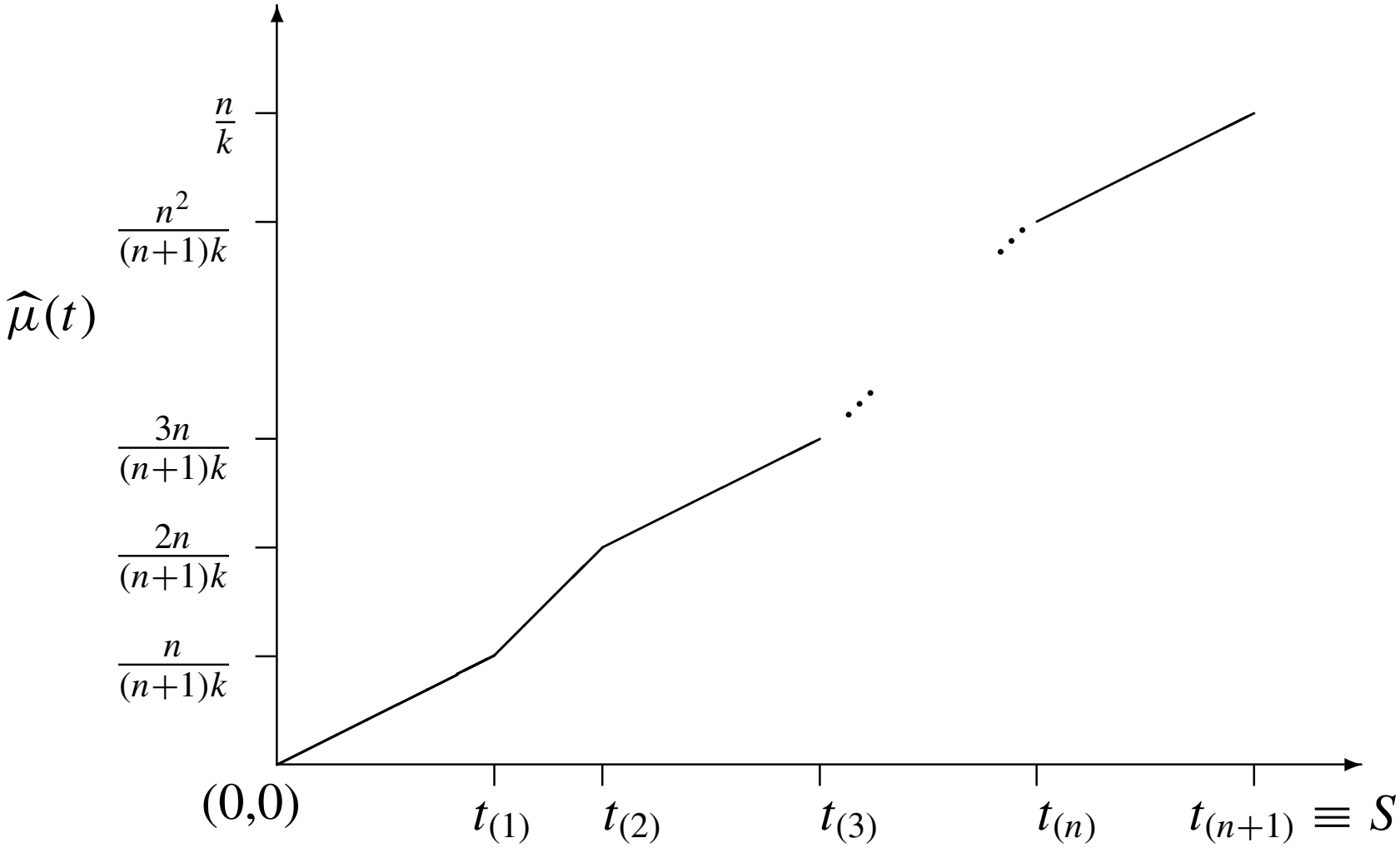
For example, if we observed $n = 250$ patient arrivals over $k = 5$ days, each with an observation interval of length $S = 480$ minutes, then

- $t_{(1)} = 2.5$ minutes means that over all 5 days, the earliest arrival occurred 2.5 minutes after the clinic opened on one of those days; and
- $t_{(2)} = 4.73$ minutes means that the second-earliest arrival occurred 4.73 minutes after the clinic opened on one of those days.
- $t_{(n)} = 478.5$ minutes means that the latest arrival occurred 478.5 minutes after the clinic opened on one of those days.

