FACULTY PAPER SERIES

DEPARTMENT OF AGRICULTURAL ECONOMICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS

FP 94-2 October 1993

TUTORIAL ON DENSITY FUNCTION ESTIMATION AND USE

James W. Mjelde, David P. Anderson, Keith Coble,
Boujemmaa Mouflih, Joe L. Outlaw, James W. Richardson,
Jeffrey R. Stokes, Vasanthara Sundarapather

Authors are Associate Professor, Research Associate, Department of Agricultural Economics, Texas A&M University; Agricultural Economist, USDA-ERS; Former Graduate Research Assistant, Assistant Research Scientist, Professor, Graduate Research Assistant, and Graduate Research Assistant, Department of Agricultural Economics, Texas A&M University.

TUTORIAL ON DENSITY FUNCTION ESTIMATION AND USE

TABLE OF CONTENTS

INTRODUCTION

One of the most important considerations when developing an economic model is the specification of stochastic events within the model. Unfortunately, this component is usually left to an *ad hoc* procedure with little discussion of the appropriateness of the density function used. Taylor (1981) states that density functions " . . . are seldom analyzed with care and attention commensurate with their impact on decision rules" (p. 1). These functions are the driving force in stochastic simulation and optimization models. In addition, Fryer states that "density estimation is possibly the most important topic in applied statistics ..." (p. 335). He further states, that unless the density function is known, characteristics of the function must be inferred from the sample, before predictions can be made. Authors such as Mjelde, Dixon, and Sonka and Rodriguez and Taylor have demonstrated the importance of including probabilities of random events within agricultural models.

The objective of this paper is to provide a tutorial on the estimation of both univariate and multivariate probability density functions (pdfs) and cumulative density functions (cdfs). A balance of theoretical and practical estimation issues is presented. Univariate density functions provide the basis for the more complicated multivariate estimation procedures. A working knowledge of univariate distributional issues is, therefore, essential to researchers interested in stochastic models. Multivariate density functions represent the more realistic case in agricultural research, but this increase in realism comes with additional estimation costs.

This tutorial is designed for the applied researcher who is familiar with basic statistical techniques. Several different procedures for estimating both univariate and multivariate density functions, covering a wide range of estimator types are illustrated. This tutorial is not meant to be all inclusive, but rather it is designed to present a range of density function estimation issues and techniques. Most of the issues and techniques presented have been discussed by previous authors, but they have not been combined into a single document that is accessible to the average applied researcher (nonstatistician). Examples are contained throughout this tutorial which illustrate and compare the different techniques. These examples are included as a guide to estimation of density functions. It must be stressed that this document should only be the starting point for density function estimation and not the ending point.

Seven major sections comprise this tutorial. The first section provides a general introduction to the tutorial. General topics concerning density function estimation are the focus of the second section. Estimation of univariate procedures comprise the third. Several techniques to compare estimated pdfs and cdfs are discussed in the fourth sections. As noted earlier, not all univariate

techniques are discussed, but rather a range of different techniques are presented. An example of using each technique is presented immediately following the discussion of a particular procedure. Each example estimates a univariate pdf for simulated yearly aquifer recharge data. This allows the procedures to be compared. Multivariate density estimation is the topic of the fifth section. Again several techniques and examples are discussed. This discussion builds on the univariate presentation. Two common procedures to incorporate probability distributions into economic models are the subject of the sixth section. Several of the distributions are compared using the aquifer examples. Finally in the seventh section, some concluding remarks are presented.

Each major section contains several subsections. Each subsection is written to be as self-contained as possible. The reader can, therefore, read the subsection(s) of interest. It is suggested that readers who are unfamiliar with density estimation understand the first two sections before moving to the multivariate section. Readers familiar with univariate density estimation can proceed to the multivariate and application sections with little difficulty. It is suggested that all readers review the first section on general estimation.

GENERAL ESTIMATION CONSIDERATIONS

A brief overview concerning density estimation is presented in this section. Topics such as unique aspects of density estimation, classification of density functions, pdf *versus* cdf estimation, and the necessity of whitening the data among others are discussed. Readers should be familiar with these topics before proceeding to the estimation sections.

Unique Aspects of Density Function Estimation

For a function to represent a pdf two conditions must be met: 1) the function needs to be nonnegative for each value within its domain, and 2) the area under the pdf must integrate to one. Conditions for a cdf are: 1) the function must be constrained to lie between zero and one, and 2) the function must be monotonically increasing (Taylor 1981). These conditions assure that the probability of obtaining a value or the cumulative probabilities obtained lie between zero and one. In addition, for both pdf and cdf estimation, the function should be flexible enough to approximate theoretical distributions which provide the foundation for statistical theory (Taylor 1981).

Problems arise when estimating density functions because of the previous conditions. Ordinary least squares (OLS) can not generally be used to estimate density functions. The inability to use OLS helps explain why density estimation is, generally, not given the same attention as other components of a model. Because of these conditions, maximum likelihood is often the technique used to estimate density functions. Fortunately, many authors have used maximum likelihood estimation to

obtain simplified expressions for some density estimators. This work helps to simplify density function estimation.

Classification of Estimation Techniques

Pdf and cdf estimation techniques can be classified in several different fashions. One such classification is parametric *versus* nonparametric estimation techniques. Parametric estimation techniques are a family of distributions which have parameters that often result in an inflexible function. An example of a parametric function is the normal distribution. The shape of the normal pdf is a single mode bell shaped curve. Parametric functions are not completely inflexible. The mean and the variance of a normal distribution do, for example, affect the location and to some extent the shape of the curve. Parametric estimators usually involve theoretical distributions, such as the normal or gamma distribution.

Nonparametric density estimators are considered more flexible than parametric estimators because they involve fewer restrictions (El harrack). These estimators let the data determine the unknown density function. Examples of nonparametric density estimators are histograms and kernel-estimators. In contrast to parametric estimators, nonparametric estimators usually do not involve a specific theoretical distribution, therefore providing increased flexibility over parametric estimation.

Discrete and continuous density functions are contained within both the parametric and the nonparametric family of density estimators. Within the parametric family the uniform and discrete uniform would be examples of continuous and discrete density functions. The aforementioned histogram estimator would be an example of a discrete nonparametric density function estimator, whereas, the kernel estimator would provide a continuous density function.

Classifying distributions by any scheme is fuzzy at best. It has been suggested that a semiparametric category be included in the parametric/nonparametric scheme. Classification schemes do not affect the estimation or use of a pdf or cdf; therefore, arguments concerning classifications of distributions and their estimators are purely schematic. Here, we adopt the parametric/nonparametric classification scheme strictly for ease of presentation.

General Advantages and Disadvantages

Advantages and disadvantages of the various techniques exist, both between and within the different types of estimators. In this section, some general advantages and disadvantages are discussed. More specific advantages and disadvantages are presented later for each particular estimator examined. In addition, not all authors agree on what is an advantage or disadvantage. An advantage, in one case, maybe a disadvantage in another situation.

Law and Kelton state that fitting a theoretical distribution is preferred to either using the data as is (incorporate each individual data point as being equally likely into the model) or estimating an empirical distribution. Reasons for their preference include:

1) a theoretical distribution smooths the data, therefore avoiding any irregularities in the data,

2) with an empirical distribution or using the data as is, it is usually not possible to generate values outside the observed data,

3) there may be a compelling reason to use a certain theoretical form, and

4) a theoretical distribution provides a compact way of representing a sample of data or a random variable.

It should be noted that Law and Kelton do not consider estimation of many of the flexible nonparametric estimators. They acknowledge that there are situations in which no theoretical distribution provides an acceptable representation of the data. In these cases, they recommend using the empirical distribution.

The preceding reasons pertain to why estimating a smooth distribution (theoretical or estimated smooth flexible density estimator) is preferred to using the data as collected (empirical and histogram). Another advantage associated with estimating a smooth distribution is that it is generally easier to handle a continuous function. This advantage is highly related to Law and Kelton's fourth reason. Maybe the most compelling advantage of continuous over discrete functions is that the estimation procedure can exploit statistical leverage that can be achieved by introducing continuity. This leverage is akin to estimating a production function instead of using each observation on yield separately. Estimating production functions is widely accepted, but this statistical leverage is not generally exploited in density estimation in agricultural research. This statistical leverage comes at the cost of increased knowledge and time necessary for estimation.

Using either a histogram or empirical distribution does have its advantage over estimating a smooth distribution. Estimating smooth distributions are usually harder than estimating a histogram or empirical distribution. There also exist instances when a histogram or empirical density function may be more appropriate than estimating a smooth distribution. One such case is data concerning the National Weather Service's long range precipitation and temperature forecasts. These forecasts are issued in three categories, above, near, and below normal. In this case, a three interval histogram would be more appropriate than trying to estimate a continuous pdf for the probability of receiving each forecast. Additionally, it is generally easier to go from a pdf to a cdf or vice versa using histograms than it is using estimated or theoretical distributions. For example, to obtain a cdf from a

55 normal pdf, either the pdf must be numerically integrated or a standard normal table must be used. On the other hand, to obtain a cdf from a histogram, the number of data points falling in each interval are summed to obtain the cdf.

The major drawback of fitting a theoretical distribution is that the pdf obtained is often inflexible. That is, the shape of the theoretical distribution may be fixed. Further, theoretical distributions generally do not allow for multimodal pdfs. Estimating a nonparametric distribution provides more flexibility in the shape of the pdf. As before, this increase in flexibility is not free. It is usually easier to fit a theoretical distribution than it is to estimate a nonparametric density function. But, fitting a theoretical distribution does have the potential advantage that the distribution is "known" for statistical testing.

This brief discussion of the advantages and disadvantages of the various types of techniques indicates that no one density estimation technique dominates the others. The appropriate estimation technique is a function of the data available, proposed use of the density function, study objectives, and the fit of the data to the estimated density function.

