NASA-TP-2389 19850008256

NASA
Technical
Paper
2389

January 1985

Spline Methods for Approximating Quantile Functions and Generating Random Samples

James R. Schiess and
Christine G. Matthews

LIBRARY COPY
LANGLEY RESEARCH CENTER
LIBRARY, NASA
HAMPTON, VIRGINIA

NASA

NASA Technical Paper 2389

1985

Spline Methods for Approximating Quantile Functions and Generating Random Samples

James R. Schiess
Langley Research Center
Hampton, Virginia

Christine G. Matthews
Computer Sciences Corporation
Hampton, Virginia

National Aeronautics and Space Administration

Scientific and Technical Information Branch

SUMMARY

Data gathered under experimental conditions are often used to obtain an analytic model for generating further data with similar random characteristics. The quantile function (inverse of the cumulative distribution function) is very useful in this respect. With this function, a probability is first randomly selected from a uniform distribution. The quantile function then associates a random value with this probability. In this paper, two spline functions (B-spline and rational spline) are used to approximate the quantile function from a sample of data. Also, an analytic representation of the quantile function is used for comparison. These approximations are used to generate random samples for simulations. Comparisons are made between the three representations for samples generated from known distributions and for a sample of experimental data. The spline representations are shown to be more accurate for multimodal and skewed samples and to require much less time to generate samples than the analytic representation.

INTRODUCTION

An important aspect of scientific research is the analysis of data gathered under experimental conditions. Because of the stochastic nature of the data, it may be desirable to determine the probability distribution of the data sample. In some instances, it is reasonable to assume that the data were sampled from a continuous population obeying the rules of a standard probability distribution, such as the normal distribution. In other instances, the distributional nature of the sample may not be obvious. For these latter cases, the researcher may represent the underlying population distribution with histograms, analytic functions, or function approximations (ref. 1). Histograms provide only a rough, discrete version of a continuous function. Analytic functions often are limited to unimodal distributions having restricted skewness and kurtosis ranges. Function approximation methods can overcome the weaknesses of the other approaches.

In seeking to sample information to describe the distribution of a population, a researcher can choose to approximate the cumulative distribution function (CDF), the quantile function (inverse of the CDF), or the probability density function (PDF). The CDF is the probability of occurrence of a value less than or equal to a given value of the random variable. To approximate either the CDF or the quantile function from a sample, the probability of occurrence of each sample value must be either known or assumed. The PDF is the first derivative of the CDF and may be fitted to histograms or sample values by various complicated classical or nonparametric methods (ref. 2). The intent of the present investigation is to study a relatively simple, straightforward representation of the relationship between random values and the corresponding probabilities.

The study reported here considers spline representations of the sample quantile function of a continuous probability distribution. These representations provide both a functional description of a random sample and a method of generating random variables. The major purpose of this research is to develop a method for generating continuous random values that are distributed similarly to a random sample of unknown origin. Determining a quantile function that adequately represents the random sample constitutes a major portion of the method. Two spline formulations are considered.

The first (ref. 3) consists of a linear combination of cubic basis splines (B-splines). The second (ref. 4) is the rational spline, which has a separate tension parameter for each subinterval defined by two knots. With a zero tension parameter, the rational spline reduces to a cubic spline; for infinite tension, the rational spline becomes a linear function. Both spline formulations based on equally spaced knots are fit in a least squares sense to the sample quantile function.

The usefulness of both spline formulations for representing the quantile function and generating samples is shown with samples obtained from standard distributions. The spline capabilities are further illustrated in an example using experimental data. The spline results for both simulated and experimental data are compared with the results from an analytic representation of the quantile function developed by Ramberg and Schmeiser (refs. 5 and 6).

SYMBOLS

a_i	coefficient of $B_i(p)$ in B-spline representation of quantile function
$B_i(p)$	B-spline function centered at $\bar{p}_i$
$f(x)$	probability density function of x
$F(x)$	cumulative distribution function of x
$F(10,9)$	F-distribution with 10 and 9 degrees of freedom
h	distance between abscissa coordinates of knots, equal to $1/k$
H_i	constant for interval i in rational spline
k	number of subintervals in spline representations
M_r	rth moment about zero
n	number of points in sample
p	probability, $0 \leq p \leq 1$
p_j	probability of jth ordered sample point, equal to $(j - 0.5)/n$
$\bar{p}_i$	abscissa of ith knot, equal to $(i - 2)h$
$Q(p)$	quantile function
$Q_B(p)$	B-spline representation of quantile function
$Q_G(p)$	generalized lambda distribution representation of quantile function
$Q_R(p)$	rational spline representation of quantile function
$Q_S(p)$	sample quantile function
t	function of probability on ith subinterval, equal to $1 - u$

T_i rational spline tension parameter for ith subinterval

u function of probability on ith subinterval, equal to $(\bar{p}_{i+1} - p)/h$

x random variable

$\bar{x}_i$ ordinate of ith knot

x_j jth ordered sample value of x

$\lambda_1, \lambda_2, \lambda_3, \lambda_4$ parameters of generalized lambda distribution

Abbreviations:

CDF cumulative distribution function

GLD generalized lambda distribution

PDF probability density function

A prime indicates derivative with respect to the independent variable.

CONTINUOUS RANDOM VARIABLES

Let x be a real, continuous random variable from a probability distribution having the cumulative distribution function (CDF) denoted by $F(x)$. The CDF defines the probability of obtaining a value less than or equal to x:

$$p = F(x) \tag{1}$$

Because $F(x)$ is a probability, it has the following properties:

1. $0 \leq F(x) \leq 1$

2. $F(x)$ is a nondecreasing function of x: $F(x_1) \leq F(x_2)$ if $x_1 < x_2$

3. $F(-\infty) = 0$ and $F(\infty) = 1$

The first derivative of $F(x)$ is the probability density function (PDF) of the distribution and is denoted by

$$f(x) = F'(x) \tag{2}$$

Because $F(x)$ has property 2, the PDF is a nonnegative function. For the normal distribution, the PDF defines the familiar bell shape of the distribution.

Quantile Functions

The inverse function of the CDF is the quantile function $Q(p)$. Hence, given a probability p, the quantile function associates a value of the random variable

$$x = Q(p) \tag{3}$$

with the probability. The major advantage to having a representation of the quantile function is that $Q(p)$ can be used to generate random values from the corresponding distribution. This is accomplished by first generating a uniform random value p and then using equation (3) to transform to a random value x. Equations (1) and (3) can be combined to give

$$F[Q(p)] = p \tag{4}$$

Differentiating both sides of equation (4) with respect to p and substituting equations (2) and (3) yields

$$f(x)\ Q'(p) = 1 \quad \text{or} \quad f(x) = \frac{1}{Q'(p)} \tag{5}$$

By evaluating equations (3) and (5) at various values of p and plotting $f(x)$ versus x, a graphical description of the PDF can be obtained from knowledge of only the quantile function $Q(p)$.

Calculation of Moments

The first few statistical moments of a distribution offer a way of characterizing the distribution with a few parameters; the normal distribution is completely defined by the first two central moments (mean and variance) (ref. 7). For many distributions, the first four moments do not completely define the distribution, but do provide good summary descriptions of the distribution. The moments are usually calculated from the PDF.

The quantile function can also be used to calculate the moments of the distribution. In terms of the PDF, the rth moment about zero is classically defined as (ref. 7)

$$M_r = \int_{-\infty}^{\infty} x^r\ f(x)\ dx \qquad (r = 1,2,3,\ldots) \tag{6}$$

Substituting equations (3) and (5) into equation (6) and changing variables yields

$$M_r = \int_0^1 [Q(p)]^r \, dp \qquad (r = 1,2,3,\ldots) \tag{7}$$

The first moment about zero is the mean; the first four moments can be used to calculate the variance, skewness, and kurtosis ("peakedness") of the distribution (ref. 7).

QUANTILE FUNCTION REPRESENTATIONS

The previous section has shown the usefulness of the quantile function for generating random values. For arbitrary random samples, the quantile function is not usually known; therefore, an approximate representation of the quantile function is needed.

Three representations of the quantile function are presented. The first two representations are function approximation methods; the third is an analytic function, the generalized lambda distribution (GLD). The GLD was chosen as a basis of comparison because it applies to a wide range of skewness and kurtosis values.

B-Spline Approximation

For the B-spline approximation, the interval $0 \leqslant p \leqslant 1$ is first partitioned into k subintervals of width $h = 1/k$. This yields $k + 1$ boundary points (knots) with coordinates $(\bar{p}_i, B_i(\bar{p}_i))$ at the following abscissa values: $0,h,2h,\ldots,1$. The B-spline centered at the knot $\bar{p}_i$ is defined in reference 3 as

$$B_i(p) = \begin{cases} \dfrac{1}{6h^3}\left[h^3 + 3h^2\left(h - |p - \bar{p}_i|\right)\right. & \\ \qquad \left. + 3h\left(h - |p - \bar{p}_i|\right)^2 - 3\left(h - |p - \bar{p}_i|\right)^3\right] & \left(|p - \bar{p}_i| \leqslant h\right) \\ \dfrac{1}{6h^3}\left(2h - |p - \bar{p}_i|\right)^3 & \left(h < |p - \bar{p}_i| < 2h\right) \\ 0 & \text{Otherwise} \end{cases} \tag{8}$$

To fully define the B-spline approximation, two additional B-splines centered at $\bar{p} = -h$ and $\bar{p} = 1 + h$ are included in the model. Figure 1 illustrates this model

for $k = 4$ (which gives $h = 0.25$ and $k + 3 = 7$ B-splines). The quantile function is approximated by

$$Q_B(p) = \sum_{i=1}^{k+3} a_i B_i(p) \tag{9}$$

where $\bar{p}_i = (i - 2)h$ for $i = 1,2,\ldots,k+3$ and a_i are the coefficients to be determined. Figure 1 shows that for any value of p, equation (9) contains only four nonzero terms on the right-hand side; thus equation (9) reduces to

$$Q_B(p) = \sum_{i=s-1}^{s+2} a_i B_i(p) \qquad (\bar{p}_s \leq p < \bar{p}_{s+1}) \tag{10}$$

For example, for $k = 4$ and $p = 0.4$, figure 1 shows that $s = 3$ in equation (10). In this study, k was 4, 8, or 16, which yield models with 7, 11, or 19 B-splines, respectively.