- We estimate the mean-value function $\mu(t)$ as follows.
  - We take $t_{(0)} \equiv 0$ and $t_{(n+1)} \equiv S$.
  - For $t_{(i)} < t \leq t_{(i+1)}$ and $i = 0, 1, \ldots, n$, we take

$$\hat{\mu}(t) = \frac{i\,n}{(n+1)k} + \left\{ \frac{n\left[t - t_{(i)}\right]}{(n+1)k\left[t_{(i+1)} - t_{(i)}\right]} \right\}. \tag{14}$$

- Equation (14) provides a basis for modeling and simulating unscheduled patient-arrival streams when the arrival rate exhibits a strong dependence on time.




Nonparametric Estimator of Mean Value Function

## Goodness-of-fit Testing for the Fitted Mean-Value Function

- In addition to the realizations of the target arrival process that were used to compute the estimated mean-value function $\widehat{\mu}(t)$, suppose we observe one additional realization

$$\{A'_i : i = 1, 2, \ldots, n'\}$$

  independently of the previously observed realizations, with the $i$th patient arriving at time $A'_i$ for $i = 1, \ldots, n'$.

- If the target arrival stream is an NHPP with mean-value function $\mu(t)$ for $t \in (0, S]$, then the transformed arrival times

$$\{B'_i = \mu(A'_i) : i = 1, 2, \ldots, n'\}$$

  constitute a homogeneous Poisson process with an arrival rate of 1.

- If the target arrival stream is an NHPP with mean-value function $\mu(t)$ for $t \in (0, S]$, then the corresponding transformed interarrival times

$$\{X_i' = B_i' - B_{i-1}' : i = 1, 2, \ldots, n'\}$$

(with $B_0' \equiv 0$) constitute a random sample from an exponential distribution with a mean of 1.

- To test the adequacy of the fitted mean-value function $\hat{\mu}(t)$ as an approximation to $\mu(t)$, apply the Kolmogorov-Smirnov test to the data set

$$\{X_i'' = \hat{\mu}(A_i') - \hat{\mu}(A_{i-1}') : i = 1, 2, \ldots n'\}$$

(with $A_0' \equiv 0$), where the hypothesized c.d.f. in the goodness-of-fit test is

$$F_{X_i''}(x) = 1 - e^{-x} \quad \text{for all} \quad x \geq 0.$$

# Generating Realizations of the Fitted NHPP

**[1]** Set $i \leftarrow 1$ and $N \leftarrow 0$.
**[2]** Generate $U_i \sim \text{Uniform}(0, 1)$.
**[3]** Set $B_i \leftarrow -\ln(1 - U_i)$.
**[4]** **While** $B_i < n/k$ **do**
**Begin**

Set $m \leftarrow \left\lfloor \dfrac{(n+1)kB_i}{n} \right\rfloor$;

Set $A_i \leftarrow t_{(m)} + \{t_{(m+1)} - t_{(m)}\} \left\{ \dfrac{(n+1)kB_i}{n} - m \right\}$;

Set $N \leftarrow N + 1$; Set $i \leftarrow i + 1$;
Generate $U_i \sim \text{Uniform}(0, 1)$;
Set $B_i \leftarrow B_{i-1} - \ln(1 - U_i)$.
**End**

NHPP Simulation Procedure of Leemis (1991)

## Advantages of Leemis's Nonparametric Approach to Modeling and Simulation of NHPPs

- It does not require the assumption of any particular form for arrival rate $\lambda(t)$ as a function of $t$.
- It provides a strongly consistent estimator of mean-value function—that is,

$$\lim_{k \to \infty} \widehat{\mu}(t) = \mu(t) \text{ for all } t \in (0, S] \text{ with probability 1.}$$

- The simulation algorithm given above, which is based on inversion of $\widehat{\mu}(t)$ so that

$$A_i = \widehat{\mu}^{-1}(B_i) \text{ for } i = 1, \ldots, N,$$

is also asymptotically valid as $k \to \infty$.

## Application to Organ Transplantation Policy Analysis

- The United Network for Organ Sharing (UNOS) applied a simplified variant of this approach in the development and use of the UNOS Liver Allocation Model (ULAM) for analyzing the cadaveric liver-allocation system in the U.S. (see Harper et al. 2000).
- ULAM incorporated models of
  - (*a*) the streams of liver-transplant patients arriving at 115 transplant centers, and
  - (*b*) the streams of donated organs arriving at 61 organ procurement organizations in the United States.

  Virtually all these arrival streams exhibited long-term trends as well as strong dependencies on the time of day, the day of the week, and the geographic location of the arrival stream.