Cdf *versus* Pdf

In most applied models, the random variable(s) of interest are discretized. In this case, two options are to use the data as is or estimate a density function. Using the data as is, usually, entails assuming equal likelihood of the occurrence of each observation of the random variable. Then the model is run using only the values observed in the data set for the random variable. A major drawback of this approach is that no values of the random variable other than those observed can be used. Estimating a density function can circumvent this drawback.

The analyst must decide between estimating a pdf or a cdf. Once an analytic expression for either function is obtained, in theory the other function can be obtained. A cdf is the integral of the pdf, whereas, a pdf can be obtained from differentiating a cdf. Although in theory these transformations hold, in practice, these transformations are not always easy. Because of this difficulty, the purpose of the density function may help in determining which form to estimate.

A cdf may be easier to use than a pdf in some economic models, because of the aforementioned discretizing of random variables. With such discretizing, the probabilities associated with the random variable usually represent a range of values. To obtain the probability associated with a range, the cumulative probability associated with the smallest endpoint is subtracted from the cumulative probability associated with the largest endpoint. This is demonstrated in Figure 1. To find the probability of the random variable associated with the range A to B, the cumulative probabilities associated with both A and B are calculated, F(A) and F(B). Subtraction of the

probability associated with A, F(A), from the probability associated with B, F(B), gives the probability associated with the range (A, B). With a pdf, the area between A and B must be determined, usually by integration to determine the probability associated with the range (Figure 2). The ease of using a cdf instead of a pdf is obvious.

A major advantage of having an analytic expression for the density function over an empirical form or histogram is that the ranges (number and size) can easily be changed within the model with little, if any, programming effort. If a histogram is used, for example, the histogram must be reestimated before changes in the ranges can occur. Such a reestimation is not necessary when an analytic expression is used.

Combining Pdfs or Cdfs

Another procedure to develop density functions involves a weighted linear combination of two or more estimated density functions. There are no restrictions placed on the type of density functions that can be combined when using this *ad hoc* procedure. Restrictions placed on the weights are that they must be: (1) nonnegative, (2) less than one, and (3) must sum to one. These restrictions force the combination function to maintain the characteristics of a density function, namely the probabilities lie between zero and one, and the area under the pdf sums to one. This *ad hoc* procedure is rarely used in empirical economic modeling.

Preestimation Issues

Some general guidelines can be given when estimating either a pdf or cdf. It is suggested that the following procedures be accomplished before the estimation of a density function begins. No set order is necessary to accomplish the suggested preestimation procedures.

Most, if not all, univariate density function estimation procedures rely on the assumption that the sample data is independent and identically distributed. When working with data that is in chronological order (time series), one of the first steps should be to determine if correlation exists between the observations. A variety of tests exists for determining if such correlations exist. Several tests are discussed in the next subsection. For other tests and more detail on the tests presented here see statistical and/or econometric texts, such as Judge *et al.* or Granger and Newbold.

Besides testing for correlation, summary statistics associated with the data should be calculated. These summary statistics are useful not only in helping to determine which distribution to estimate, but also aid in determining the fit of the estimated density function to the data. At the very least, summary statistics which should be calculated are the mean, mode, median, standard deviation, coefficient of kurtosis, coefficient of skewness, and range (maximum and minimum). In addition, stem and leaf, boxplot diagrams, and quantile summaries are useful in examining a data set. Most

general introductory statistics texts discuss the calculation of these summary statistics. Creating several histograms may also help in determining the appropriate density function. Histograms are discussed in detail below.

Use of summary statistics in helping to determine which distributions to consider is not a statistical test, but does have statistical underpinnings. For example, if the data's mean, mode, and median do not coincide, assuming a normal distribution may be inappropriate. This information along with the coefficient of skewness may suggest that a gamma distribution is more appropriate. A histogram plot may indicate the distribution is multimodal. In this case, one of the nonparametric forms of estimation may be more appropriate than assuming an unimodal theoretical distribution. Another important consideration is theoretical and physical considerations underlying the data. These considerations may suggest which distribution is appropriate.

As noted earlier, determining if any correlation exists and calculating summary statistics should occur before any density estimation occurs. This information helps to determine which density function(s) to estimate. Estimation of several pdfs and cdfs are discussed. The researcher should not be confined to these forms, as many other density functions exist which may be appropriate.

Testing and Correcting for Trends and Correlation

The need to analyze the data for trends and/or correlations can easily be illustrated. Consider Figure 3 where the data clearly shows an upward trend. Assuming the trend is not spurious, failure to take the trend into account will give a higher sample variance than is appropriate. The variance after adjusting for the trend (variance around the trend) is less than the variance around the mean. Failure to account for the trend may lead to a more dispersed distribution of the stochastic variable than is appropriate. Incorporating such dispersion into an economic model leads to a "riskier" decision setting than appropriate.

Correlation between observations must also be analyzed. To illustrate the need to consider correlation between observations, consider December and January grain prices. A high price in December will normally be associated with a "high probability" of a high price in January, whereas, a low December price is associated with a "high probability" of a lower January price. This relationship is not exact, but the December price does affect the distribution of prices in January. Failure to account for this relationship in developing density functions using monthly prices leads to "poor" decision models.

The question is how does one account for time trends and/or correlations when estimating pdfs or cdfs. The easiest procedure to account for a time trend is to estimate the following model

$$x_t = \alpha + \beta t + \epsilon_t \qquad (1)$$

where x_t is the t^{th} observation on the variable of interest and t is a time trend. Significance of $\hat{\beta}$ indicates the presence of a time trend. Normal statistical tests are used in determining the importance of the trend in the data. Any univariate density function estimator can be used to estimate the distribution of the estimated error terms, $\hat{\epsilon}_t$. Under the assumption of the OLS model, the error distribution gives rise to the distribution of x_t.

First-order autocorrelation between observations can be tested by estimating the model

$$x_t = \alpha + \beta\, x_{t-1} + \epsilon_t . \qquad (2)$$

As before, statistics such as the significance of $\hat{\beta}$, R-squared, and the F-test, can be used to determine the strength of the inter-temporal correlation between observations.

Estimating equations (1) and (2) provide easy tests for trends and correlations. The presence of such factors complicates the estimation of density functions. A conditional pdf or cdf (conditional on time or lagged value) must be estimated. A simple procedure to overcome the estimation of a conditional cdf or pdf is to estimate the pdf or cdf associated with the error term. Estimation of the distribution of the error term is no more difficult than estimation of the original density function for x. Complications do arise in using the distribution. In most cases, these complications are, however, minor. See the multivariate section concerning the use of an estimated error term and for a more complete discussion of the necessary procedures or see Van Tassel, Richardson, and Conner.

COMPARISON OF ESTIMATED DENSITY FUNCTIONS

Unfortunately, there are no "goodness-of-fit" measures or procedures which compare estimated density functions which are applicable to all methods of estimating pdfs and cdfs. This lack of accessible measures contributes greatly to the usual manner in which density functions are estimated in applied work. Usually, only one density function is estimated and it is reported with little discussion. Contrast this to estimated production functions where usually the authors discuss estimation techniques, goodness-of-fit measures, estimated coefficients, and alternative models considered. As noted earlier, probabilities are one driving force behind stochastic models. Several basic reasons account for how density functions are estimated and reported.

First, typical graduate courses for nonstatisticians do not cover density function estimations or the subject is only rudimentarily discussed. Second, restrictions imposed on density function estimations (previously discussed) make the problem inherently more difficult. Third, most statistical tests are based on some form of the residuals, usually the error sum of squares (ESS). Some of the density function estimation techniques, for example histograms, do not readily lend themselves to a ESS analysis. Fourth, some of the measures developed rely on tabulated values based on known distributions. As such, many estimated density functions cannot be compared using these tests. Tabulated values which compare an estimated kernel function to an estimated hyperbolic tangent function are, for example, not available. The theory associated with goodness-of-fit measures are theses in themselves; therefore, only a sampling of comparison techniques are discussed. This is not to understate the importance of the choice of the density function issue, but rather to emphasize the complexities involved. Only a brief overview of some of the more common techniques is presented. Readers are encouraged to obtain more complete information concerning comparison techniques.

Many methods have been proposed to distinguish between pdfs and cdfs. Taylor (1983) suggests, for example, that Schwarz's information criteria can be used to determine which polynomial terms to include when estimating the hyperbolic tangent function. Nelson and Preckel use both an information test on the likelihood function and the likelihood ratio for specification tests. Kenkel, Buzby, and Skees use three different tests: chi-squared, Kolmogorov-Smirnov, and Anderson-Darling. Their results "...indicated that there was a substantial disagreement between the ranking generated by the various measures of goodness of fit" (Kenkel, Buzby, and Skees p. 13).

Statistical tests such as Schwarz information criteria normally rely on the logs of the likelihood function and a penalty for the number of parameters. As such, these tests are more appropriate for some density function estimation techniques than others. Such tests are, for example, not very useful when comparing different histograms or kernel estimators. Such statistical comparison tests are not discussed here. Only simple procedures to compare pdfs and cdfs applicable to most density functions are discussed.

Comparison Methods

Comparing the moments of the estimated density functions to calculated sample moments is one method of comparing the appropriateness of the estimated density function. For the density to accurately reflect the data, it is obvious that the moments need to be similar. Another technique is to simply plot the different density functions and compare them. The more similar the different estimated density functions are, the more confidence one has in the sense that the choice of function is not as critical. Further, the procedure may eliminate functions which appear to not represent the

situation being modeled. As with most of the techniques used to compare density functions, comparing moments and density function plots involve a great deal of subjectivity in determining the appropriate density function. Nevertheless, these simple comparisons are useful when comparing density functions. The following procedures expand on these simple comparisons.