Rational Spline Approximation

The rational spline is a cubic function defined between two knots so that the function and first two derivatives are continuous at the knots. Associated with each knot subinterval is a distinct tension parameter T_i that allows the curvature to range from that of a cubic spline to that of a linear function. The rational spline was originally derived by Späth (ref. 4); the formulation used to represent the quantile function is derived in reference 8 and is defined by

$$Q_R(p) = u\bar{x}_i + H_i\left(\frac{u^3}{T_i t + 1} - u\right)\bar{x}_i'' + t\bar{x}_{i+1} + H_i\left(\frac{t^3}{T_i u + 1} - t\right)\bar{x}_{i+1}'' \qquad (\bar{p}_i \leq p \leq \bar{p}_{i+1}) \tag{11}$$

where

$$H_i = \frac{h^2}{2\left(T_i^2 + 3T_i + 3\right)}$$

and $u = (\bar{p}_{i+1} - p)/h$ and $t = 1 - u$. Equation (11) defines a rational spline as a function of the ordinates ($\bar{x}_i$ and $\bar{x}_{i+1}$) and second derivatives ($\bar{x}_i''$ and $\bar{x}_{i+1}''$) of the function at these knots.

From equation (11), it can be seen that with no tension ($T_i = 0$) the rational spline is a cubic spline; as T_i becomes large, the function tends to the line joining the two knots. The rational spline is well-defined over the entire subinterval as long as $T_i > -1$. For each subinterval, the tension can be adjusted manually or by means of an automated algorithm described in reference 9. That algorithm increments the tension on a given subinterval until the rational spline deviates from the line joining the knots by no more than a user-specified amount.

Generalized Lambda Distribution

The analytic function representation of the quantile function used in this study is the generalized lambda distribution (GLD) developed by Ramberg and Schmeiser (refs. 5 and 6). The GLD quantile function has the form

$$Q_G(p) = \lambda_1 + \frac{p^{\lambda_3} - (1 - p)^{\lambda_4}}{\lambda_2} \qquad (0 \leqslant p \leqslant 1) \tag{12}$$

Specifying the four lambda parameters defines the mean, variance, skewness, and kurtosis of the distribution through the use of equation (7); analytic expressions for the moments are given in reference 6. Symmetric distributions are specified by setting $\lambda_3 = \lambda_4$.

The GLD was chosen for comparison in this study for two reasons. First, the GLD is defined for a wide range of skewness and kurtosis values. Specifically, the kurtosis can range from 1.83 to 22.21 for a symmetric (zero skewness) distribution and from 1.84 to 73.76 if the absolute value of the skewness lies in the interval from 0.001 to 4.646. Note the limitation that the GLD is not defined for all combinations of skewness and kurtosis values in these ranges. Second, tables are given in references 5 and 6 for finding approximate values of λ_3 and λ_4 corresponding to given skewness and kurtosis values. The two remaining lambda parameters can be calculated from the given mean and variance and the tabulated values of λ_3 and λ_4 through the use of equations in reference 6.

FITTING METHODS

The methods used to fit the three quantile representations to a sample are described in this section. The three representations differ sufficiently to require a different method for each representation.

Let x_j $(j = 1,2,\ldots,n)$ be a sample of random values from an unknown distribution ordered from smallest (x_1) to largest (x_n). Since both spline representations are function approximation, or curve fitting, approaches to approximating the quantile function, it is necessary to define the abscissa, p_j $(j = 1,2,\ldots,n)$, corresponding to the random values in the sample. The resulting relationship, $x_j = Q_S(p_j)$, is called the sample quantile function. Standard practice is to assume that the probabilities p_j are uniformly distributed for $0 \leqslant p \leqslant 1$ (ref. 10);

however, there is no standard definition for the probabilities themselves. For this study, they are defined to be

$$p_j = \frac{j - 0.5}{n} \qquad (j = 1,2,\ldots,n) \tag{13}$$

This definition places the p_j symmetrically about the midpoint, 0.5. Since $p_1 > 0$ and $p_n < 1$, equation (13) allows for the possible existence of values in the distribution smaller than x_1 and larger than x_n; however, equation (13) also has the disadvantage that the tail areas to the left of p_1 and to the right of p_n are not defined by the sample. In the terminology of function approximation, random values in the tails must be evaluated by extrapolation.

B-Spline Fit

The B-spline approximation is obtained by a linear least squares fit of $Q_B(p)$ to the sample quantile function. This requires finding the coefficients a_i $(i = 1,2,\ldots,k+3)$ for $k + 3 < n$ which minimize

$$\sum_{j=1}^{n} [Q_S(p_j) - Q_B(p_j)]^2 = \sum_{j=1}^{n} \left[x_j - \sum_{i=s-1}^{s+2} a_i B_i(p_j) \right]^2 \tag{14}$$

where $\bar{p}_s \leq p_j < \bar{p}_{s+1}$. The method of minimizing equation (14) can be found in any text on least squares or regression analysis, such as reference 11.

Rational Spline Fit

The rational spline fit to the sample quantile function is similar to the B-spline fit. For the rational spline approximation, the $2(k + 1)$ unknown parameters in equation (11) are $\bar{x}_i$ and $\bar{x}_i''$ $(i = 1,2,\ldots,k+1)$; however, solving for the abscissa and second derivative of the quantile function at each knot does not ensure that the first derivative is continuous at the knot. This continuity requires that the first derivative of the rational spline on the left side of an interior knot and the first derivative of the rational spline on the right side of the same interior knot both have the same value at that knot.

Therefore, using least squares, the rational spline fit to the sample quantile function consists of minimizing

$$\sum_{j=1}^{n} [Q_S(p_j) - Q_R(p_j)]^2$$

with respect to $\bar{x}_i$ and $\bar{x}''_i$ $(i = 1,2,\ldots,k+1)$, subject to the first derivative of $Q_R(p)$ being continuous at $\bar{p}_i$ $(i = 2,3,\ldots,k)$. The solution to this constrained least squares problem is given in reference 8.

Generalized Lambda Distribution Fit

The parameters of the GLD are determined from sample values of the mean, variance, skewness, and kurtosis. Since these four quantities summarize the characteristics of the sample, details of the distribution of the sample are lost when fitting the GLD. However, the GLD does maintain the location and general shape of the sample distribution.

As indicated previously, two of the GLD parameters (λ_3 and λ_4) can be determined from the sample skewness and kurtosis through the use of tables in reference 6. However, the sample skewness and kurtosis may not occur in the tables in many instances. In this situation, the Levenberg-Marquardt algorithm (ref. 12) is used to find the two parameter values that best fit the GLD skewness and kurtosis equations to the sample values. Since the Levenberg-Marquardt algorithm is a nonlinear technique, starting values of the parameters are required; these can be obtained from the tables in reference 6 by finding tabulated skewness and kurtosis values close to the sample values and using the corresponding parameter values. After the parameters λ_3 and λ_4 are determined, the remaining two parameters can be calculated with the sample mean and variance from equations given in reference 6.

RANDOM NUMBER GENERATION

The major purpose of this research is to develop a method for generating continuous random values which are distributed similarly to a random sample of unknown origin. A quantile function that adequately represents the random sample constitutes a major portion of the method.

Since the domain of the quantile function is the unit interval, the quantile function transforms uniformly distributed random values in the interval $0 \leqslant p \leqslant 1$ into the range of the distribution of interest. This approach to generating random values requires a standard method for generating uniform random values in the unit interval. For the study presented here, the uniformly distributed random values were generated using the computer subroutine GGUBS of reference 13.

RESULTS FOR KNOWN DISTRIBUTIONS

In this section, the B-spline, rational spline, and GLD representations are compared with each other for quantile functions of samples generated from known probability distributions. The use of known distributions allows the approximated quantile functions and PDF's to be compared with the actual quantile function and PDF of the underlying population distribution. Moments of random samples generated by the approximations are also compared.

Fitting the Distributions

The representations are compared in two ways. First, the quantile functions and PDF's are graphically compared with the actual functions. This provides a global

view of the trends of these functions and indicates intervals in which the fits are best. A graphical comparison is also relatively independent of the fitting method, since neither sums of squared residuals nor statistical moments are being compared.

Second, the first four moments of the sample and the approximating distributions are compared. Since the GLD is fit to the sample moments, it appears most favorable in this comparison. However, the first four moments do offer a second way of comparing the global fit of the representations and also provide a simple statistical description of a probability distribution. For the two spline approximations, the moments are calculated by numerically integrating equation (7) for $r = 1,2,3,4$ with Simpson's rule. The moments of the GLD are calculated from equations given in reference 6.

Known distributions.- Three continuous distributions are considered in this study. First, the standard normal distribution, which has a mean of 0 and variance of 1 is chosen since it is symmetric and frequently used in research. Second, a mixture of normal distributions, in which two-thirds of the sample is from a standard normal distribution and one-third is from a normal distribution having mean of 4 and variance of 1, is studied. This distribution was chosen because it has two modes and is asymmetric. Finally, an F-distribution with 10 and 9 degrees of freedom (F(10,9)) is considered since it is unimodal but asymmetric with large skewness and kurtosis. These three distributions are representative of the range of sample distributions encountered in experimental studies. Samples from the three distributions were generated using the computer software described in reference 13. All results presented in this section and the next were obtained on the Control Data Corporation (CDC) CYBER 175 computers at Langley Research Center.

Choosing the number of knots.- Before proceeding with the comparison of the three representations of the distributions, an issue to be settled is the number of knots required to yield a good spline approximation. This question is answered by comparing the B-spline approximations of samples of size 25 and 100 for 7, 11, and 19 knots, since the choice of the number of knots is the only manner in which the B-spline approximations of a sample quantile can differ. Table I presents a comparison of these three B-spline representations for a sample from a standard normal distribution. For a sample of size 25, the 19-knot case is not considered because the number of unknowns is close to the number of data points. The table indicates that increasing the number of knots from 7 to 11 or 19 does not significantly improve the comparison between the calculated moments and the moments of the sample. In fact, in most instances, the accuracy of the moments degrades as the number of knots increases. Comparison of the results for the two samples indicates that sample size is more important than the number of knots; the fit is better for the large sample size. Figures 2 and 3 illustrate this degradation as the number of knots increases from 7 to 19. As the number of knots increases, the quantile approximation tends to become rough because it is fitting the random error in the sample. Figures 4 and 5 illustrate the same phenomenon for the PDF derived from the approximation. Figure 5 also illustrates discontinuities that may occur in the PDF at knot locations. These spikes occur apparently because the PDF is the result of plotting the quantile function versus the reciprocal of the first derivative.