## Handling Arrival Processes Having Trends or Cyclic Effects

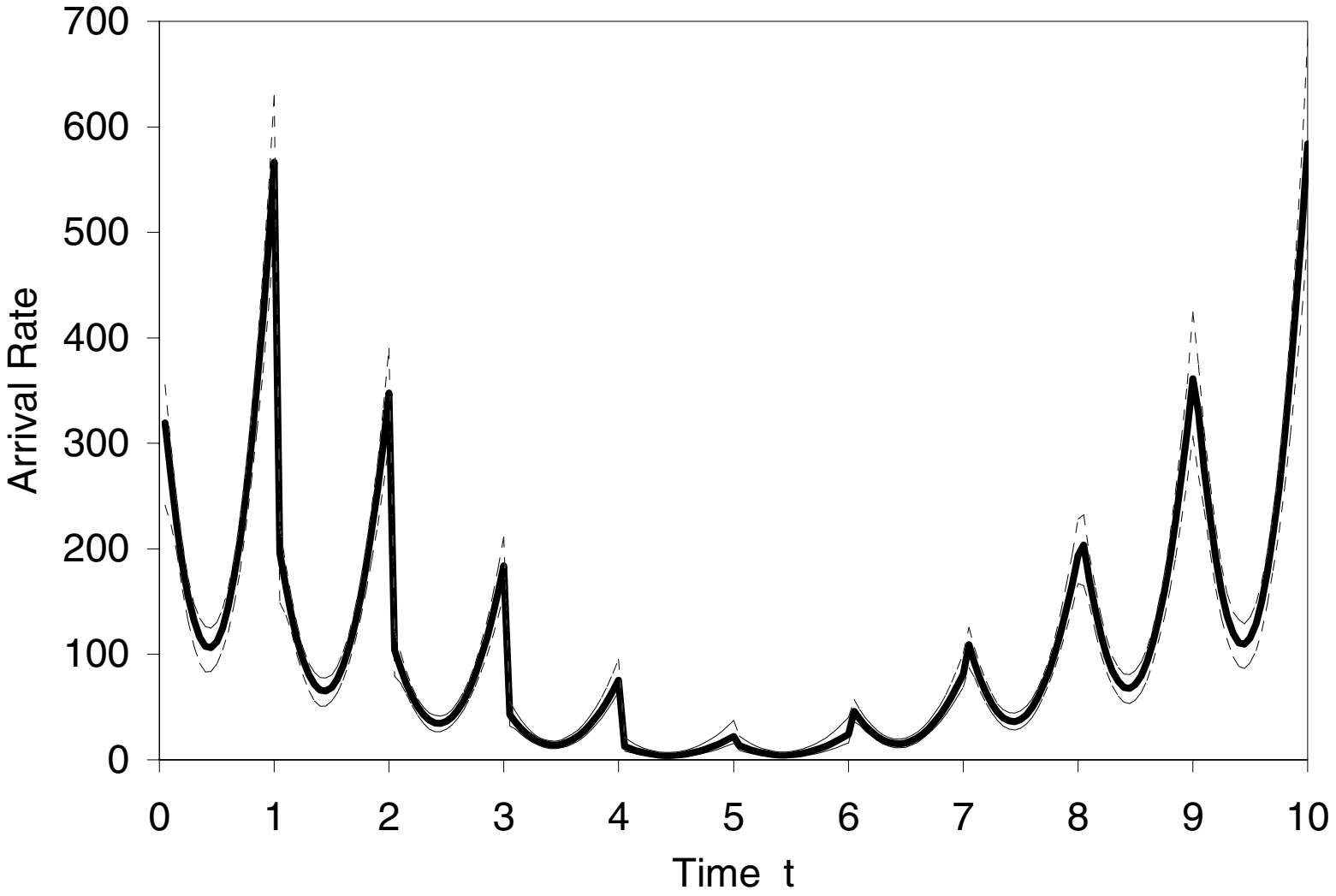
- Kuhl, Sumant, and Wilson (2006) and Kuhl, Deo, and Wilson (2008) develop a smooth, flexible, "semiparametric" method for modeling and simulating NHPP arrival processes given one or more process realizations.

- Designed for arrival processes that may exhibit
  - A long-term trend or
  - Nested periodic phenomena (such as daily and weekly cycles),