Q-Q Plots: A quantile-quantile (Q-Q) plot is best used to aid in the visual inspection or comparison of the shape of two cdfs. Such a plot involves graphing the quantiles of one distribution against the quantiles of another distribution (Chambers *et al.*). Quantiles are the percentiles expressed in fractions instead of percentages. The .25 quantile is a number that divides the data into two groups such that .25 fraction of the observations fall below this value. A quantile must range between zero and one. The numerical value representing the quantile is denoted by Q(q), where $0 \leq q \leq 1$ represents the quantile. Another way of viewing a quantile is that q represents the cumulative probability and Q(q) represents the data numerical value which gives a cumulative probability equal to q.

To create a Q-Q plot, the cumulative density functions of the two distributions to be compared are necessary. In Figure 4, the idea of a Q-Q plot is illustrated. A quantile, q, is chosen. Next, for each cumulative distribution, G(x) and F(x), the value of the data which gives the cumulative density equal to q is found G(Q(q)) and F(Q(q)). This process is repeated for the number of quantiles of interest. The Q-Q plot is then created by graphing the values of G(Q(q)) and F(Q(q)) for the values of q. In practical terms, one axis is the range associated with G(x) and the other axis is associated with F(x). The paired points G(Q(q)) and F(Q(q)) are then plotted. If the distributions are identical, the points G(Q(q)) and F(Q(q)) lie on a 45 degree line. Deviations from this 45 degree line indicate how much and where the two distributions deviate from each other.

To illustrate the use of a Q-Q plot, consider the data given in Table 1. A Q-Q plot of each series' empirical cdf is presented in Figure 5 (see the later section on estimation of empirical cdf). Because the two series have an identical number of observations, the Q-Q plot of the empirical cdfs is simply a plot of the two sorted series (Chambers *et al.*). This plot indicates that the empirical cdfs have a similar "shape" for the smaller values in the series, but at larger data values, the cdfs diverge. Another important aspect, that of scaling, is illustrated in this simple example. Because Q-Q plots are used to compare the shapes of the cdfs, the data must be scaled to place each distribution in approximately the same range, series 2 was scaled by dividing by 20. Scaling is data specific, but a potential necessary component in creating appropriate plots. Other scaling procedures such as taking logarithms should be explored.

Table 1. Sorted Data and Associated Probabilities Used to Illustrate the Q-Q and P-P Plots.

Series 1	Series 2[1]	Cumulative Probability[2]	Probability Series 1[3]	Probability Series 2[3]
5	100	0.1	0.1	0.10
10	180	0.2	0.2	0.22
17	300	0.3	0.3	0.33
23	450	0.4	0.4	0.42
24	500	0.5	0.5	0.46
26	600	0.6	0.6	0.52
30	700	0.7	0.7	0.60
35	900	0.8	0.8	0.70
40	980	0.9	0.9	0.75
41	1000	1.0	1.0	0.76

1) In all calculations this series is scaled by a factor of 1/20.

2) Calculated using equation (52). Cumulative probability for each series is identical for each series, based on the observation number.

3) Calculated using equation (54) and series 1 for the x-value. The assumed lower bound on x is zero. These are the paired values graphed in Figure 7.

Comparisons with unequal numbers of observations or between distributions other than the empirical can be made using Q-Q plots. The concept presented in Figure 4 of obtaining a q value and then finding the value which gives this cumulative probability must be found for each cdf. These paired values are then plotted. Mathematically, the procedure involves solving the equations

$$\begin{aligned} G(Q(q)) &= q, \text{ and} \\ F(Q(q)) &= q \end{aligned} \qquad (3)$$

or

$$\begin{aligned} Q(q) &= G^{-1}(q), \text{ and} \\ Q(q) &= F^{-1}(q). \end{aligned}$$

The procedure is to find the values of the distribution in which the cumulative probability equals q. The difficulty in using equation (3) is that it is not always easy to obtain the inverse of the cumulative function or obtain the cdf from a theoretical pdf. An example of the latter is obtaining the cdf for the normal distribution. Chambers *et al.* provide a more thorough discussion of constructing Q-Q plots and provide approximations for quantile functions for some theoretical distributions.

P-P Plots: Probability-probability (P-P) plots assist in comparing the "shape" of two cdfs and are similar in concept to the Q-Q plots. The idea of a P-P plot is illustrated in Figure 6. For all values of x to be considered, G(x) and F(x) are determined. A P-P plot consists of graphing the values for G(x) and F(x). Again, for identical distributions the plots will lie on a 45 degree line on the graph. As can be seen in Figure 6, scaling of the distributions may be an important component.

The P-P plot for the data series in Table 1 is presented in Figure 7. In this Figure, equation (54) is used to find the probabilities for each series. The x-value (see Figure 7 or equation 54) associated with each cumulative probability is the series 1 value. As with the Q-Q plot, the P-P plot shows the density functions diverge at higher values for x.

Both Q-Q and P-P plots can be used to compare the shapes of any two distributions. Difficulties arise with some distributions in obtaining the cdf. With a P-P plot, the inverse of the cdf is not necessary, thus alleviating some of the problems associated with Q-Q plots.

Kolmogorov-Smirnov Test: Several tests have been developed that are concerned with the distance between the two comparison cdfs. One such test is the Kolmogorov-Smirnov (K-S) test. The K-S test compares the hypothesized distribution to the empirical distribution using the statistic

$$D_n = \sup |F_n(x) - G(x)| \qquad (4)$$

where D_n is the K-S statistic based on the sample size n, $F_n(x)$ is the sample empirical cdf, G(x) is the hypothesized distribution, x is the sample (data) observation, and sup | | represents the absolute value of the supremum. This statistic is simply the largest vertical distance between the empirical cdf and the hypothesized cdf. The K-S order statistic is

$$w = n D_n \quad \text{for } n \leq 40$$
$$\text{or} \quad w = \sqrt{n}\, D_n \quad \text{for } n > 40. \tag{5}$$

Values for the order statistic are compared to tabulated probabilities (Kraft and Eeden). If the tabulated probabilities, P_w, associated with w is less than the preassigned α level, then the null hypothesis that the sample comes from G(x) is rejected. Large deviations of $F_n(x)$ from G(x) lead to larger values for D_n, therefore, larger values for w. Large values for w are associated with smaller P_w, which increases the likelihood of rejecting the null hypothesis. The K-S statistic is sample size dependent and only compares the hypothesized distribution to the empirical distribution.

The K-S test assumes the empirical cdf is given by

$$F_n(X) = \frac{\text{number of } X_i \leq X}{n}. \tag{6}$$

Calculation of D_n is as follows, for each observation i, the following values are calculated

$$F_n(X_i)^+ = \frac{i}{n} \quad \text{and}$$
$$F_n(X_i)^- = \frac{i-1}{n}. \tag{7}$$

Then, for each observation, the absolute value of the two differences $F(x_i)^+ - G(x_i)$ and $F_n(x_i)^- - G(x_i)$ are found. D_n is the largest of these differences in absolute value (Kraft and Eeden).

An 1 - α confidence band around the empirical distribution can also be calculated using the K-S statistic (El harrack; and Kraft and Eeden). A preassigned α level is determined. From this level, a w is obtained from the K-S tables. Finally, a difference D_n is determined using equation (5) where w and n are known. The confidence band is found by adding and subtracting the calculated distance, D_n, to each point of the empirical cdf. Obviously, the upper limit of the band is set equal to one when the band is greater than one. Likewise, the lower limit of the band is set equal to zero when it is less than zero. It can easily be determined by graphing the confidence bands along with the various cdfs if the comparison cdfs are completely located within the band.

Unfortunately, the K-S test can only be used to determine if the estimated cdf is consistent with the empirical cdf. It can not be used to directly determine if two nonempirical (for example a kernel cdf and a Beta distribution) are consistent. One may possibly determine if the kernel is consistent with the empirical and if the beta is consistent with the empirical. From these two comparisons, it may be inferred that the kernel and the beta are consistent with each other. The power of this inference is weak and caution in doing such a procedure is highly stressed. Finally, the actual probability associated with the K-S test when doing such multiple comparisons is unknown to our knowledge.

Other Tests: Other tests based on differences between various cdfs exist, such as the Cramēr-Von Mises statistic and Anderson-Durling statistic (Durbin). Tabulated values for these statistics rely on the assumption of a specific theoretical distribution, for example, the normal distribution. As such, these tests are only applicable when a tabulated table for the theoretical distribution assumed exists. Not all distributions have been tabulated. Stephens suggests that the Cramēr-Von Mises and Anderson-Durling statistics are preferable to the K-S statistic when assuming the normal distribution as the theoretical distribution.

Other tests such as a chi-squared test can be used as a goodness-of-fit test. Tests such as this require the n observations to be classified into intervals. Selection of such intervals may affect the inference from the chi-squared test statistics. The selection of the intervals is, therefore, a major weakness of the chi-squared test.

Only a very brief introduction concerning the selection of the appropriate density function has been presented. Each procedure, whether heuristic or statistical in nature, has its drawbacks. For density function estimation procedures which rely on maximum likelihood estimation, procedures such as Schwarz's information criteria are readily available and should be used. These goodness-of-fit measures can be found in most statistical or econometric texts. Readers are encouraged to obtain additional information concerning model selection and density function estimation.

UNIVARIATE DENSITY ESTIMATION

Several parametric and nonparametric estimation techniques are discussed for the univariate case. As noted earlier, by no means is the discussion complete, but rather it is meant to be representative of the different techniques available and commonly used. Parametric approaches usually involve fitting observed data to a theoretical distribution. Nonparametric approaches involve using the observed data to estimate a flexible function which is then used to define the density function.