The effect of the number of knots for the B-spline fit was also tested on the mixture of normal distributions and the F-distribution with the result in each case that the best fit was obtained for seven knots. Hence, in all the results to follow, seven knots (at -0.25, 0, 0.25, 0.50, 0.75, 1.00, 1.25) were used for the B-spline fits. Analogously, the rational spline is fit with the five knots at 0, 0.25, 0.50, 0.75, 1.00.

Fits to standard normal distribution.- Results for the standard normal distribution are shown in figures 6 to 9 and table II. For n = 25, figures 6 and 7 indicate that the representations are all skewed because the small sample is skewed. The GLD fit to the PDF (fig. 7) is slightly better than the spline fits. Table II indicates that the GLD exactly fits the moments of the sample, as expected, with the rational spline fit second best overall. For the sample of size 100, figure 8 indicates that all three fits are about equally good. The GLD representation of the PDF shown in figure 9 is excellent. The two spline approximations show the basic shape of the normal PDF with the B-spline fit closer near the peak and the rational spline fit closer in the tails. Note the "shoulders" on the left of the rational spline fit, which result from discontinuities occurring at the knots. For the larger sample size, table II shows the GLD fit to the moments to be exact and the rational spline fit to be closer overall than the B-spline fit. For each of the three generated samples and the experimental data presented later, the rational spline tension was adjusted in order to improve the smoothness of the PDF without appreciably degrading the fit of the quantile approximation.

Fits to mixture of normal distributions.- Results for the mixture of normal distributions are presented in figures 10 to 13 and table III. For the small sample size, all the quantile fits are smooth and fairly good (fig. 10). Although the Levenberg-Marquardt algorithm converged to the GLD fit, the third and fourth lambda parameters are larger than specified in reference 6. Furthermore, the GLD representation of the PDF tends to oversmooth by removing one of the modes (fig. 11), which illustrates the inability of the GLD to accurately represent multimodal distributions. The spline approximations are very similar to each other. Table III again shows the GLD fit to be exact and the rational spline fit to be very close to the sample moments. The B-spline fit is not as close, particularly to the skewness and kurtosis. For the sample of size 100, the fits illustrated in figure 12 are fairly close. The GLD representation of the PDF (fig. 13) is poor, clearly indicating that the GLD is inaccurate for bimodal distributions. In this case, the Levenberg-Marquardt algorithm could not converge on the GLD any closer than shown in table III; the fourth lambda parameter was much larger than allowed by reference 6. This poor fit is due either to the bimodal nature of the distribution or to the larger skewness. The spline approximations suggest the second peak (fig. 13) but still show shoulders at two of the knots. The rational spline fit is slightly better than the B-spline fit.

Fits to F-distribution.- The final example uses samples from the F-distribution with 10 and 9 degrees of freedom. For the sample of size 25, figures 14 and 15 show that the GLD does not fit the sample quantile function, although table IV indicates an exact fit. Examination of the lambda parameters, however, shows them to be larger than the allowable values (ref. 6). The poor quality of this fit may be due to the large kurtosis of this distribution. The spline fits to the quantile function (fig. 14 and table IV) are fairly close, but fits to the PDF are poor (fig. 15), apparently because of the small size of the sample. For the larger sample size (n = 100), the GLD deviates from the sample quantile function in the left tail and is fairly close elsewhere (fig. 16). Figure 17 shows that the GLD representation of the distribution extends too far to the left. Although table IV indicates an exact fit, the lambda parameters are too negative and violate the mathematical theory of the GLD (ref. 6). The difficulty appears to be the large kurtosis or a skewness-kurtosis combination that is not feasible for the GLD. Both spline fits to the sample quantile (fig. 16) are close. The B-spline fit to the PDF (fig. 17) has a modal value that is too large and two pointed shoulders. The rational spline approximation of the PDF is closer with a modal value that is not excessively large. All three PDF

representations are shifted to the right because of the particular sample. The rational spline moments (table IV) are good approximations of the sample moments.

Generating Random Values

In this section, the three representations are examined by comparing the moments of random samples generated by each representation. Using the parameters estimated from the larger samples (n = 100) in the previous section, 1000 samples each of sizes 25 and 100 were generated. The first four mean moments of the 1000 samples were calculated and compared in order to determine the tendency of each method to reproduce the moments of the original sample. The procedure is outlined as follows:

1. A uniform random number generator (ref. 13) was used to generate 1000 "seeds."

2. Each seed was used by the same generator to produce a sample of n uniform random numbers.

3. The n uniform random numbers were transformed by one of the three quantile representations to the interval of interest.

4. The first four moments (mean, variance, skewness, and kurtosis) of the transformed sample were calculated.

5. The mean values of the 1000 sets of moments were calculated.

For the results presented here, the two spline methods required at least one-third less computer time than the GLD to generate a random number on a CYBER 175 computer.

Table V presents the results of the quantile representations of the standard normal distribution. For both sample sizes, all the mean moments of the GLD are closest to the sample moments. Except for both the variance and the mean in the small sample case, the rational spline moments are closer than those of the B-spline.

On the other hand, the superiority of one particular method of generating samples from the mixture of normal distributions (table VI) is less obvious. Overall the B-spline moments appear to be slightly closer than the moments of the other methods. Since the mean and variance of a sample can be adjusted to any desired values by a linear transformation, the most critical moments for comparison are the skewness and kurtosis. With the exception of the kurtosis for the small sample, the B-spline and rational spline methods produce samples having mean skewness and kurtosis values closer to the sample moments than does the GLD. Therefore, on the basis of the comparison of the two highest moments, the two spline methods are superior to the GLD for this example. Much of this superiority can be attributed to the inability of the GLD to fit a bimodal distribution as noted previously.

Although the estimated values of the GLD parameters for the sample of size 100 from the F-distribution are more negative than prescribed in reference 6, random samples were generated with the GLD for comparison purposes. Table VII shows that the mean skewness and kurtosis of the GLD samples are especially in error. However, the mean moments from the two spline approximations for the larger sample size are very good. For the smaller sample size, the B-spline moments are slightly worse than for the larger sample, and the rational spline moments have decreased considerably.

SURFACE WIND EXAMPLE

This example compares the three representations when applied to experimental data. The data consist of the two horizontal components (east-west and north-south) of surface wind velocities measured in meters per second. For purposes of illustration, the wind components are treated as two independent samples of wind velocities. Before analyzing the data, each sample was ordered from smallest to largest in order to associate the sample values with the corresponding probabilities (eq. (13)) which define the sample quantile function.

Table VIII and figures 18 to 21 compare the fit of the three representations to the samples. Figures 18 and 20 indicate that the three representations generally fit the sample quantile function quite well. The only exception appears to be the GLD fit to the north-south component (fig. 20) for $p < 0.3$, which diverges somewhat. Table VIII shows the GLD fits to the moments to be exact. The rational spline fits are somewhat worse than the GLD fits and generally better than the B-spline fits.

In figures 19 and 21, the PDF representations are plotted along with the sample histogram expressed as a relative frequency. The histograms were constructed by subdividing the range of the east-west and north-south samples into 11 and 14 subintervals, respectively. Hence, the histogram is a rough approximation of the sample PDF. The peaks of the three representations are located near the histogram peaks. In each case, the unimodal shape of the GLD PDF tends to smooth the smaller peaks in the histogram. The spline PDF's again have shoulders, which may be due to discontinuities in the plots or, because of their location, may actually reflect better fits to the sample distribution. Clearly, the rational spline PDF appears to be better behaved than the B-spline PDF. Because of the crude nature of these histograms and the behavior of the spline PDF's on the previously generated samples, it is not clear whether the GLD or the rational spline PDF is more representative of the sample PDF.

Samples of the two wind components were generated using the parameters for the three representations obtained in fitting the two samples. Table IX presents the mean moments of 1000 generated samples of size 100 for each representation. With the exception of the kurtosis of the distribution of north-south components, the GLD moments are closer to the original sample moments than are the moments of the two spline approximations. For this particular application, the two spline approximations generate samples that are about equally good. Apparently, the GLD is superior in this application because the skewness and kurtosis values are not excessively large and the combinations lie in the range of values tabulated in reference 6.

CONCLUDING REMARKS

The research presented here has examined the feasibility of using B-spline and rational spline functions to approximate the quantile function of a random sample. The spline approximations provide means not only for representing the distribution of the sample but also for generating random values having statistical properties like those of the original sample.

Included in this study was a comparison of the spline approximations with the generalized lambda distribution (GLD), an analytic representation of the distribution. Comparison of these representations of samples from three standard distributions indicate that the GLD provides a better fit if the distribution is nearly symmetric and unimodal or the combination of skewness and kurtosis are reasonably

close to values tabulated for the GLD. In contrast, since the spline functions can be fit to essentially any set of data, the spline approximations are better for samples from skewed or multimodal distributions. Furthermore, the spline approximations are more easily fit by linear least squares, whereas the GLD requires the use of a nonlinear method which may not converge. However the spline approximations of the probability density function are generally inadequate, since they often exhibit false peaks at values corresponding to spline knots.

As methods of generating random values, the two spline approximations were at least one-third faster than the GLD. The ability of each method to produce random values having statistical properties similar to the original sample is directly related to how well the approximation fits the sample.

The rational spline is recommended as an alternative to the GLD for several reasons. First, it can be fit to an arbitrary random sample. The fitting method for the rational spline uses each individual sample value, rather than summary statistics. The distinct tension parameter associated with each knot interval provides a flexibility not available with other methods; with all the tension parameters set to zero, the rational spline reduces to a cubic spline. Finally, the rational spline can generate samples faster than the GLD.

Langley Research Center
National Aeronautics and Space Administration
Hampton, VA 23665
October 16, 1984

REFERENCES

1. Schmeiser, Bruce: Methods for Modelling and Generating Probabilistic Components in Digital Computer Simulation When the Standard Distributions Are Not Adequate: A Survey. 1977 Winter Simulation Conference, Volume 1, Harold Joseph Highland, Robert G. Sargent, and J. William Schmidt, eds., Nat. Bur. Stand., 1977, pp. 50-57.

2. Tapia, Richard A.; and Thompson, James R.: Nonparametric Probability Density Estimation. Johns Hopkins Univ. Press, c.1978.

3. Prenter, P. M.: Splines and Variational Methods. John Wiley & Sons, Inc., c.1975.

4. Späth, Helmuth (W. D. Hoskins and H. W. Sager, transl.): Spline Algorithms for Curves and Surfaces. Utilitas Mathematica Pub. Inc., 1974.

5. Ramberg, John S.; and Schmeiser, Bruce W.: An Approximate Method for Generating Symmetric Random Variables. Commun. ACM, vol. 15, no. 11, Nov. 1972, pp. 987-990.