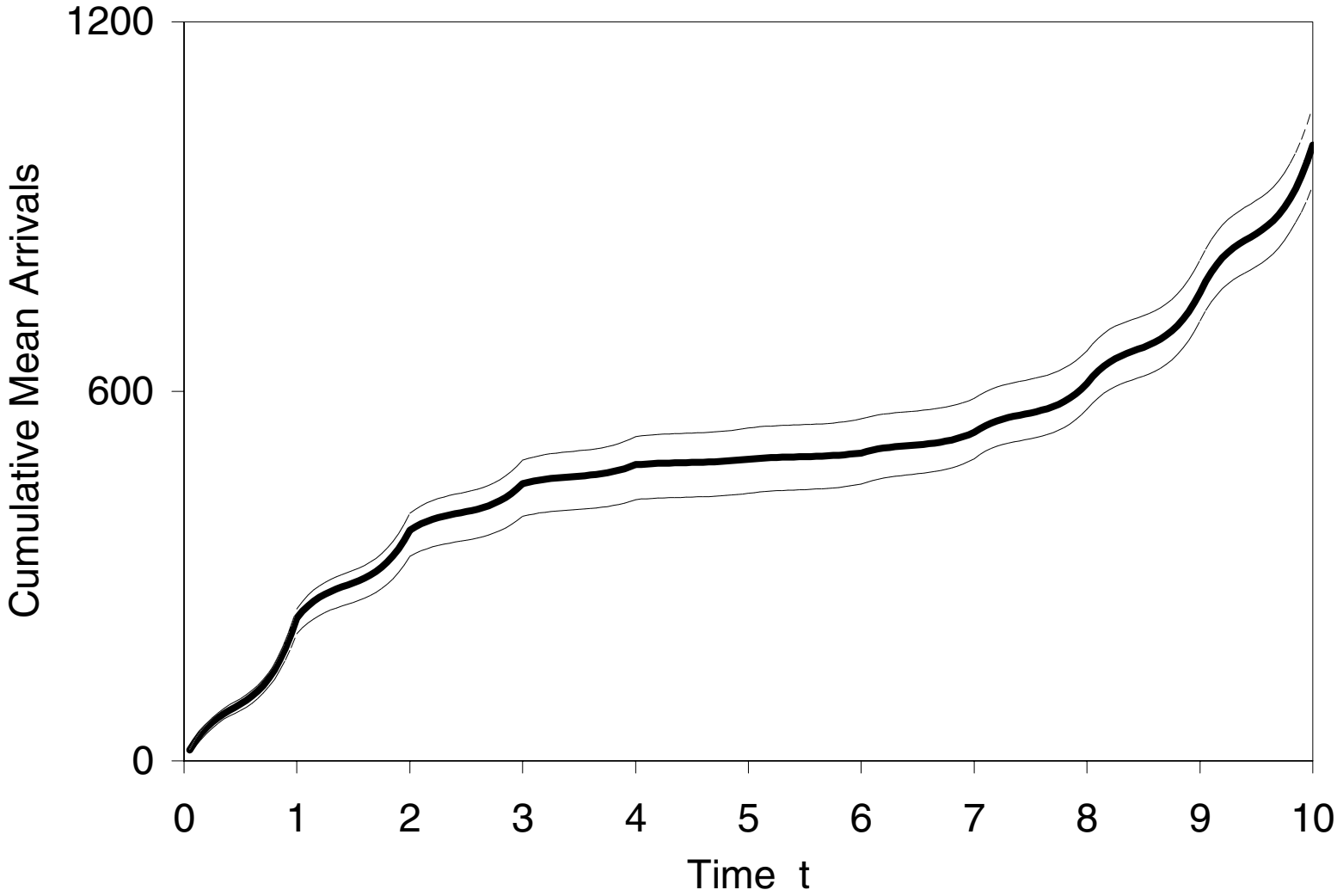
  where the latter effects might not necessarily possess the symmetry of sinusoidal oscillations.

- In the case of *known* nested periodic components a multiresolution procedure is applied to the observed process data.