Nonparametric approaches obtain their flexibility by placing less constraints on the density function than parametric approaches. When the underlying distribution is unknown, the nonparametric approach is an appealing approach to density estimation. This approach allows the data to determine the distribution, rather than forcing a distribution on the data. Parametric approaches discussed are the 1) normal distribution, 2) gamma distribution, 3) exponential distribution, and 4) beta distribution. Nonparametric approaches discussed are 1) histogram, 2) empirical distribution, 3) J-distribution, 4) kernel estimators, and 5) hyperbolic tangent approach. Each section is fairly self-contained and the reader could easily proceed to the section(s) of interest, after reviewing the preestimation section.

Preestimation Aquifer Example

For each univariate approach considered, a distribution for simulated aquifer recharge levels is estimated. Using a single data set allows the different distributions and estimation approaches to be compared. In this subsection, the simulated data is briefly described. This description follows the preestimation considerations previously discussed. A listing of the data is presented in Appendix A.

A graph of yearly recharge data is presented in Figure 8. No trend appears to be present in the data. To statistically test if a trend is present, a regression of recharge as a function of time was performed. The following estimated equation was obtained

$$\begin{aligned} RC_t &= -7110.5 + 3.95\ t \\ &\qquad\quad (3.58) \\ R^2 &= 0.02 \end{aligned} \tag{8}$$

where RC_t is recharge at year t and the standard error of the coefficient is in parenthesis under the estimated coefficient. This regression supports the contention that no linear time trend is present.

A second regression performed on the data is to test if a Markovian relationship (correlation) exists between the recharge levels. To test for a first order Markovian relationship, the following estimated equation was obtained

$$\begin{aligned} RC_t &= 577.7 + 0.105\ RC_{t-1} \\ &\qquad\quad (0.13) \\ R^2 &= 0.01 \end{aligned} \tag{9}$$

where the standard error of the coefficient is in parenthesis. As with the trend analysis, the estimated equation supports the assumption that the data points are independent. It was expected that yearly recharge levels are independent, as the recharge depends on yearly weather conditions. Generally, it is not felt yearly weather conditions are related. If climatic change is occurring, this belief may be

changed. Based on the two regressions, it is assumed that the observations are independent; therefore, no adjustments are made to the data (unless otherwise noted) for use in the estimation procedures described below.

Summary statistics associated with the aquifer recharge data are presented in Table 2. These statistics indicate that the data are peaked and skewed to the right, relative to the normal distribution. A wide range of recharge levels are contained in the data, with the data ranging from 43.7 to 2003.5 acre feet (ac ft). The mean recharge is approximately 635 ac ft. A fitted distribution should retain, at least in part, these characteristics of the sample data. Summary statistics associated with the estimated distributions are presented for comparison purposes. A stem and leaf plot and a box plot are presented in Figure 9. These plots also show the skewness of the data toward high recharge levels.

The Normal Distribution

One of the most common parametric distribution used in statistics is the normal distribution. A univariate normal probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (10)$$

where μ is the mean and σ is the standard deviation of the population, π is the constant pi, and e is the base of the natural logarithms. Several important characteristics are associated with the normal distribution. One is that the distribution is continuous and symmetric around its mean (Kmenta). This implies that the mean, mode, and median are all equal, and that the mean divides the area under the normal curve in half. The distribution is unbounded, it extends from negative to positive infinity. Maximum height of the normal pdf is obtained where $x = \mu$. Points of inflection occur at $\mu \pm \sigma$.

The familiar bellshaped curve associated with the normal pdf is illustrated in Figure 10. This distribution is fully described by its mean and variance, that is, no other information is necessary to describe the distribution (Pindyck and Rubinfeld). Maximum likelihood estimates for the parameters, μ and σ^2 are the sample mean, $\bar{x}$, and sample variance, $\hat{\sigma}^2$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ and}$$
$$\hat{\sigma}^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (11)$$

where n is the number of observations and x_i is the i^{th} sample observation. The sample mean is an unbiased estimator of μ, but $\hat{\sigma}^2$ is a biased estimator of σ^2. An unbiased estimator of σ^2 is

$$\hat{\sigma}_u^2 = \hat{\sigma}^2 \left(\frac{n}{n-1}\right) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (12)$$

which is an adjusted variance (Hastings and Peacock). For large sample sizes, this adjustment has little effect on the parameter estimate.

Another important characteristic of the normal distribution is that its probabilities can be found by using the standard normal distribution. The standard normal distribution has a mean of zero and a standard deviation of one. Any normally distributed variable can be transformed into a standard normal variable. Probabilities associated with any normally distributed random variable x may be found by transforming x into a standard normal by the following formula

$$z = \frac{(x - \mu)}{\sigma} . \qquad (13)$$

The probabilities associated with x can then be found by using the standard normal (z) table found in most statistics textbooks. A disadvantage of this approach is that not all values for z are contained in these tables.

For a normal distribution, any value further than three standard deviations from the mean in either direction has a probability that is essentially zero. Coefficients of skewness and kurtosis for the normal distribution are zero and three. Calculation of these coefficients along with examining the mean, mode, and median of the sample are all *ad hoc* procedures to determine the appropriateness of assuming normality. A more in depth theoretical discussion of the normal distribution can be found in most mathematical statistics textbooks.

Ease of estimation, known statistical properties, and general acceptance of the normal distribution are advantages associated with using the normal distribution. The need to numerically integrate equation (10) or use the standard normal to obtain the cdf and fixed bell-shape pdf are disadvantages associated with the normal distribution. These disadvantages may make using the normal distribution in an applied model difficult.

Table 2. Summary Statistics Associated with the Simulation Aquifer Recharge Data.

Statistic	Value	Statistic	Value
Mean	635.52	Minimum	43.70
Mode[a]	-	Maximum	2003.50
Median	527.80	Coefficient of Kurtosis	1.26
Standard Deviation	422.55	Coefficient of Skewness	1.12

a) No observation occurs more than once.

Aquifer Application: Using the sample mean and variance (Table 2) as estimates of the true mean and variance, the estimated normal pdf is

$$f(x) = \frac{1}{422.55\ \sqrt{2\pi}}\ e^{-\frac{(x - 635.52)^2}{2(422.55)^2}}. \qquad (14)$$

A graph of the estimated pdf is presented in Figure 11. Aquifer recharge sample data ranged from 43.7 to 2003.5. The largest observation of 2003.5 is greater than three standard deviations to the right of the mean, whereas, the lowest observation of 43.7 is only 1.6 standard deviations to the left of the mean. Further, the coefficients of skewness and kurtosis for the normal distribution, zero and three, differ from the sample estimated coefficients of 1.12 and 1.26. These statistics indicate assuming normality for the recharge data may be inappropriate.

Two Parameter Gamma Distribution

The gamma distribution is just one of a family of exponential distributions. This distribution has played an important role in agricultural and medical research. The gamma distribution is frequently the pdf employed when modeling waiting times, for example, waiting for death or a defect to occur. It is also useful when the random variables are all nonnegative. This distribution is included as an example of a parametric distribution which has some degree of flexibility in the shape of its pdf. The two parameters which define this distribution provide this flexibility (Hogg and Craig).

A two parameter gamma pdf is given by

$$f(x \mid \alpha, \beta) = \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)}\ x^{\alpha - 1}\ e^{-x/\beta}, \qquad 0 \le x \le \infty \qquad (15)$$

where α and β are nonnegative parameters, x is the random variable, e is the base of the natural logarithm, and $\Gamma(\alpha)$ is the gamma function (Hogg and Craig). Values for the two parameters, α and β, define the shape of the function (see Figure 12). Several important distributions are special cases of the gamma pdf, depending on the values for the two parameters. The exponential distribution is, for example, a gamma distribution with $\alpha = 1$ and β nonnegative (see the next section). Another example of a special case of a gamma distribution is the chi-squared distribution with r degrees of freedom. In this case, $\alpha = r/2$ (r is any positive integer) and $\beta = 2$.

Other characteristics of the gamma distribution are as follows. The mean of the gamma distribution is equal to $\alpha\beta$, whereas, the variance is equal to $\alpha\beta^2$ (Hogg and Craig). These two characteristics provide one method (moment generating method) of obtaining parameter estimates for

the gamma distribution. The coefficient of skewness for the gamma distribution is $2\alpha^{-1/2}$, whereas, the coefficient of kurtosis equals $3 + 6/\alpha$.

The last important component of equation (15) is the gamma function, $\Gamma(\alpha)$. Several methods exists for finding values for this function. The gamma function is (Hogg and Craig)

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha - 1} e^{-y} \, dy. \tag{16}$$

Values for the gamma function could be obtained by performing the necessary integration, but this procedure is computationally long and burdensome. A simpler procedure is to obtain the values for $\Gamma(\alpha)$ from tables of logarithms of the complete Γ-function (Brownlee; and Pearson). Another approach is using Pearson's approximation for $\Gamma(\alpha)$. His approximation is

$$\Gamma(\alpha) = \sqrt{2\pi} \left\{ \frac{\sqrt{p^2 + p + 1/6}}{e} \right\}^{p + 1/2}, \quad \text{where } p = \alpha - 1. \tag{17}$$

Software packages also exist which can also be used to calculate the gamma function. One such package, Axum, is used to estimate the pdfs in this section.

The easiest method to obtain estimates for the two parameters is the moment matching or moment generating method. This method uses the sample mean and variance and the definition for the mean and variance of the gamma function (previously mentioned) to estimate α and β. Previously, it was noted that the mean and variance for the gamma distribution are

$$\begin{aligned} \mu &= \alpha \beta, \text{ and} \\ \sigma^2 &= \alpha \beta^2. \end{aligned} \tag{18}$$

Using $\bar{x}$ and $\hat{\sigma}^2$ (the sample mean and variance) as estimates for the true mean and variance (μ and σ^2), the following estimators for α and β are obtained

$$\begin{aligned} \hat{\beta} &= \hat{\sigma}^2/\bar{x}, \text{ and} \\ \hat{\alpha} &= \bar{x}^2/\hat{\sigma}^2. \end{aligned} \tag{19}$$

Derivation of the mean and variance formula's can be found in many statistics books (e.g. Bain; Hogg and Craig; Hasting and Peacock).