6. Ramberg, John S.; and Schmeiser, Bruce W.: An Approximate Method for Generating Asymmetric Random Variables. Commun. ACM, vol. 17, no. 2, Feb. 1974, pp. 78-82.

7. Kendall, Maurice; and Stuart, Alan: The Advanced Theory of Statistics. Volume 1 - Distribution Theory, Fourth ed. Macmillan Pub. Co., Inc., c.1977.

8. Schiess, James R.; and Kerr, Patricia A.: Rational Spline Approximation With Automatic Tension Adjustment. NASA TP-2366, 1984.

9. Frost, Charles E.; and Kinzel, Gary L.: An Automatic Adjustment Procedure for Rational Splines. Comput. & Graphics, vol. 6, no. 4, 1982, pp. 171-176.

10. Parzen, Emanuel: Data Modeling Using Quantile and Density-Quantile Functions. Some Recent Advances in Statistics, J. Tiago de Oliveira and Benjamin Epstein, eds., Academic Press Inc., 1982, pp. 23-52.

11. Draper, N. R.; and Smith, H.: Applied Regression Analysis, Second ed. John Wiley & Sons, Inc., c.1981.

12. Brown, K. M.; and Dennis, J. E.: Derivative Free Analogues of the Levenberg-Marquardt and Gauss Algorithms for Nonlinear Least Squares Approximation. Numer. Math., vol. 18, no. 4, 1972, pp. 289-297.

13. IMSL Library Reference Manual - Volume 1, ed. 9. IMSL LIB-009 (Rev.), IMSL, Inc., c.1982.

TABLE I.- COMPARISON OF MOMENTS FOR B-SPLINE FITS TO STANDARD NORMAL RANDOM SAMPLES

	Mean	Variance	Skewness	Kurtosis
Population	0	1.0	0	3.0
Sample size n = 25				
Sample	-0.436	0.604	0.273	2.221
7 B-splines	-.498	.603	.183	2.381
11 B-splines	-.498	.607	.187	2.318
Sample size n = 100				
Sample	0.012	0.991	0.024	2.848
7 B-splines	-.009	.989	-.017	2.784
11 B-splines	-.012	1.002	-.076	2.968
19 B-splines	-.013	.999	-.046	2.969

TABLE II.- COMPARISON OF MOMENTS CALCULATED FROM THE THREE FITS TO STANDARD NORMAL RANDOM SAMPLES

	Mean	Variance	Skewness	Kurtosis
Population	0	1.0	0	3.0
Sample size n = 25				
Sample	-0.436	0.604	0.273	2.221
7 B-splines	-.498	.603	.183	2.381
Rational spline[a]	-.438	.606	.248	2.246
GLD	-.436	.604	.273	2.221
Sample size n = 100				
Sample	0.012	.991	0.024	2.848
7 B-splines	-.009	.989	-.017	2.784
Rational spline[a]	.013	.998	.035	2.895
GLD	.012	.991	.024	2.848

[a]5 knots.

TABLE III.- COMPARISON OF MOMENTS CALCULATED FROM THE THREE FITS TO RANDOM SAMPLES FROM MIXTURE OF NORMAL DISTRIBUTIONS

	Mean	Variance	Skewness	Kurtosis
Population	1.333	4.556	0.488	2.086
Sample size n = 25				
Sample	1.182	3.960	0.180	2.050
7 B-splines	1.026	3.976	.103	2.258
Rational spline[a]	1.182	3.955	.172	2.063
GLD	1.182	3.960	.180	2.050
Sample size n = 100				
Sample	1.161	4.548	0.677	2.395
7 B-splines	1.122	4.457	.685	2.472
Rational spline[a]	1.163	4.562	.683	2.474
GLD	1.161	4.548	.583	2.445

[a]5 knots.

TABLE IV.- COMPARISON OF MOMENTS CALCULATED FROM THE THREE FITS TO RANDOM SAMPLES FROM F(10,9)-DISTRIBUTION

	Mean	Variance	Skewness	Kurtosis
Population	1.286	1.124	5.203	127.32
Sample size n = 25				
Sample	1.355	1.563	3.178	13.934
7 B-splines	1.246	1.407	3.615	16.580
Rational spline[a]	1.374	1.897	4.010	23.453
GLD	1.355	1.563	3.178	13.935
Sample size n = 100				
Sample	1.287	1.411	2.544	9.866
7 B-splines	1.287	1.639	3.189	13.868
Rational spline[a]	1.287	1.407	2.488	9.470
GLD	1.287	1.411	2.544	9.866

[a]5 knots

TABLE V.- COMPARISON OF MEAN MOMENTS FOR 1000 SAMPLES SIMULATED FROM THE THREE FITS TO A SAMPLE OF SIZE 100 FROM THE STANDARD NORMAL DISTRIBUTION

	Mean	Variance	Skewness	Kurtosis
Sample (n = 100)	0.012	0.991	0.024	2.848
Sample size n = 25				
Simulations:				
7 B-splines	-0.016	1.144	-0.137	2.341
Rational spline[a]	-.020	1.161	.014	2.527
GLD	.019	.942	.031	2.711
Sample size n = 100				
Simulations:				
7 B-splines	-0.036	1.022	-0.101	2.511
Rational spline[a]	-.028	1.036	-.071	2.710
GLD	.012	.987	.014	2.834

[a]5 knots.

TABLE VI.- COMPARISON OF MEAN MOMENTS FOR 1000 SAMPLES SIMULATED FROM THE THREE FITS TO A SAMPLE OF SIZE 100 FROM THE MIXTURE OF NORMAL DISTRIBUTIONS

	Mean	Variance	Skewness	Kurtosis
Sample (n = 100)	1.161	4.548	0.677	2.395
Sample size n = 25				
Simulations:				
7 B-splines	0.913	4.553	0.654	2.462
Rational spline[a]	0.932	4.645	.610	2.468
GLD	1.148	4.340	.548	2.454
Sample size n = 100				
Simulations:				
7 B-splines	1.106	4.522	0.656	2.387
Rational spline[a]	1.103	4.553	.650	2.426
GLD	1.163	4.498	.568	2.443

[a]5 knots.

TABLE VII.- COMPARISON OF MEAN MOMENTS FOR 1000 SAMPLES SIMULATED FROM THE THREE FITS TO A SAMPLE OF SIZE 100 FROM THE F(10,9)-DISTRIBUTION

	Mean	Variance	Skewness	Kurtosis
Sample (n = 100)	1.287	1.411	2.544	9.866
Sample size n = 25				
Simulations:				
7 B-splines	1.243	1.291	2.408	9.247
Rational spline[a]	1.202	1.280	1.922	6.860
GLD	1.292	1.386	1.464	5.125
Sample size n = 100				
Simulations:				
7 B-splines	1.264	1.366	2.429	9.345
Rational spline[a]	1.264	1.367	2.439	9.446
GLD	1.290	1.403	1.860	7.606

[a]5 knots.

TABLE VIII.- COMPARISON OF MOMENTS CALCULATED FROM THE THREE FITS TO EXPERIMENTAL WIND SAMPLES

	Mean	Variance	Skewness	Kurtosis
East-west wind component				
Sample (n = 100)	-1.101	4.976	-0.366	2.515
7 B-splines	-1.101	4.956	-.348	2.426
Rational spline[a]	-1.101	4.988	-.362	2.539
GLD	-1.101	4.976	-.366	2.515
North-south wind component				
Sample (n = 100)	-0.907	23.485	-0.608	4.042
7 B-splines	-.908	23.290	-.648	4.000
Rational spline[a]	-.904	23.516	-.584	4.189
GLD	-.907	23.485	-.608	4.042

[a]5 knots.

TABLE IX.- COMPARISON OF MEAN MOMENTS FOR 1000 SAMPLES OF SIZE 100 SIMULATED FROM THE THREE FITS TO WIND SAMPLES

	Mean	Variance	Skewness	Kurtosis
East-west wind component				
Sample (n = 100)	-1.101	4.976	-0.366	2.515
7 B-splines	-1.194	5.241	-.379	2.436
Rational spline[a]	-1.200	5.323	-.426	2.581
GLD	-1.112	4.961	-.359	2.504
North-south wind component				
Sample (n = 100)	-0.907	23.485	-0.608	4.042
7 B-splines	-1.157	25.854	-.756	4.013
Rational spline[a]	-1.157	25.975	-.733	4.105
GLD	-.920	23.157	-.572	3.858

[a]5 knots.

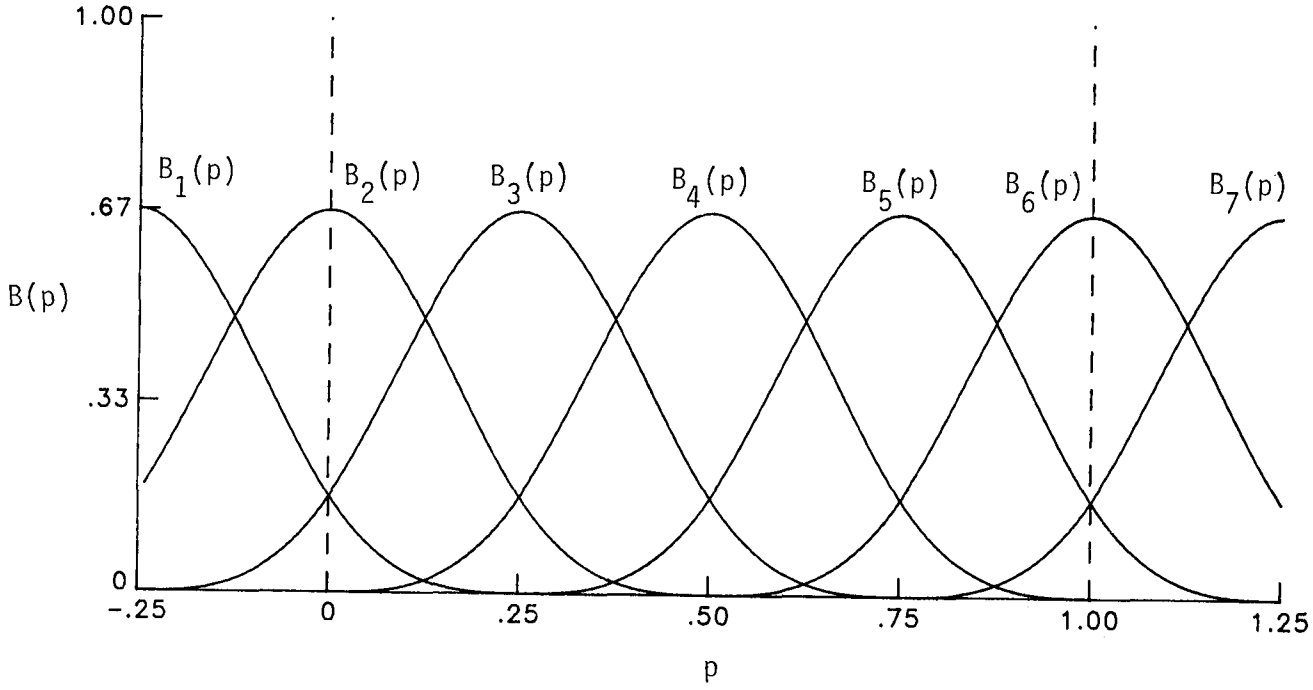

Figure 1.- Basis splines for approximation with seven B-splines.