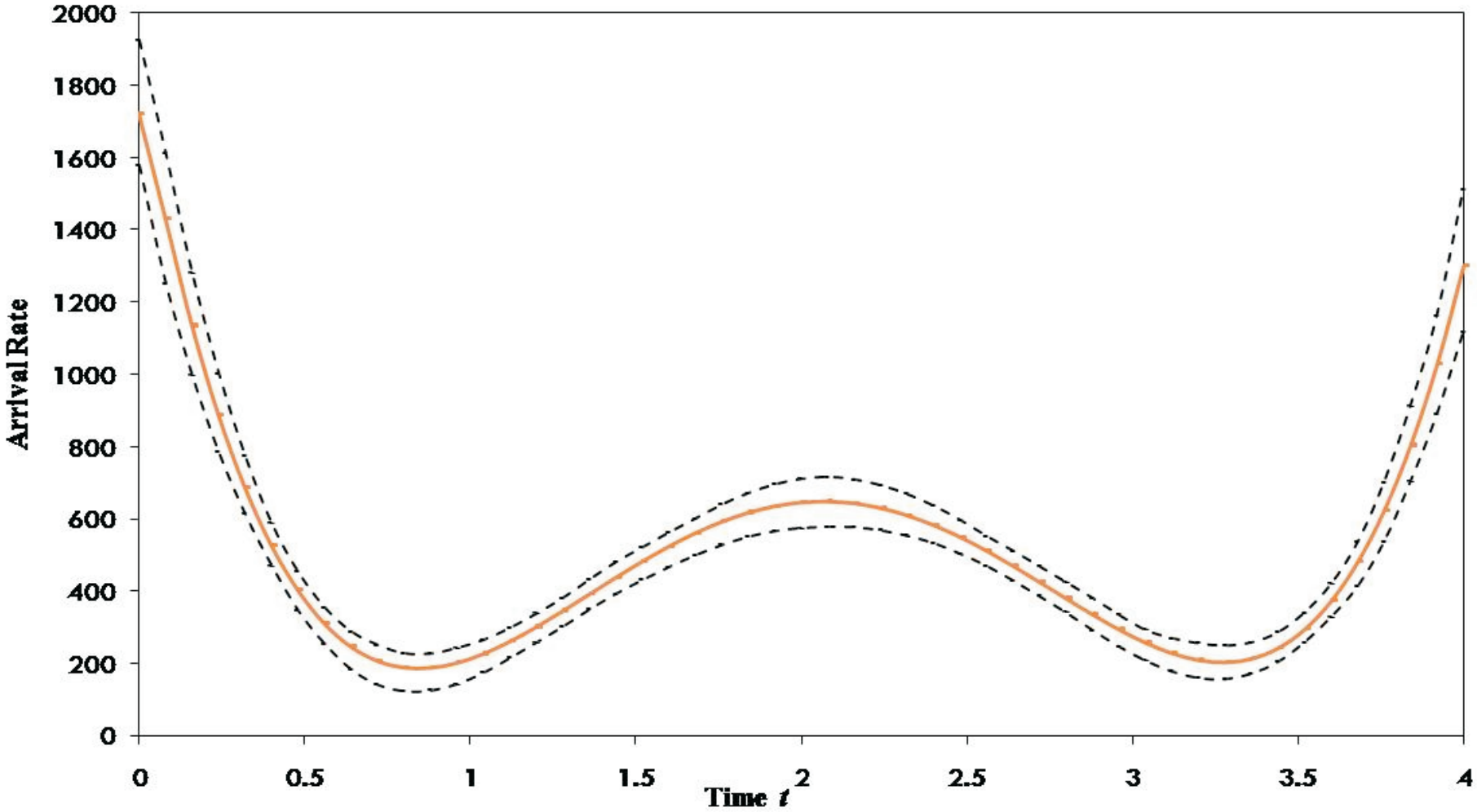
# Fitted Rate Function over 100 Replications of a Test Process with One Cyclic Rate Component and Long-term Trend