A second method to obtain estimates for the parameters is through the use of maximum likelihood estimation. The likelihood function for the two-parameter gamma function is

$$L(\alpha, \beta) = f(x_1|\alpha, \beta)\ f(x_2|\alpha, \beta)\ \ldots\ f(x_n|\alpha, \beta)$$

$$= \frac{1}{\beta^{n\alpha}\ [\Gamma(\alpha)]^n}\ \left[\prod_{i=1}^{n} x_i\right]^{\alpha - 1}\ e^{-\sum_{i=1}^{n} \frac{x_i}{\beta}}. \qquad (20)$$

To obtain maximum likelihood estimates (MLE) for the gamma function, the logarithmic form is normally used for ease of computation. This form is

$$\ln L(\alpha, \beta) = -n\alpha \ln\beta - n\ln[\Gamma(\alpha)] + (\alpha - 1) \ln \prod_{i=1}^{n} x_i - \sum_{i=1}^{n} \frac{x_i}{\beta}. \qquad (21)$$

Taking the partial derivative with respect to the two parameters and setting the resultant equations equal to zero and solving gives the MLE. The partial derivative with respect to β set equal to zero

$$\frac{\partial L(\alpha, \beta)}{\partial \beta} = -\frac{n\ \alpha}{\beta} + \sum_{i=1}^{n} \frac{x_i}{\beta^2} = 0. \qquad (22)$$

This equation can be simplified by multiplying each side of the equation by $\beta^2/\Sigma x_i$. This simplification gives the MLE for β as

$$\hat{\beta}_{MLE} = \frac{\bar{x}}{\hat{\alpha}}. \qquad (23)$$

Both maximum likelihood and the moment matching method give the same estimator for β (with same rearrangement).

Setting the partial derivative of equation (21) with respect to α equal to zero gives the MLE for α. This partial derivative is

$$\frac{\partial L(\alpha, \beta)}{\partial \alpha} = -n\ln\beta - \frac{n\partial\ln[\Gamma(\alpha)]}{\partial \alpha} + \ln \prod_{i=1}^{n} x_i$$

$$= -n\ln\beta - \frac{n\partial\ln[\Gamma(\alpha)]}{\partial \alpha} + \sum_{i=1}^{n} \ln x_i = 0. \qquad (24)$$

Substituting the MLE for β into equation (24) and rearranging terms gives

$$\frac{\partial\ln[\Gamma(\alpha)]}{\partial \alpha} - \ln\alpha = \frac{1}{n} \sum_{i=1}^{n} \ln x_i - \ln\bar{x}. \qquad (25)$$

Finding exact MLE for α using equation (25) is not "easy" because this equation contains the gamma function; therefore, approximation techniques have been developed.

Sinha approximates the partial derivation in equation (25) by $\ln \alpha - (2\alpha)^{-1}$. This approximation gives a point estimate for α of

$$\hat{\alpha}_{MLE} = \frac{1}{2\left[\ln\bar{x} - \frac{1}{n}\sum_{i=1}^{n} \ln x_i\right]} \tag{26}$$

Given a point estimate for α, equation (23) can be used to find $\hat{\beta}_{MLE}$.

Other approximations have been developed. Two such approximations depend on the fact that the term $(1/n)\ \Sigma \ln x_i$ in equation (25) is the logarithm of the geometric mean of the sample, that is

$$\ln \tilde{x} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i. \tag{27}$$

Substituting, equation (27) into equation (25) and multiplying by -1 yields

$$\ln\alpha - \frac{\partial \ln[\Gamma(\alpha)]}{\partial \alpha} = \ln\bar{x} - \ln\tilde{x}. \tag{28}$$

From equation (28) it is apparent that α depends only on the ratio of $\bar{x}$ and $\tilde{x}$. The arithmetic mean and geometric mean are, therefore, jointly completed, sufficient statistics for estimates of α and β (El harrack). Using this result, Bain gives an approximation for α which is

$$\alpha^{*}_{MLE} = \frac{1 + (1 + 4/3\ M)^{1/2}}{4M} \tag{29}$$

$$\text{where} \qquad M = \ln\left(\frac{\bar{x}}{\tilde{x}}\right). \tag{30}$$

Estimates for β are found using α^{*}_{MLE} and equation (23).

A second approximation based on the native of the arithmetic and geometric mean is an approximation given by Bain and Grice and Bain. This approximation comes from work by Wilk, Gnanadesikan, and Huyett. To use this approximation, Q is estimated by

$$Q = \left[1 - \left(\frac{\bar{x}}{\tilde{x}}\right)^{-1}\right]^{-1}. \tag{31}$$

Point estimates for α are then obtained by finding the value for α in tables which relate Q to α. (Bain, Table 1 p 326). A disadvantage of this approximation over the other two is that the Q - tables are necessary, that is the estimation relies on published tables and can not be completely computerized.

Finally, MLE of α and β could also be found by using nonlinear estimation techniques and numerical equations (22) and (24). Such procedures are beyond the scope of this paper when approximations which appear to provide satisfactory estimates are available. This omission is not to trivialize such MLE procedures.

Potential ease of estimation and the ability to represent different pdf shapes are the advantages of the gamma distribution. The necessity to evaluate the gamma function and numerically integrate the gamma pdf to obtain its cdf are the major disadvantages of this distribution.

Aquifer Application: Gamma distributions are estimated using the moment approach and Sinha's and Wilk's approximations for MLE. Estimated parameters along with summary statistics for the three distributions are presented on Table 3. Inserting these estimated values into equation (15) gives the estimated pdfs. The three estimated pdfs are graphed in Figure 13. Comparing the summary statistics and the graphs indicate that there appears to be only small differences in the estimation techniques. El harrack's findings are similar in that the approximation techniques are very close to one another. It appears that ease of estimation and use may be an important criteria when choosing which gamma approximation procedure to employ for estimation purposes.

Exponential Distribution

The exponential distribution is a special case of the gamma distribution (see previous section) and Weibull distribution with $\alpha = 1$ and β nonnegative in both cases (Freund and Walpole). This distribution is often used in analyses where one wants to calculate a waiting time before the first success or the waiting times between successes. The probability density function for the exponential is

$$f(x) = \begin{cases} 1/\beta\, e^{-x/\beta} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases} \qquad (32)$$

where $\beta > 0$. An advantage of using the exponential distribution is that the cdf can easily be obtained from the pdf. The cdf for the exponential distribution is

$$F(x) = \begin{cases} 1-e^{-x/\beta} & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases} \qquad (33)$$

The exponential distribution ranges from zero to positive infinity with a mean of β and a variance of β^2 (Law and Kelton). The general shape of the exponential pdf is illustrated in Figure 14. From the graph of the pdf, the exponential nature of this function is clear. It should be obvious why this function is often referred to as a decay function, with a high probability for lower x values, and lower probabilities associated with high values of x.

The maximum likelihood estimator for the parameter, β, is the sample mean, $\bar{x}$ (Hastings and Peacock). As expected, this is the same estimator derived through using either the moment estimating procedure or maximum likelihood for the gamma distribution (see gamma distribution discussion), when $\alpha = 1$. Any positive random variable ($x > 0$) can be used in equations (32) or (33) along with the estimate for β to obtain values for the pdf or cdf. Ease of estimation and use, along with a resulting continuous function, are advantages of the exponential function. Further, the exponential distribution's statistical properties are also known. Another advantage to the exponential distribution is the ease with which random variables can be derived from this function. Finally, the ease of going from the pdf to cdf or *vice versa* is a major advantage to using this distribution. A restrictive pdf shape is the major disadvantage to using this function.

Aquifer Application: The sample mean of 635.52 is the point estimate for β in the exponential distribution (Table 2). This estimate gives a pdf and cdf of

$$f(x) = \begin{cases} \frac{1}{635.52} e^{-x/635.52} & \text{if } x \geq 0 \\ 0 & \text{elsewhere} \end{cases} \quad (34)$$

and

$$F(x) = \begin{cases} 1 - e^{-x/635.52} & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases} \quad (35)$$

Summary statistics for the estimated exponential distribution are: 1) mean of 635.52, 2) standard deviation of 635.52, 3) coefficient of skewness equal to two, and 4) coefficient of Kurtosis equal to nine. Coefficients of skewness and kurtosis are fixed for the exponential distribution because of the assumption that $\alpha = 1$ (see gamma distribution discussion). The estimated exponential pdf and cdf are graphed in Figures 15 and 16.

Beta Distribution

The beta distribution is a flexible parametric distribution commonly used in empirical work (Law and Kelton). It has proven useful where a skewed bell-shaped, a J-shaped, or a U-shaped

Table 3. Estimated Parameter Values and Associated Summary Statistics for The Gamma Distribution.

	Estimation Technique		
	Moment	MLE Approximation	
Parameter Value	Matching	Sinha	Wilk
Alpha	2.26	2.09	2.25
Beta	280.95	303.42	282.49
Gamma (alpha)	1.14	1.04	1.13
Summary Statistics[1]			
Mean	635.52	635.52	635.52
St. Dev.	422.55	439.12	423.74
Skewness	1.33	1.38	1.33
Kurtosis	5.65	5.86	5.67

1) Calculated summary statistics, using the estimated alpha's and Beta. See Table 2 for the summary statistics associated with the data.

distribution is suggested. In empirical research where random events influence production, the skewed bell-shaped beta distribution has been used by various authors including Day, Nelson, and Nelson and Preckel.