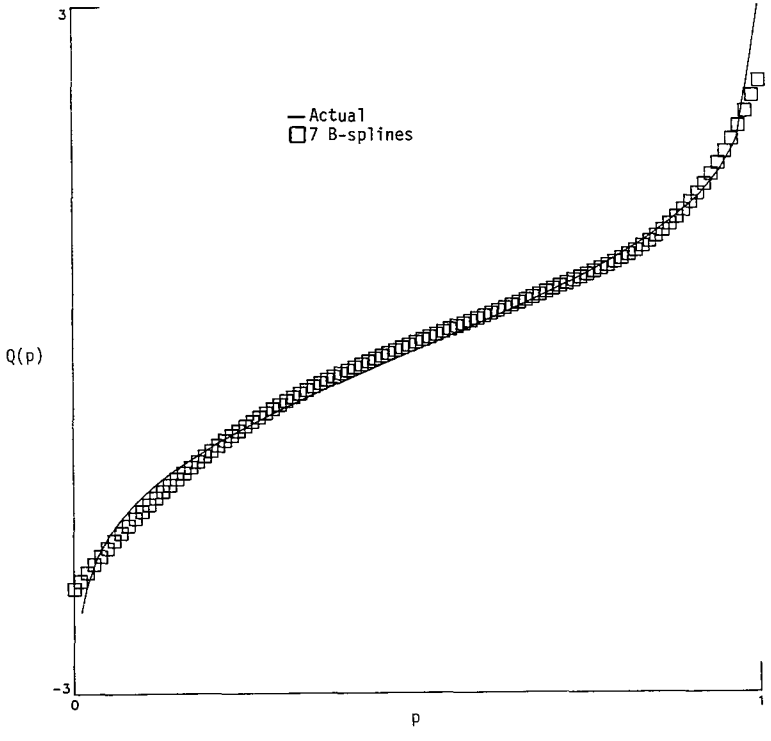

Figure 2.- Comparison of standard normal quantile function with quantile function from approximation (n = 100) with seven B-splines.

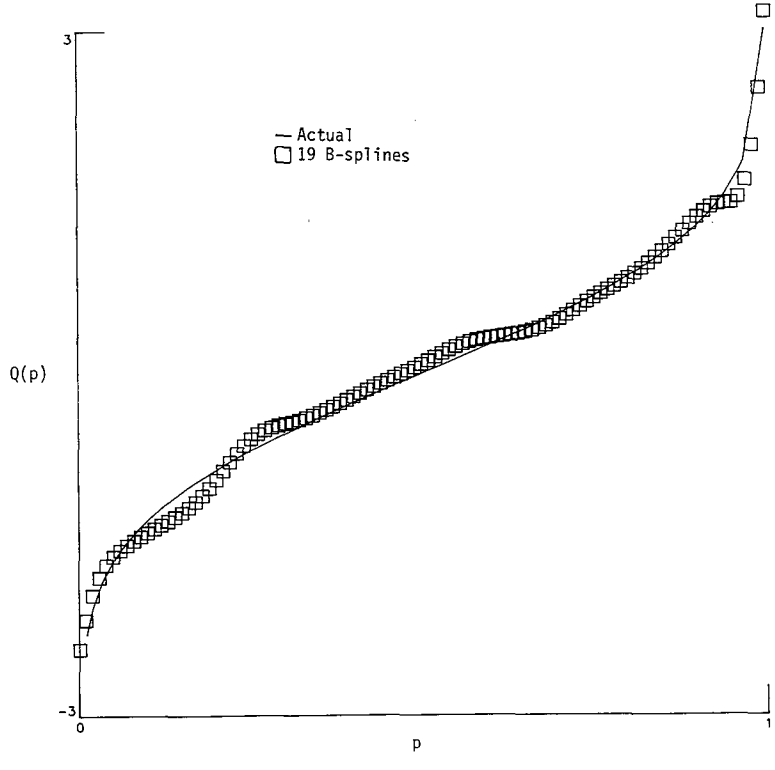

Figure 3.- Comparison of standard normal quantile function with quantile function from approximation (n = 100) with 19 B-splines.

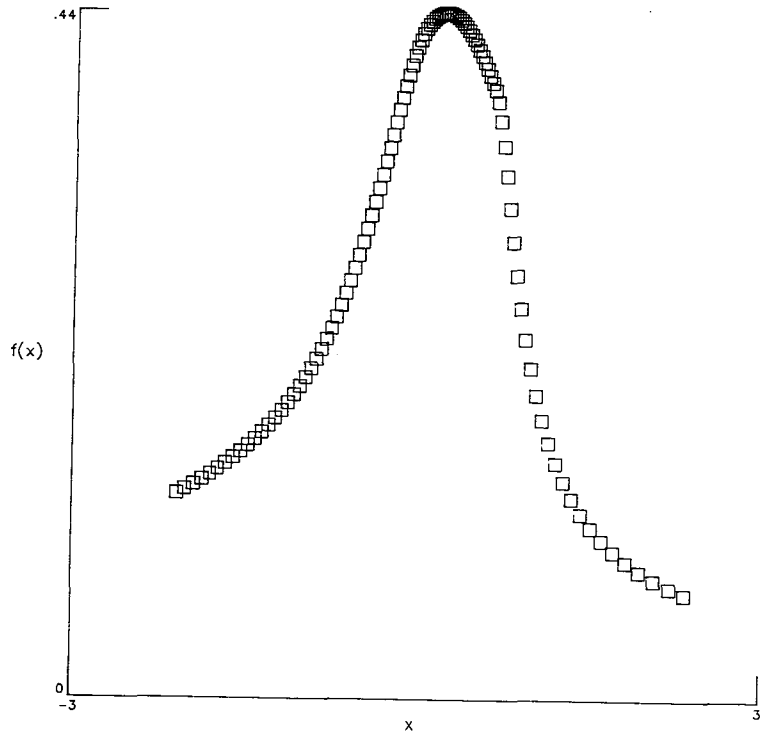

Figure 4.- Standard normal probability density function derived from approximation (n = 100) with seven B-splines.

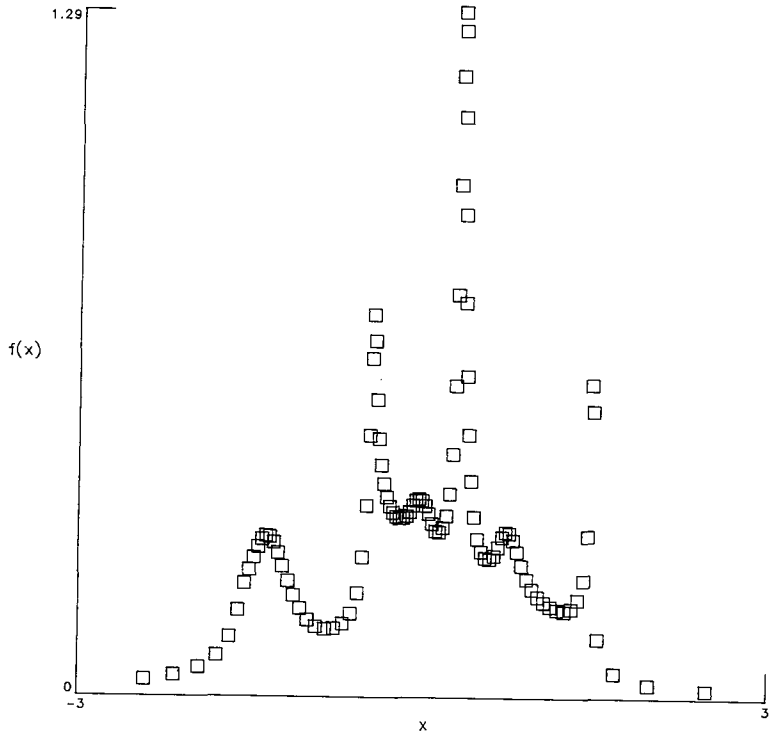

Figure 5.- Standard normal probability density function derived from approximation (n = 100) with 19 B-splines.

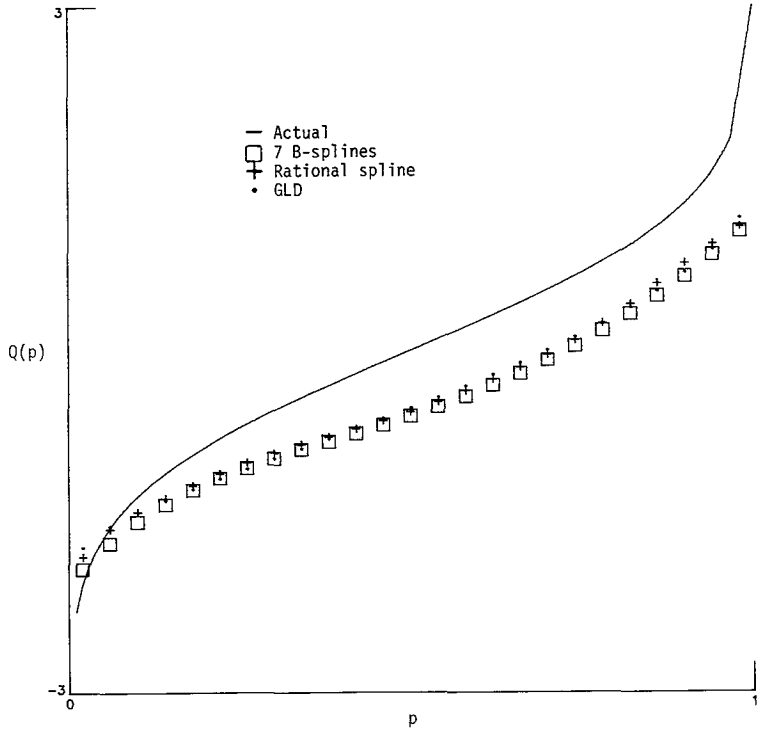

Figure 6.- Comparison of standard normal quantile function with the three quantile approximations. n = 25.