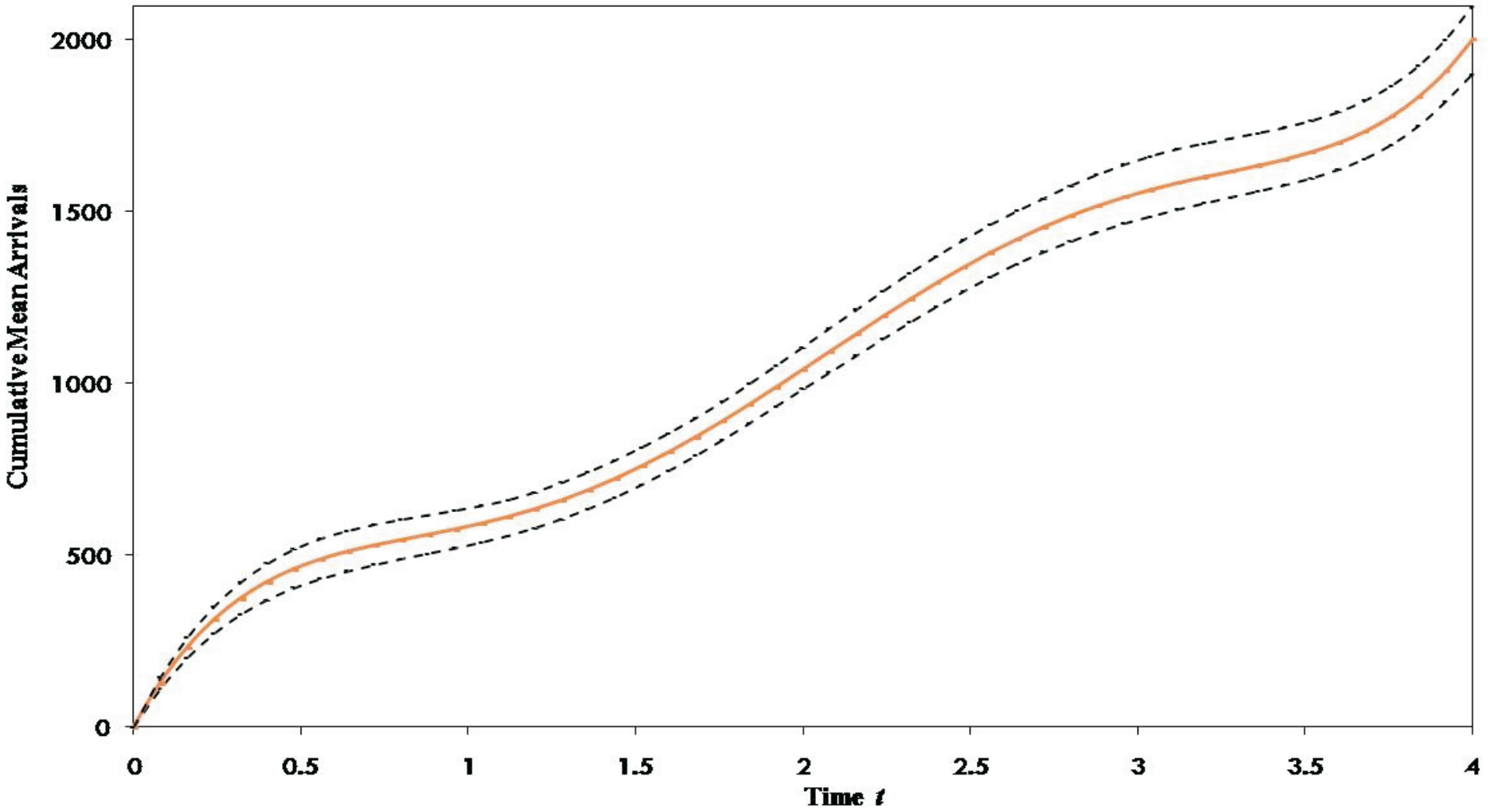
# Fitted Mean-Value Function over 100 Replications of a Test Process with One Cyclic Rate Component and Long-term Trend



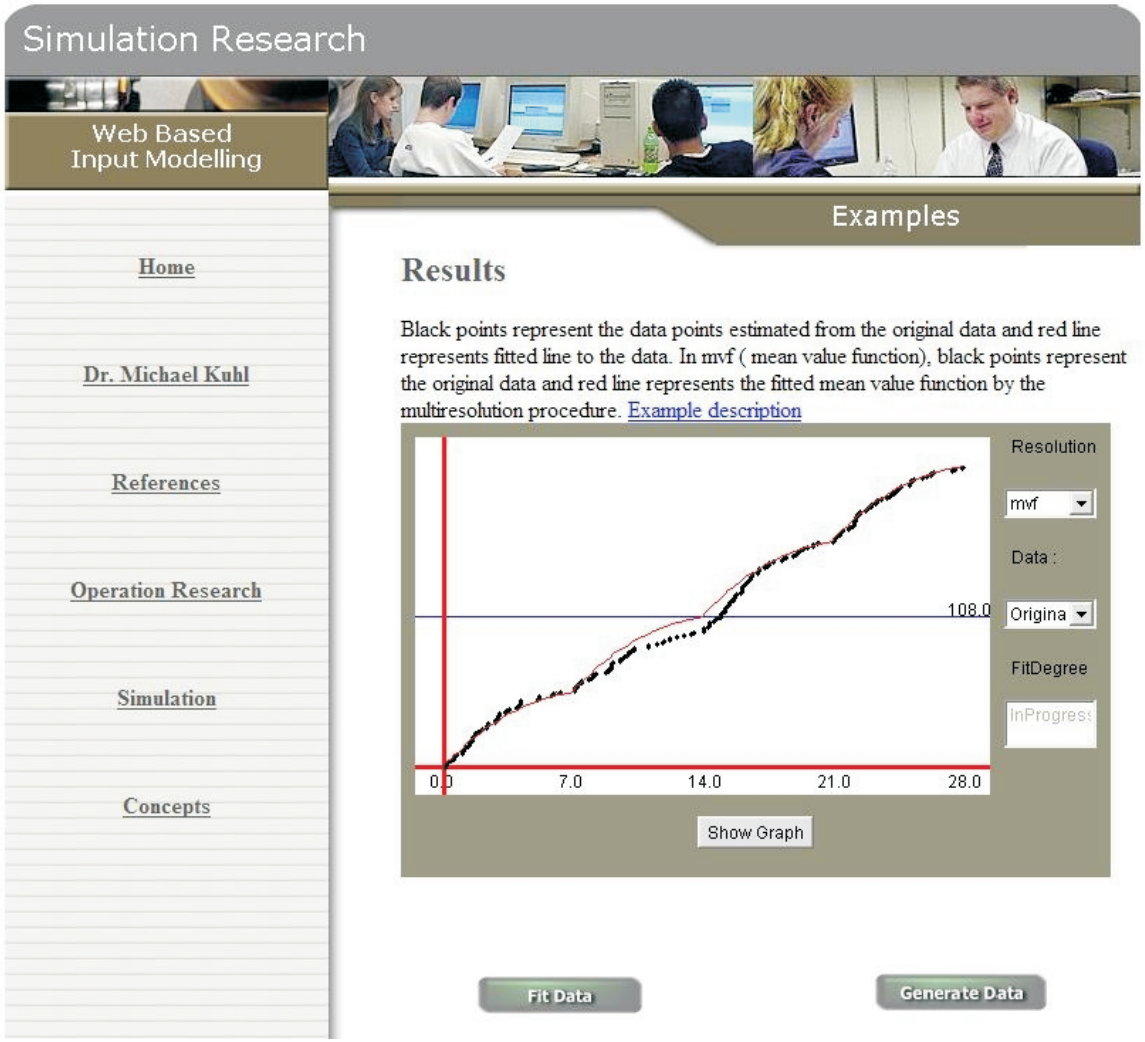
# Fitted Rate Function over 100 Replications of a Test Process with Multiple Observed Process Realizations