Kendall and Stuart note that the beta distribution is a member of the Pearson family of distributions which are completely determined by their first four moments. This family of distributions also includes the gamma, t, and normal distribution as special cases (Johnson and Kotz). The beta distribution is sometimes referred to as a Pearson Type I distribution and has the probability density function

$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{(x-a)^{p-1}(b-x)^{(q-1)}}{(b-a)^{p+q-1}} \quad \text{for } a \le x \le b \tag{36}$$

where a and b are the lower and upper bound, p and q distribution parameters, and $\Gamma(\bullet)$ is the gamma function. Conditions placed on the parameters are that $p > 0$, and $q > 0$. If the transformation $y = (x - a)/(b - a)$ is made the probability distribution is defined over the zero-one interval and is referred to its standardized form. The beta distribution in standard form is

$$f(y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{(q-1)} \quad \text{for } 0 \le y \le 1. \tag{37}$$

The beta distribution's shape is determined by the values of the two parameters p and q. Both p and q must be greater than one for the distribution to be unimodal. When $p = q = 1$ the rectangular (or uniform) distribution arises and when p and/or q are below one a U-shaped or J-shaped distribution results. If both p and q are greater than one, then magnitude of the two parameters in relation to each other determines the skewness of the beta distribution: 1) $p < q$ implies positive skewness, 2) $p = q$ implies symmetry, and 3) $p > q$ implies negative skewness. Several potential shapes the beta distribution may take are illustrated in Figure 17.

A standard beta distribution's r^{th} moment about zero is

$$\frac{p^{[r]}}{(p+q)^{[r]}} \tag{38}$$

where $s^{[r]} = s(s + 1). . .(s + r - 1)$ is the ascending factorial. The mean and variance are, therefore

$$E(y) = \frac{p}{(p+q)}, \text{ and} \tag{39}$$

$$\text{var}(y) = \frac{pq}{(p+q)^2 (p+q+1)}. \tag{40}$$

Estimators for the beta distribution parameters are generally obtained through sample moments or numerical approximations. Estimation is substantially easier when the bounds, a and b, are known. If the bounds, a and b, are known, p and q may be obtained from the first and second moment by using the following equations

$$p + q = \frac{\left[\frac{\bar{x} - a}{b - a}\right]\left[1 - \frac{\bar{x} - a}{b - a}\right]}{\left[\frac{\hat{\sigma}^2}{(b - a)^2}\right] - 1}, \text{ and} \tag{41}$$

$$p = \frac{\left[\frac{\bar{x} - a}{b - a}\right]^2\left[1 - \frac{\bar{x} - a}{b - a}\right]}{\frac{\hat{\sigma}^2}{(b - a)^2}} - \frac{\bar{x} - a}{b - a} \tag{42}$$

where $\bar{x}$ is the mean and $\hat{\sigma}^2$ is the sample variance (Johnson and Kotz).

When one or both bounds are unknown the maximum likelihood procedures are, generally used. The maximum likelihood estimators for p and q may be obtained from solving the following two equations

$$\psi(p) - \psi(p+q) = \ln G_1, \text{ and} \tag{43}$$

$$\psi(q) - \psi(p+q) = \ln G_2. \tag{44}$$

where $\Psi(\bullet)$ is the digamma function

$$G_1 = \left(\prod_{i=1}^{n} x_i\right)^{1/n}, \text{ and } G_2 = \left[\prod_{i=1}^{n} (1 - x_i)\right]^{1/n}. \tag{45}$$

Solutions to these equations are not easily obtained though differentiation; therefore, numerical methods are generally used. Numerical solution techniques may be used to solve equations (43) and (44), or to maximize the log-likelihood function directly. When a and/or b are unknown iterative maximum likelihood procedures may still be used, but the computational burden is increased. Appropriate starting values help to reduce this computation burden. Fortunately, in many cases prior information can be used to provide good starting values for a and/or b. Approximations of the parameters p and q based on computed pairs G_1 and G_2 may be obtained by using tables based on the work of Beckman and Tietjen.

Advantages of the beta distribution are its flexibility to represent different shaped pdfs and the known statistical properties. Ease of estimation when the bounds are known is another advantage of this distribution. One major disadvantage of this function occurs when the bounds are unknown. In such a case, estimation is more difficult. Assumptions on specific values for the bounds may also affect the shape of the distribution. That is, one assumption may give a J-shaped distribution, whereas, another assumption may give a U-shaped distribution. Another disadvantage is that the distribution relies on the gamma function.

Aquifer Application: Beta distributions are estimated for the aquifer data using two procedures. First, equations (41) and (42) are used to estimate p and q assuming that a and b are 0 and 2003.5. A lower bound of 0 is the case of no rainfall, whereas, 2003.5 is the highest observed aquifer recharge level in the data set. The second procedure is the maximum likelihood estimation. Four alternative maximum likelihood results are presented to illustrated the influence that bound specification may have on the resulting parameter estimates. In all four maximum likelihood estimations the lower bound is assumed to be zero. In the first three maximum likelihood estimations the upper bound is imposed at 101, 125, and 150 percent of the highest observed aquifer level. In the fourth maximum likelihood estimation the upper bound is estimated along with the parameters p and q.

The resulting parameter estimates are reported in Table 4 and plots of the resulting distributions are shown in Figure 18. Using the moments estimator the values of p and q are 1.5 and 3.36. The maximum likelihood estimates vary substantially across the four specifications of the upper bound. In the three models with the upper bound imposed, both p and q increase monotonically as the upper bound is increased. When the upper bound is estimated along with the parameters p and q the estimated upper bound is 11336.53 which is more than five times the highest observed aquifer level. From Figure 18 it can be seen that as the upper bound is increased the distribution becomes more positively skewed.

Two potential problems arose when estimating or plotting the beta distribution. For some unknown reason, maximum likelihood estimates gave a U to J-shaped distribution when the upper bound was set equal to the highest value in the data set. Setting the upper bound equal to 101% of the highest value gave the shape displayed in Figure 18. Second, the estimated beta pdf's blew-up for recharge values equal to zero; this is the reason Figure 18 starts at a very small positive number.

Histogram

A histogram is a graphical plot of the underlying density function of the data. It provides a consistent estimate of the true underlying pdf. To create a histogram the data values must be sorted

in ascending order and placed into adjacent intervals. There are several methods for finding the number of intervals and the interval width. The starting value of the histogram must also be chosen. The lowest observation in the data is often used as the starting value, but starting values smaller than the lowest observed data point can be used. With this assumption, the height of each interval is given by the number of observations which fall in each interval range. Mathematically, a histogram is given by

$$f(x) = \frac{1}{n} \text{ (number of } x_i \text{ in the same interval as } x) \qquad (46)$$

where x_i is the i^{th} observation and n is the number of observations. Histogram widths or bins are

$$[x_o + mh,\ x_o + (m+1)\,h] \qquad (47)$$

where x_o is the lower bound, m the interval (or bin) number, and h the interval width. Usually these intervals are closed on the left side and open on the right side. Being closed on one side and open on the other allows observations which fall on an interval break to be placed into a single interval. One method of determining the number of intervals is Sturge's Rule. Sturge's Rule is

$$k = 1 + 3.322 \log_{10} n \qquad (48)$$

where k is the number of intervals. Once the number of intervals has been determined, the size of each interval must be calculated using the lower and upper bounds. Each interval range is then calculated by adding the size of the interval to the lowest value associated with the interval. If the lower bound to be considered is, for example, 5 and the interval size 10, the first histogram interval is from 5 to 15; the second interval is 15 to 25. This process is repeated until the upper bound is reached. In developing histograms, it is usually assumed that each observation has an equal chance of occurring.

Other authors have proposed that the interval width, rather than number of intervals, be calculated from the data. Scott suggests the formula

$$h = 3.49\hat{\sigma}\, n^{-1/3} \qquad (49)$$

for deriving the interval width. In this formula $\hat{\sigma}$ is the sample standard deviation. The number of intervals are found by adding the interval width to the starting point of the histogram, with the remaining intervals calculated by successively adding the interval width until the upper bound is

Table 4. Estimated Parameters of the Beta Distribution for the Aquifer Application.

Upper Bound Specification	Upper Bound	p	q
	Moment Estimation		
100% of Highest Observation	2003.5	1.5	3.36
	Maximum Likelihood Estimation		
101% of Highest Observation	2023.50	1.258	2.482
125% of Highest Observation	2504.38	1.638	4.714
150% of Highest Observation	3005.25	1.771	6.535
Estimated	11336.53	2.136	35.967

reached. Terrell proposed using the maximal smoothing principle to obtain the interval width. His expression is

$$h_{ms} = 20^{-1/3} \left[E(x) - \frac{(a + b)}{2} \right]^{2/3} (b - a)^{5/3} n^{-1/3} \qquad (50)$$

where, E(x) is the expected value for x (normally the sample mean, $\bar{x}$, is used), and a and b represent the data range. The maximal smoothing principle pertains to data ranges [a, b] which satisfy the following condition

$$\frac{7a + 3b}{10} \leq E(x) \leq \frac{3a + 7b}{10}. \qquad (51)$$

Another procedure commonly used in estimating histograms is a subjective approach to both lower and upper bounds and interval width. This approach is particularly useful when the intervals are predetermined, that is, the data is collected in interval form. The interval widths and bounds are predetermined in this case.

The interval width and lower bound are critical in estimating histograms. If the interval width is too small then the histogram will be too rough. If the interval width is too large then the histogram will be over smoothed, equivalent to having a large variance. A histogram that is over smoothed is statistically equivalent to having a large bias (Scott). Choice of lower bounds can also have a major effect on the shape of a histogram. The effect of interval widths and lower bounds are examined relative to the empirical example discussed later.

These methods for determining the number of intervals or interval width are meant to be rules of thumb, or a starting point in defining a histogram. Developing a histogram is always a subjective process with the investigator determining what seems to best fit the data and objectives of the study.