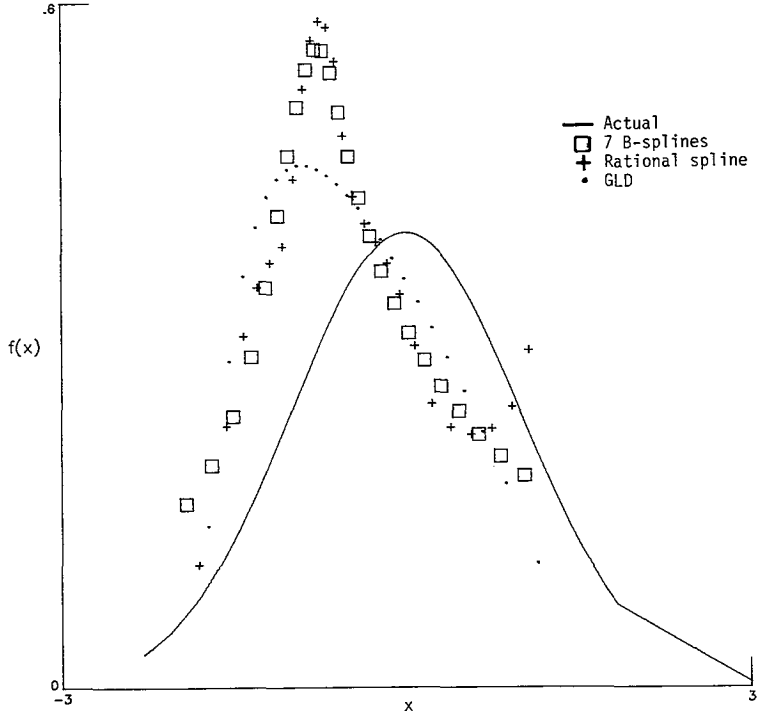

Figure 7.- Comparison of standard normal probability density function with probability density functions derived from the three quantile approximations. n = 25.

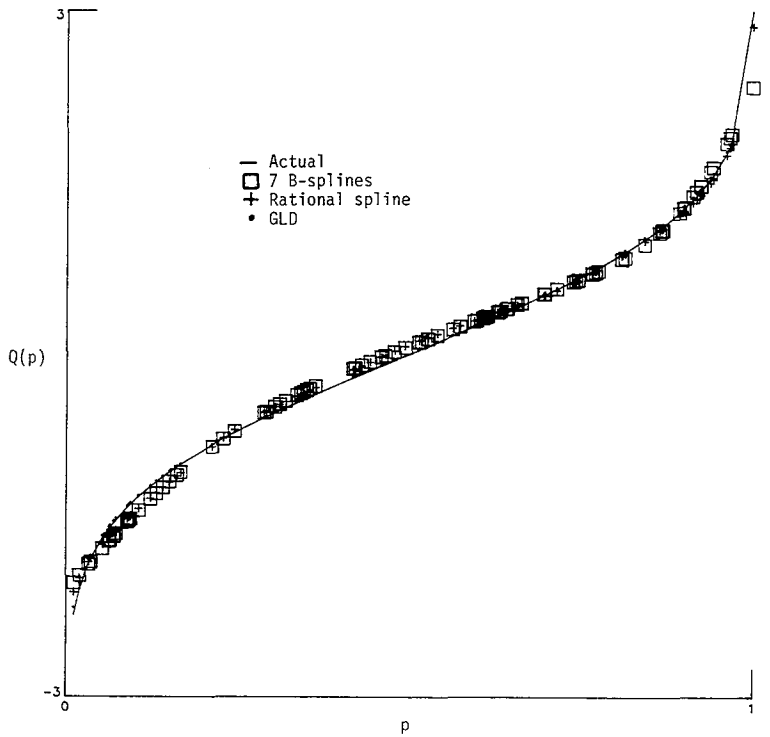

Figure 8.- Comparison of standard normal quantile function with the three quantile approximations. n = 100.

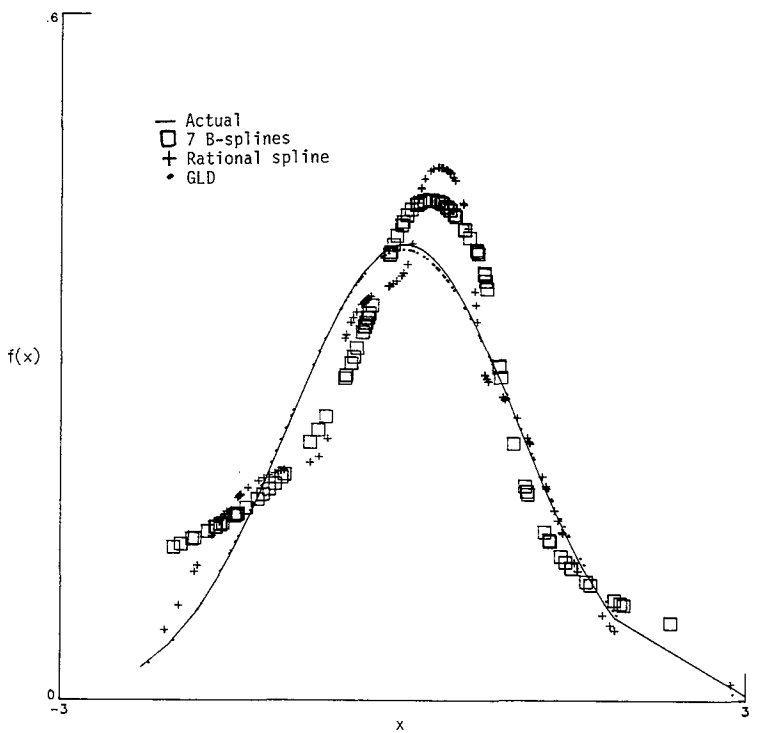

Figure 9.- Comparison of standard normal probability density function with probability density functions derived from the three quantile approximations. n = 100.

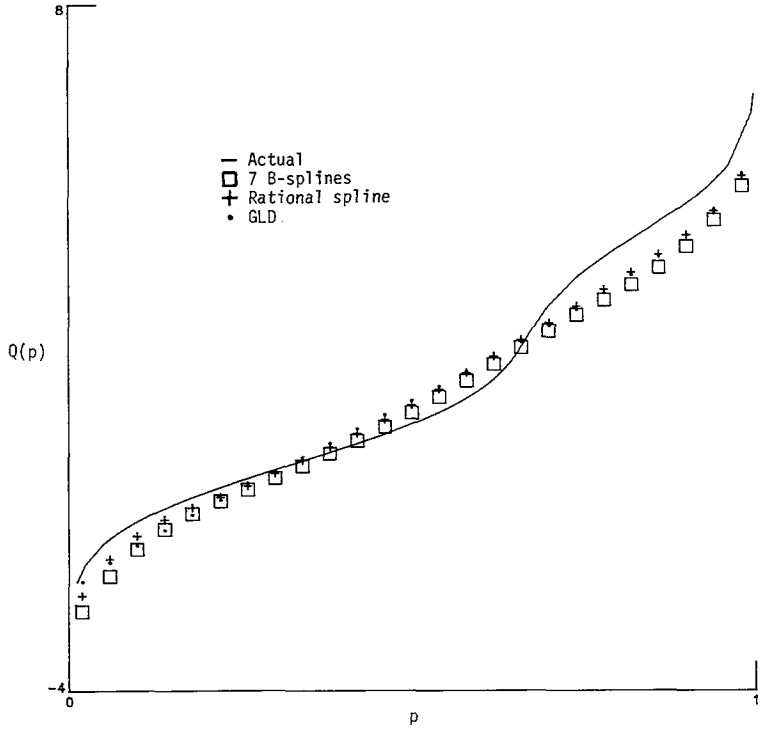

Figure 10.- Comparison of quantile function for mixture of normal distributions with the three quantile approximations. n = 25.

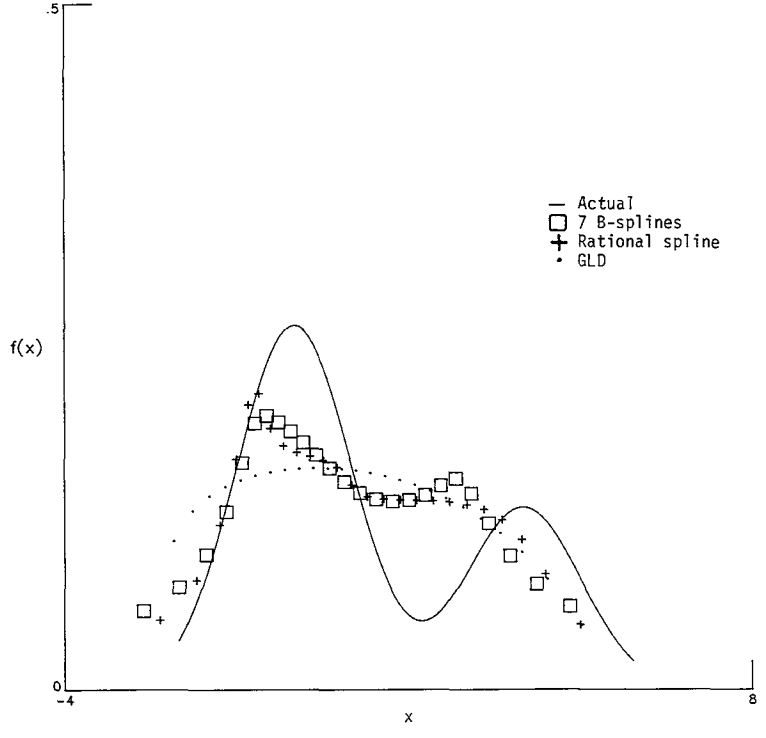

Figure 11.- Comparison of probability density function for mixture of normal distributions with probability density functions derived from the three quantile approximations. n = 25.