# Fitted Mean-Value Function over 100 Replications of a Test Process with Multiple Observed Process Realizations



# Web-based NHPP Input Modeling Software




<www.rit.edu/simulation>

# IV. Application of Beta Distributions to Medical Decision Making

Xu et al. (2009) develop a decision tree model for determining the cost effectiveness of cesarean delivery upon maternal request (CDMR) for women having a single childbirth without indications.

- Their model compares CDMR to trial of labor (TOL) considering all possible short- and long-term outcomes and resulting consequences for the mother and neonate.
- This yields a decision-tree model with over 100 chance events.

## Application of Beta Distributions to Medical Decision Making (Cont'd)

- For each parameter in their model, Xu et al. use either literature-based or expert opinion–based estimates for $\hat{a}$, $\hat{m}$, and $\hat{b}$.
- The outcome probabilities and utilities are fitted to beta distributions using the rapid estimates (4) of the required shape parameters.
- For each fitted beta distribution with user-specified endpoints $\hat{a}$ and $\hat{b}$ and shape parameters $\hat{\alpha}_1$ and $\hat{\alpha}_2$ computed via (4), the theoretical mode (3) matched the target value $\hat{m}$ with error less than $10^{-8}$.

# Application of Beta Distributions to Medical Decision Making (Cont'd)

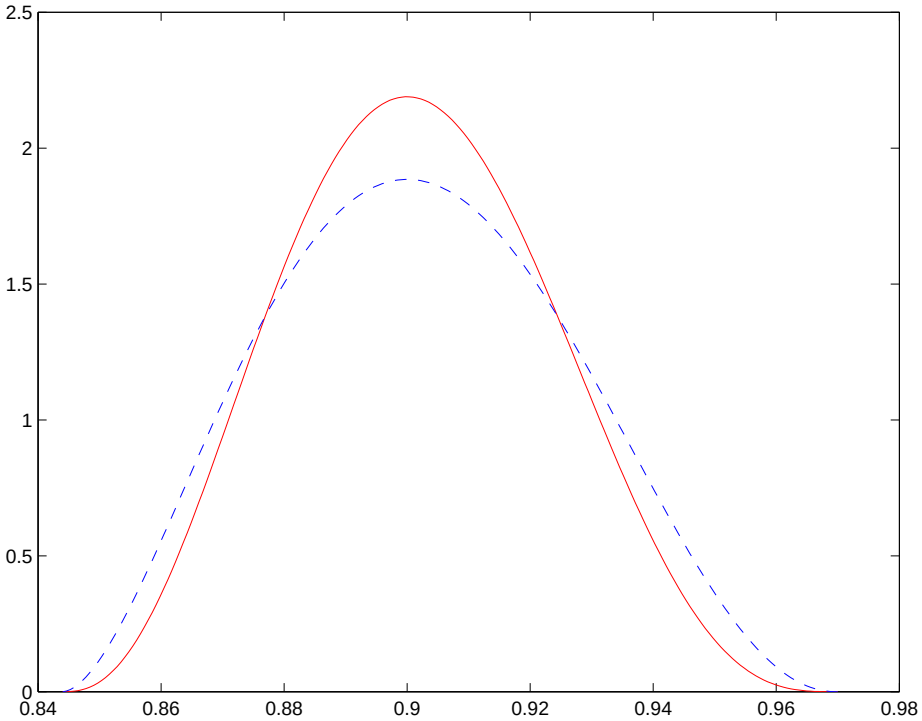
- While the "minimum" and "maximum" values are the smallest and largest values found in the available literature, we recognize that the true lower and upper limits for each target distribution may be outside of this range.
- To account for this possibility, we explore the effect of taking relevant offsets from $\hat{a}$ and $\hat{b}$ expressed as a fraction $\psi$ of the difference $\hat{b} - \hat{a}$:

$$\hat{a}' = \max\left\{\hat{a} - \psi\left(\hat{b} - \hat{a}\right), 0\right\} \text{ and } \hat{b}' = \min\left\{\hat{b} + \psi\left(\hat{b} - \hat{a}\right), 1\right\}.$$

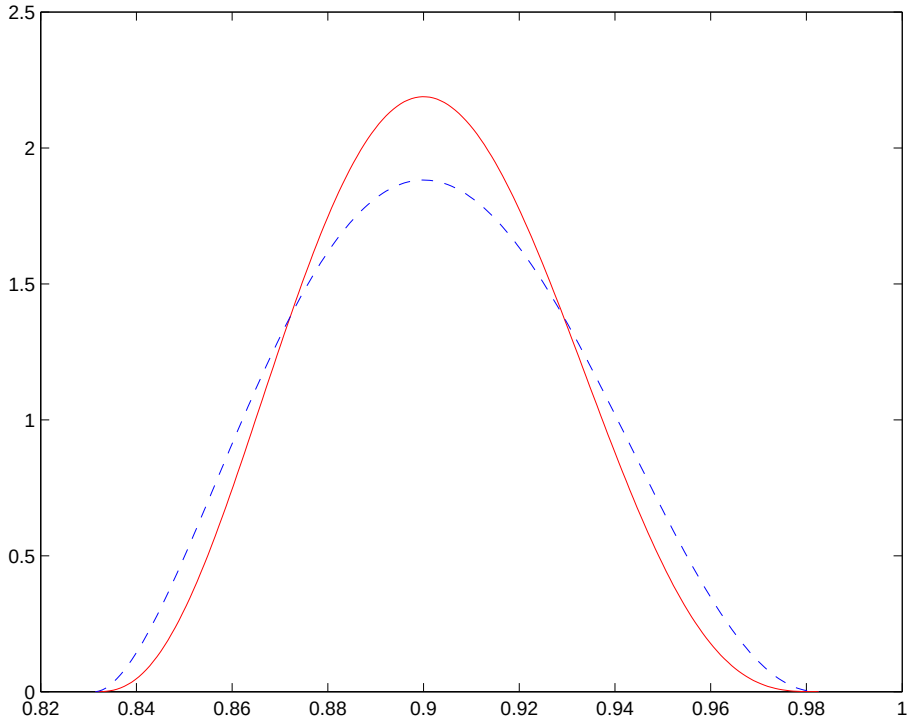
- We varied $\psi$ from 0 to 0.1 to examine the effect on the cost-effectiveness comparison of CDMR and TOL.

## Application of Beta Distributions to Medical Decision Making (Cont'd)

- For $\psi \in [0, 0.02)$ , there was a significant difference in the effectiveness of CDMR and TOL (i.e., the 95% confidence interval for the mean difference in the utility between CDMR and TOL did not include zero) when using beta distributions fitted by either (4) or the RiskPert method.
- For $\psi \in [0.02, 0.07]$ , there was a significant difference in the effectiveness of CDMR and TOL only when using beta distributions fitted via (4).
- For $\psi > 0.07$ , the difference in effectiveness of CDMR and TOL was not significant for either method of fitting beta distributions.




**(a)** $P(\text{Vag}),\ \psi = 0.0$




**(b)** $P(\text{Vag}),\ \psi = 0.10$

Beta distributions fitted to $P(\text{Vag})$, the probability of vaginal delivery, where the solid red line is the fit using Equation (4) and the dashed blue line is the RiskPert fit.

# V. Conclusions and Recommendations

- The common thread running through this tutorial is the focus on robust input models that are
  - computationally tractable and
  - sufficiently flexible to represent adequately many of the probabilistic phenomena that arise in many applications of discrete-event stochastic simulation.
- Notably absent is a discussion of Bayesian input-modeling techniques—a topic that will receive increasing attention in the future.
- Additional material on input modeling is available via

  <www.ise.ncsu.edu/jwilson/more_info>.