The major advantage of a histogram is that it is the easiest form of presenting a pdf (Ott). An additional advantage is the ease at which a cdf can be obtained. The cdf associated with a histogram is

$$F(x_j) = \frac{m_j}{n} \qquad (52)$$

where m_j is the total number of observations in the first interval through interval j. A stair step function is given by this cdf, as it is based on the intervals associated with the histogram. For data which is categorical, a histogram may be the appropriate pdf or cdf estimator.

Disadvantages of the histogram approach include its subjectiveness and lack of an analytic form. If the interval width, number of intervals and/or bounds are to be changed, one must recalculate the pdf by reclassifying each observation. No statistical leverage is obtained in smoothing the histogram.

Aquifer Application: Using the aquifer recharge data, the effect of different starting points and interval widths on a histogram are illustrated. Using Sturge's rule, the number of intervals is seven (rounded to nearest integer). If the upper and lower bounds are 43.7 and 2003.5 (the minimum and maximum of the data set) interval width is rounded to 280 ac ft. Note, using this interval width, the upper bound becomes 2003.7. The histogram developed using Sturge's rule is illustrated in Figure 19.

To illustrate the importance of starting values, Figure 20 contains a histogram with seven intervals, but the lower bound is zero (interval width of 286.29 ac ft.). A lower bound of zero assumes negative recharge is not possible. The probability of the second interval increases when the interval size is increased. The probabilities of the other intervals adjust slightly.

If Scott's formula is used, an interval width of 388 acre feet is obtained. Using this interval width and a starting value of zero acre feet, six intervals are necessary to cover the data range. This pdf is graphed in Figure 21. Scott's Rule weighs the lower aquifer recharge rates more than the other 1interval generation techniques; thereby, reducing the probability of higher recharge rates.

Continuous Empirical Distribution

Similar to the histogram, estimating an empirical cdf has the advantage of being one of the easiest approaches to obtaining a density function. Estimating an empirical cdf is similar to that of estimating a histogram. The major difference is that when estimating a histogram, an interval width is used, whereas, in estimating an empirical distribution, the number of intervals and their width are determined by the number of data points. Empirical cdfs can be estimated using several procedures, depending on the assumptions the analyst makes. Two different procedures are discussed in this section. The two approaches differ on their assumption concerning the smallest observation. Differences in the cdfs between the two approaches are for the most part limited to the lower end (left side) of the cdf. As the number of observations increase, the differences between the two approaches become less noticeable.

To estimate an empirical cdf, the first step is to sort the data from the lowest to the highest value. Usually, it is assumed each observation has an equal probability of occurring which is 1/n, where n is the total number of observations. This effectively gives the number of intervals equal to n. The difference between the two approaches discussed here is how the lowest value is treated. Let

$x_{(i)}$ denote the i^{th} smallest of the ordered x_j's, such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Under the definition that a cdf is the probability of obtaining a value less than or equal to x, an empirical cdf can be given by

$$F(x_{(i)}) = \begin{cases} \frac{i}{n} & \text{for all } x_{(i)} < x_{(n)} \\ 1 & \text{if } x_{(i)} \geq x_{(n)} \end{cases} \qquad (53)$$

where i is the ordered observation number. To clarify, if there are three ordered observations, say 3, 5, and 7, (then i = 1, 2, 3, and n = 3), the estimated empirical cdf using equation (52) is .33 or (1/3), .67 or (2/3), and 1.0 or (3/3).

The second approach assumes that there is zero probability of realizing a value less than $x_{(1)}$. Under this assumption, the empirical cdf becomes

$$F(x_{(i)}) = \begin{cases} 0 & \text{if } x_{(i)} < x_{(1)} \\ \frac{i-1}{n-1} & \text{if } x_{(1)} \leq x_{(i)} \leq x_{(n)}. \\ 1 & \text{if } x_{(1)} \geq x_{(n)} \end{cases} \qquad (54)$$

Using the same example as with the first approach, the difference between the two approaches becomes obvious. The cumulative probabilities associated with 3, 5, and 7, are 0 or (1-1)/(3-1), .5 or (2-1)/(3-1), and 1.0 or (3-1)/(3-1). A major difference in the cumulative probabilities associated with the values of 3 and 5 is apparent. In both approaches, the upper bound is $x_{(n)}$, this is why the cumulative probability associated with 7 is identical.

Difficulties exist in using the first approach to find cumulative probabilities associated with x values not contained in the data set. As long as the x value in question lies between $x_{(1)}$ and $x_{(n)}$ the difficulties vanish. The problem is with x values smaller than $x_{(1)}$. To obtain probabilities associated with any x value the following equations can be used

$$F(x) = \begin{cases} 0 & \text{for } x < \ell \\ \frac{0}{n} + \frac{x-\ell}{n(x_{(1)}-\ell)} & \text{for } x \leq x_{(1)} \\ \frac{i}{n} + \frac{x - x_{(i)}}{n(x_{(i+1)} - x_{(i)})} & \text{for } x_{(i)} \leq x \leq x_{(i+1)} \\ 1 & \text{for } x \geq x_{(n)}. \end{cases} \qquad (55)$$

where ℓ is the lower bound of the x range not necessarily the smallest observation on x. These equations use linear interpolation to obtain the probabilities associated with x values not in the data set. A lower bound for use in the interpolation part of the equation is not obvious. An additional assumption as to what this value is must be made.

The second approach to estimating the empirical distribution can also be used to find the cumulative probability for any value of x not contained in the data set used to estimate the cdf. In this case the cdf is

$$F(x) = \begin{cases} 0 & \text{if } x < x_{(1)} \\ \left(\frac{i-1}{n-1}\right) + \left(\frac{x - x_{(i)}}{(n-1)\,(x_{(i+1)} - x_{(1)})}\right) & \text{if } x_{(i)} \le x < x_{(i+1)} \\ 1 & \text{if } x \ge x_n \end{cases} \quad (56)$$

Linear interpolation is used in equation (56) between the $x_{(i)}$'s to estimate the cumulative probability for x values not contained in the data set. The second approach sets the lower bound for x values at $x_{(1)}$, whereas, the first approach assumed ℓ was the lowest possible value. This is the major difference between the two approaches.

A major disadvantage to using the empirical distribution is that random variables generated from it can never be less than $X_{(1)}$ (or ℓ) or more than $X_{(n)}$. This means that if the historical data (including ℓ) does not encompass the whole range of feasible values, then using the empirical distribution will truncate the distribution and one could never realize values below the observed minimum (or ℓ) and above the observed maximum value.

A greater problem with the continuous empirical distribution is the fact that the end points $x_{(1)}$ and $x_{(n)}$ are not observed with their true frequency, i.e., (1/n). For large sample sizes of 100 or more, this is not a problem; but for agriculture, where sample sizes of 10 to 20 are often used, the end points will never be observed in simulation. To correct this short coming, Richardson developed the J-Distribution.

J-Distribution: The J-Distribution was developed to overcome a major short coming of the continuous empirical distribution when dealing with small samples of actual data (Richardson). This distribution is an extension of the continuous empirical cdf. It differs from the continuous empirical distribution in how the highest and lowest values are treated during simulation. The Use Section of this paper describes use of distributions in simulation models in greater detail. The first step in using the J-Distribution is to sort the n data values from the lowest to the highest value. The lowest value

is, however, called $x_{(2)}$ and the highest value is referred to as $x_{(n+1)}$, (i.e., $x_{(2)}$, $x_{(3)}$, . . ., $x_{(n+1)}$). A value of $1-x_{(2)}$ is assigned to $x_{(1)}$ and the value of $1+x_{(n+1)}$ is assigned to $x_{(n+2)}$. In simulation, the J-Distribution is treated as if there are n+1 intervals and when a stochastic x value less than $x_{(1)}$ is observed, it is assigned the value of $x_{(2)}$. Similarly, when an x value greater than $x_{(n+1)}$ is observed it is assigned the value of $x_{(n+1)}$. In this way, the J-Distribution spreads out some of the variation from the interior intervals to the two end points. Given that the end points were observed with a 1/n probability, they should be observed with about the same frequency in simulation. The end points will only be observed for the continuous empirical cumulative distribution when the random number generator (U(0,1)) returns precisely 1.0000 and 0.0000 so each event occurs less than 0.001 percent of the time. Observation of end points is crucial for evaluating some agricultural distributions, such as, crop yields for crop insurance purposes or prices that are truncated by a loan rate.

For a sample size of 10 observed values, the end points for the J-Distribution will be observed 9.0909 percent of the time, making the bias 0.909 percent; if a sample size of 20 is used to develop the distribution, the end point bias is only 0.24 percent. Values for x in the $x_{(1)}$ to $x_{(2)}$ interval will occur 9.09 percent for a 10 interval cdf and 4.76 percent for a 20 interval cdf. Refer to the Use Section of this paper for a discussion of estimation procedure in simulation and its application to aquifer recharge data.

Aquifer Application: Estimated empirical cdf's using both approaches are graphed in Figure 22. Only small differences between the two approaches are evident. As expected, these differences are most notable at the lower end of the cdf. For any given x_i, the probability of realizing that value less than it can be determined by reading off the horizontal axis up to the cdf and across to the vertical axis. The probability of a recharge level less than 600, for example, is approximately 55 percent. So the probability of a recharge greater than 600 would be (1 - .55) or 45 percent. As one can readily see, the probabilities associated with high recharge levels are relatively low.

Kernel Estimators

Kernel estimators can best be described as a statistical procedure to smooth histograms. Wertz provides a heuristic motivation for the derivation and use of the kernel estimator. This derivation relies on the mean value theorem for integrals and mean square error. Wertz's derivation starts with the definition of a pdf, which is, a pdf is the derivative of the cdf. His first step is to apply the mean value theorem for integrals to obtain the variance and mean for estimation of the pdf (derivative) when the empirical cdf is substituted for the "true" cdf in the derivative. Using this result

to obtain the mean square error associated with the estimation procedure, a specific kernel estimator is obtained. This derivation provides the kernel estimator with a statistical basis.