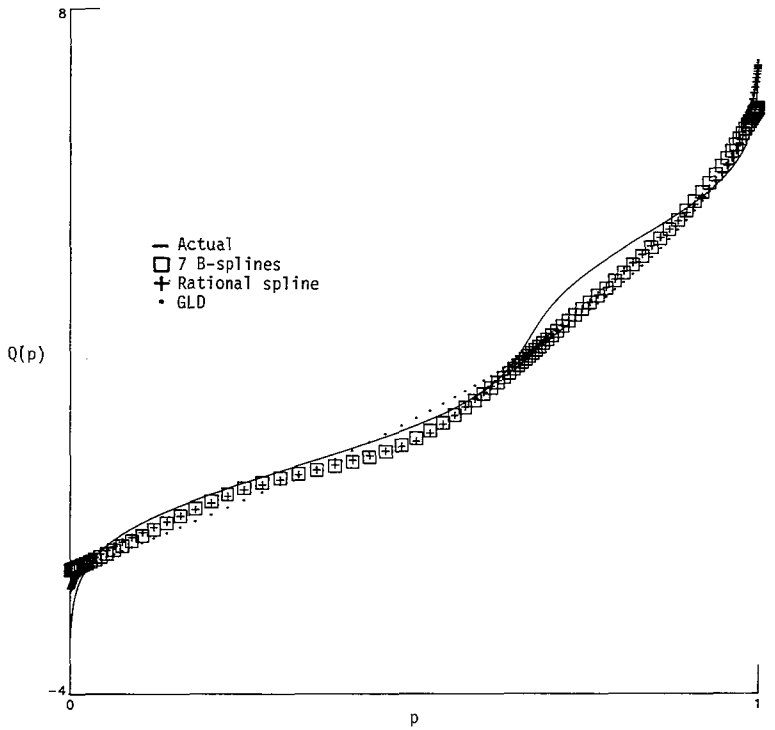

Figure 12.- Comparison of quantile function for mixture of normal distributions with the three quantile approximations. n = 100.

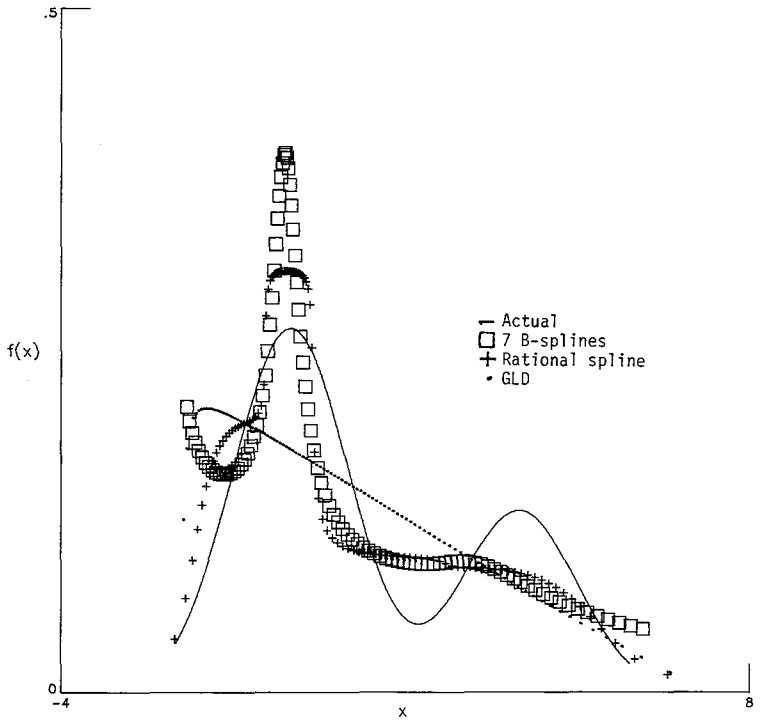

Figure 13.- Comparison of probability density function for mixture of normal distributions with probability density functions derived from the three quantile approximations. n = 100.

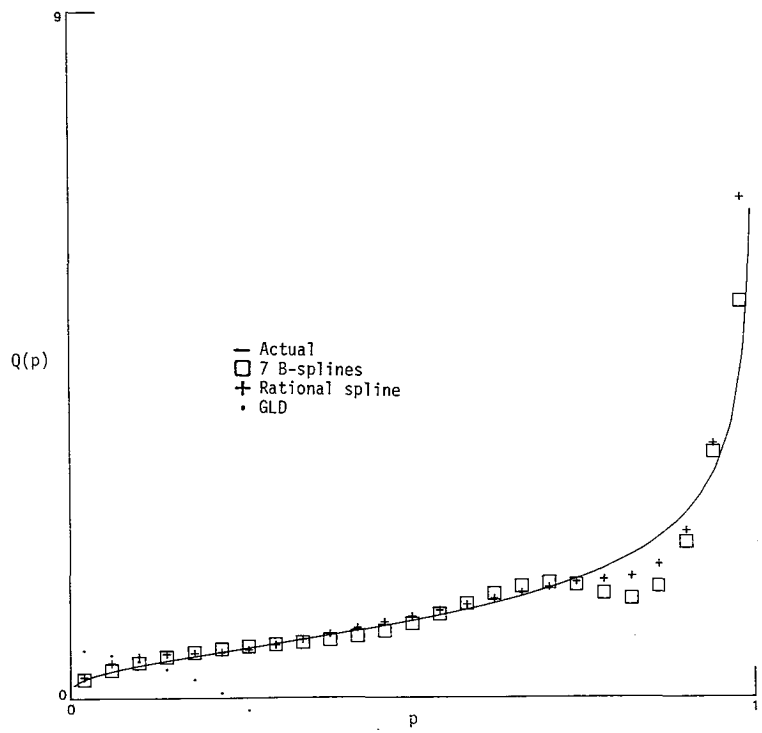

Figure 14.- Comparison of F(10,9) quantile function with the three quantile approximations. n = 25.

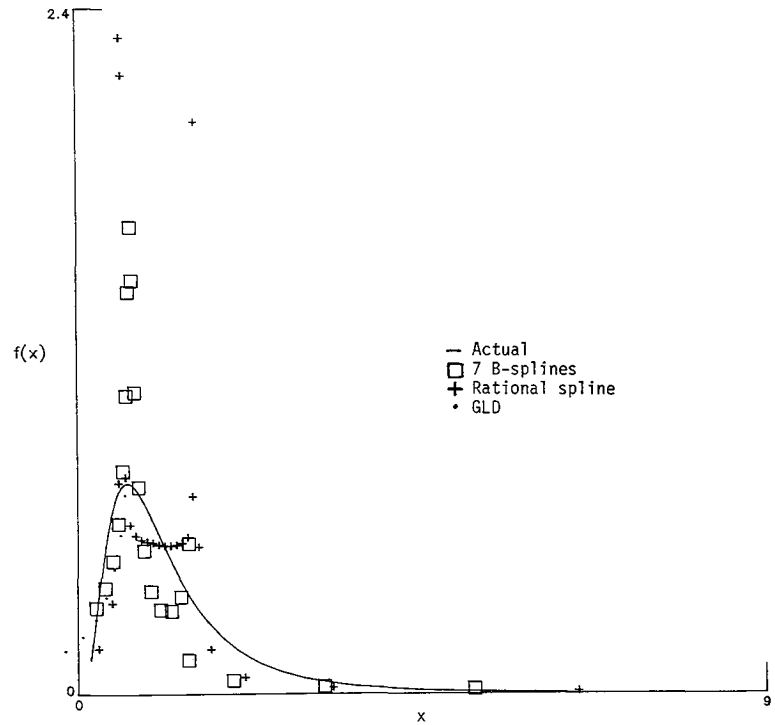

Figure 15.- Comparison of F(10,9) probability density function with probability density functions derived from the three quantile approximations. n = 25.

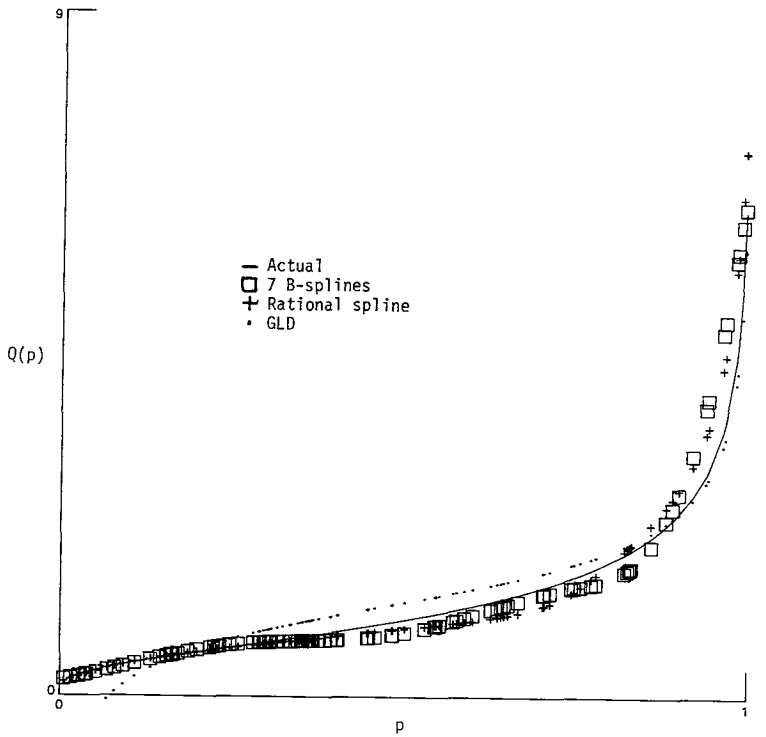

Figure 16.- Comparison of F(10,9) quantile function with the three quantile approximations. n = 100.

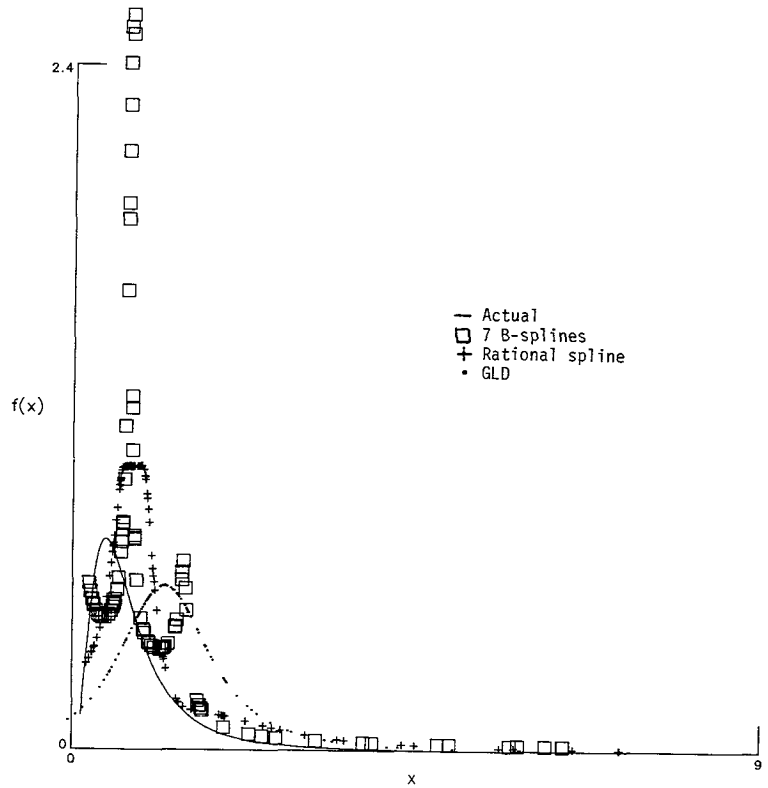

Figure 17.- Comparison of F(10,9) probability density function with probability density functions derived from the three quantile approximations. n = 100.

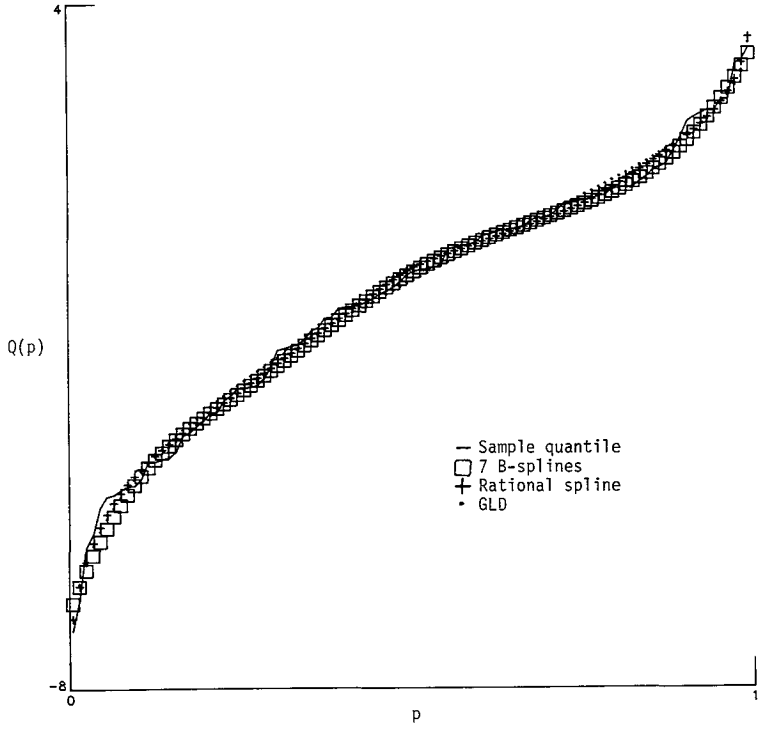

Figure 18.- Comparison of sample quantile function with the three quantile approximations for the east-west wind component.

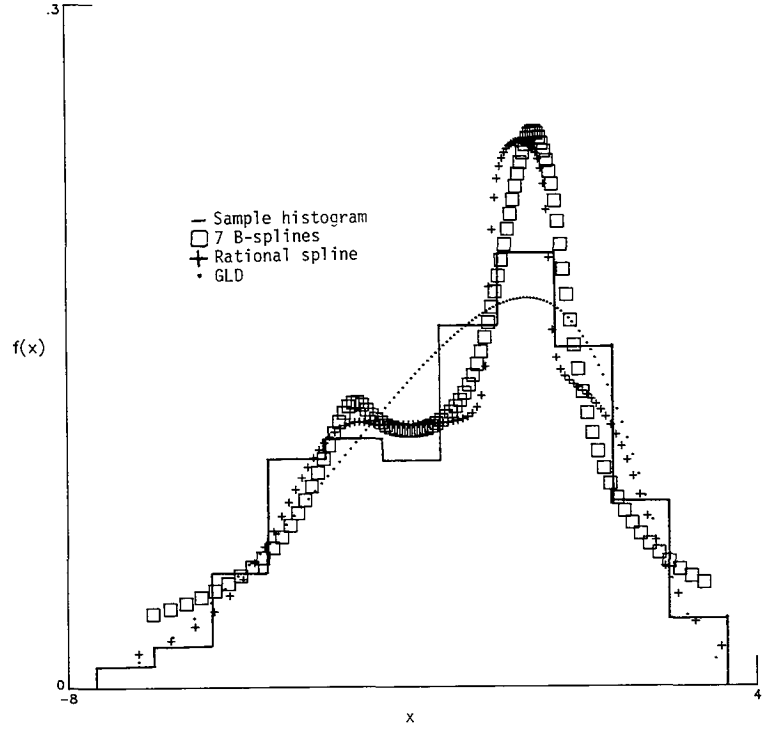

Figure 19.- Comparison of sample histogram with probability density functions derived from the three quantile approximations for the east-west wind component.

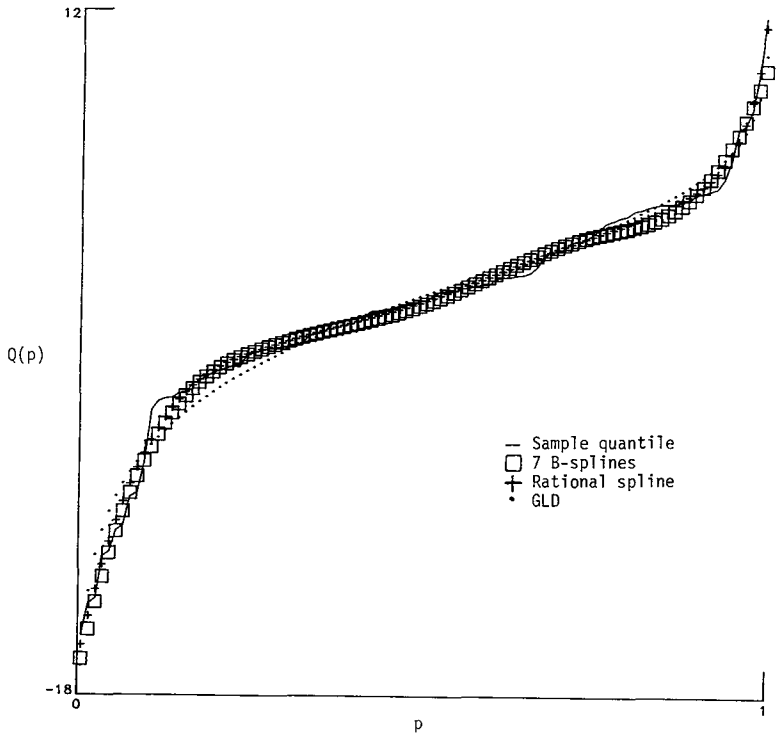

Figure 20.- Comparison of sample quantile function and the three quantile approximations for the north-south wind component.

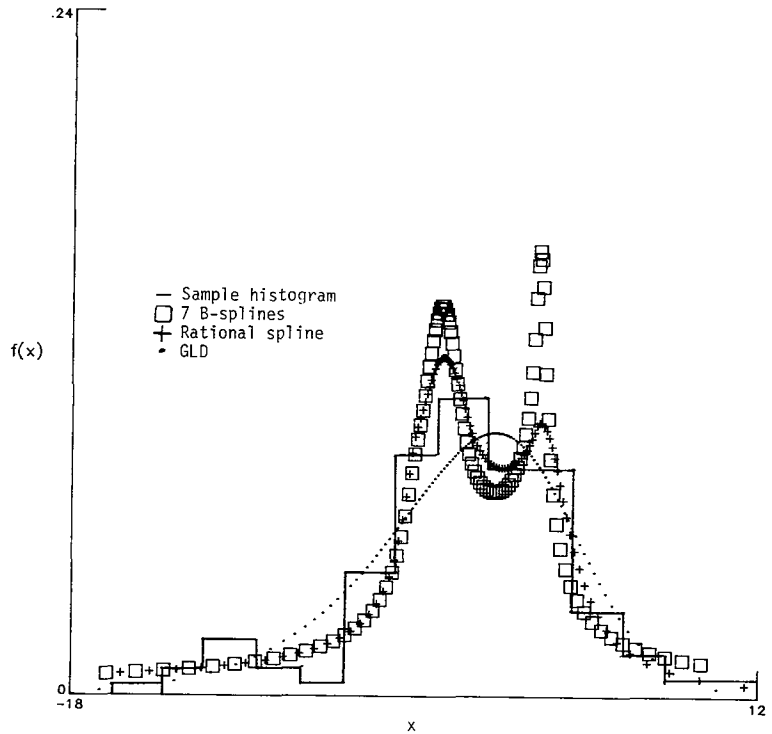

Figure 21.- Comparison of sample histogram and probability density functions derived from the three quantile approximations for the north-south wind component.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA TP-2389		

4. Title and Subtitle	5. Report Date
SPLINE METHODS FOR APPROXIMATING QUANTILE FUNCTIONS AND GENERATING RANDOM SAMPLES	January 1985
	6. Performing Organization Code: 505-31-83-02
7. Author(s): James R. Schiess and Christine G. Matthews	8. Performing Organization Report No.: L-15850
	10. Work Unit No.
9. Performing Organization Name and Address: NASA Langley Research Center, Hampton, VA 23665	11. Contract or Grant No.
	13. Type of Report and Period Covered: Technical Paper
12. Sponsoring Agency Name and Address: National Aeronautics and Space Administration, Washington, DC 20546	14. Army Project No.

15. Supplementary Notes

James R. Schiess: Langley Research Center, Hampton, Virginia.
Christine G. Matthews: Computer Sciences Corporation, Hampton, Virginia.

16. Abstract

Two cubic spline formulations are presented for representing the quantile function (inverse cumulative distribution function) of a random sample of data. Both B-spline and rational spline approximations are compared with analytic representations of the quantile function. It is also shown how these representations can be used to generate random samples for use in simulation studies. Comparisons are made on samples generated from known distributions and a sample of experimental data. The spline representations are shown to be more accurate for multimodal and skewed samples and to require much less time to generate samples than the analytic representation.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Random numbers Simulation Probability distribution	Unclassified - Unlimited Subject Category 65

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	32	A03

*For sale by the National Technical Information Service, Springfield, Virginia 22161

NASA-Langley, 1985

National Aeronautics and
Space Administration

Washington, D.C.
20546

Official Business
Penalty for Private Use, $300

THIRD-CLASS BULK RATE

Postage and Fees Paid
National Aeronautics and
Space Administration
NASA-451

U.S.MAIL

NASA

POSTMASTER: If Undeliverable (Section 158
Postal Manual) Do Not Return