The general formula for a kernel estimator is

$$f(x) = \frac{1}{nb} \sum_{i=1}^{n} K((x - x_i)/b), \tag{57}$$

where n is the number of observations, x_i is the i^{th} observation, K is a measurable function called the kernel, and b is a positive number called bandwidth, window width, or the smoothing parameter. The kernel function, K(w), must be a symmetric function satisfying the following conditions

$$\int_{-\infty}^{+\infty} K(w)\, dw = 1, \tag{58}$$

$$E(w) = \int_{-\infty}^{+\infty} w\, K(w)\, dw = 0, \text{ and} \tag{59}$$

$$E(w^2) = \int_{-\infty}^{+\infty} w^2\, K(w)\, dw = k_2 \neq 0 \tag{60}$$

where $w = (x - x_i)/b$. Condition (58) forces the area under K(w) to sum to unity, a requirement of a pdf. Conditions (59) and (60) state that the kernel function has a mean of zero and a positive variance, k_2. The variance falls out of condition (60) because the mean equals zero; therefore, condition (60) represents the second moment. The function f(x) will be a probability density inheriting all the continuity and differentiability properties of K(w). Rosenblatt considered the estimation of the density f(x) from the observed data and noted that for any non-negative K(w) satisfying condition (58), a consistent estimator of the density function can be obtained.

The kernel estimator can be considered as a sum of "bumps" placed at the observations. The shape of the bumps are determined by the kernel function, K(w), and the bandwidth, b. Larger values for b cause the function to be smoother, whereas, smaller values for b result in the function spiking at each observation (see Figure 23). As shown in Figure (23), the choice of bandwidth is an important step when using kernel estimators.

In examining several different kernel estimators, Silverman concluded that their efficiencies relative to the Epanechnikov kernel is close to one. Silverman reports, for example, that the efficiency between the Epanechnikov and Gaussian kernel functions is 0.95. This conclusion implies

that, on the basis of mean integrated square error (MISE), the choice of which kernel function to employ is not very important. That is, using different kernel functions will not provide significantly different estimated pdfs. Ease of use may be an important factor in determining which kernel function to employ. An Epanechnikov kernel function may be preferable, for example, to a Gaussian kernel if a cdf is desired (El harrack). Several commonly used kernel functions and relevant properties, as reported by Wertz, can be found in Table 5.

As noted earlier, the choice of b remains an important issue. Several different methods of determining an "optimal" value for the smoothing parameter have been proposed. One such method relies on the MISE. MISE is the appropriate measure to examine how closely the estimated kernel is to the true pdf (Boyd and Steele; Bullock; Rosenblatt; and Silverman). The MISE measure is

$$\begin{aligned} MISE(f) &= \int MSE_x(f)\, dx = \int Bias^2 + \int Var \\ &= \int \{Ef(x) - f(x)\}^2\, dx + \int Var\ f(x) \end{aligned} \tag{61}$$

where $MSE_x(f) = \{Ef(x) - f(x)\}^2 + \text{var } f(x) = Bias^2 + Variance$. Shown in equation (61) is the trade-off between bias and variance. This trade-off can be manipulated by adjusting the amount of smoothing (changing the value for b). Attempts to minimize the bias will increase the variance and *vice versa*. While the bias will be independent of sample size it is dependent on both the bandwidth and the kernel function.

Several authors (Rosenblatt; Silverman; Devroye and Penrod) have suggested that an optimal value for b can be obtained by minimizing an approximate MISE. This approximation involves a change in variables, namely $y = x - bw$, and employing a Taylor's series expansion of MISE. The approximate MISE obtained is:

$$MISE = \frac{1}{4} b^4 (k^2)^2 \int f''(x)^2\, dx + n^{-1} b^{-1} \int K(w)^2\, dw \tag{62}$$

where k^2 is the variance of the kernel function and f'' is the second derivative of f. An optimal value for b can then be found by minimizing the MISE. Differentiating equation (62), setting the derivative equal to zero, and solving gives the following approximately optimal value for b

$$b_{opt} = (\sigma^2)^{-2/5} \left[\int K(w)^2\, dw\right]^{1/5} \left[\int f''(x^2)\, dx\right]^{-1/5} n^{-1/5}. \tag{63}$$

Table Three Kernels and Their Properties.[1]

nel	K(w)	$\int K(w)dw=1$	$\int w^2K(w)dw=1$	$\int K^2(w)dw=$
Epanechnikov	$(3/4\sqrt{5})(1-(1/5)w^2)$ for $\lvert w\rvert<\sqrt{5}$; 0 otherwise	yes	yes	$3/5\sqrt{5}=$ 0.2683
Triangular	$1-\lvert w\rvert$ for $\lvert w\rvert<1$; 0 otherwise	yes	yes	$\sqrt{6}/9=$ 0.2722
Gaussian	$1/(\sqrt{2\pi})e^{-(1/2)w^2}$	yes	yes	$1/(2\sqrt{\pi})=$ 0.2821

1) Source: Wertz, p.34

Unfortunately, the value for b_{opt} depends on both the kernel function employed and the unknown "true" pdf. Silverman reports that for the Gaussian kernel and a "true" normal distribution, b_{opt} is

$$b_{opt} = 1.06\ \hat{\sigma}\ n^{-1/5} \tag{64}$$

where $\hat{\sigma}$ is the sample standard deviation. Using an Epanechnikov kernel and a normal "true" distribution, El harrack reports a value for b_{opt} of

$$b_{opt} = 1.049\ \hat{\sigma}\ n^{-1/5}. \tag{65}$$

Another procedure for choosing the smoothing parameter has been proposed by Terrell. He suggests using the maximal smoothing principle which, for a kernel with unit variance, is

$$b_{ms} = 3\ (35)^{-1/5}\ \hat{\sigma}\ (\int k^2)^{1/5} n^{-1/5}. \tag{66}$$

Other methods for choosing the smoothing parameter not discussed here include the least squares cross validation, the likelihood cross validation, the test graph, simulation method and the Scott-Tapia-Thompson method (Roeder; Silverman; Tapia and Thompson; and Staniswalis). These methods provide some structure and generally have a statistical basis to choosing the smoothing parameter. One method with little statistical basis is the subjective choice method. In this method, a plot of several pdfs using different smoothing parameter values is made and the parameter value which seems most appropriate is chosen.

As indicated by Silverman, the choice of the smoothing parameter appears to be more important than the choice of the kernel in density estimation. Several different methods of choosing the parameter have been proposed. Unfortunately, most methods rely on assumptions concerning the true underlying pdf and subjectiveness on the researchers' part. Fryer concludes that the methods for choosing the smoothing parameter are computationally long. He states his personal preference ". . . is to plot the estimates for f(•) for several values of h_n" (refers to b in our notation) "and subjectively choose - usually taking h just large enough to eliminate bumps at outlying observations" (Fryer p. 350).

An advantage of the kernel estimator is its statistical basis and accepted use. Ease of estimation provides another advantage over techniques which require maximum likelihood estimation procedures. The major drawback associated with kernel estimators is the lack of obtaining an analytic form (equation) which can be easily incorporated into stochastic models. The lack of such form also precludes the performing of any statistical tests on parameters and the developing factors such as

elasticities. Kernel estimators remain useful, however, in estimating an unknown pdf for applications such as illustrative purposes and forecasting.

Clarifying Example: At this point, a simplified example may be useful in explaining the kernel estimator; the use of equation (57) is not intuitively obvious the first time it is applied. Consider estimating a Gaussian kernel pdf using a data set which contains three observations, 1.0, 1.5, and 3.0. The following discussion outlines the necessary steps.

One of the first steps is to determine the bandwidth to be used. Using the "optimal" bandwidth given by Silverman (equation (64)) a value of 0.723 is obtained for b. Next, values for x in equation (57) must be determined. Any values for x can be used. It is suggested that the values for x cover the range of the data set. Here, we will use values for x ranging from zero to four in increments of 0.25 (arbitrarily chosen). Next, each individual kernel must be calculated. Each observation, x_i, and x value considered have a kernel associated with them. For the Gaussian kernel, the individual kernel centered at the i^{th} observation is

$$\frac{1}{nb} K(w_i) = \frac{1}{nb} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x - x_i)^2}{b}} \qquad (67)$$

where x is any point, x_i is the observed data point, and b is the bandwidth. Calculated kernels for the example are presented in Table 6.

To clarify the individual kernels, consider the column of numbers associated with the first observation, $x_i = 1.0$. The values in this column are calculated by changing the values for x (from zero to four in increments of 0.25) in equation (67) and leaving the value for x_i fixed at 1.0. Values obtained when x_i equals 1.5 and 3 are calculated in a similar fashion. Kernel estimators for pdfs are obtained by summing the individual kernels. This sum is presented in the last column in Table 6. A graph of the individual kernels and their sum is presented in Figure 24. Programming kernel estimation can easily be accomplished.

Aquifer Application: Two different kernel estimators, the Epanechnikov and Gaussian kernels are used to estimate the probability density function associated with the aquifer recharge data. Silverman's "optimal" smoothing parameter is 200.96 for the Gaussian kernel and 198.88 for the Epanechnikov kernel. Using these parameters, both estimated kernels are graphed in Figure 25. As expected, the two estimated pdfs are very similar. This was expected given Silverman's estimate of an efficiency of 0.95 between the two kernels.

Hyperbolic Tangent Function

Taylor's (1981, 1983, 1984, and 1987) hyperbolic trigonometric (HT) transformation is a procedure for the empirical estimation of both unconditional and conditional cdfs. Upon differentiation the pdf is obtained. The HT transformation satisfies the conditions placed on cdfs.

The motivation for the HT transformation stems from the hyperbolic tangent

$$\tanh v = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} \quad \text{for all } v \; -\infty \leq v \leq \infty \tag{68}$$

where v is any real number and tanh is the hyperbolic tangent operator which ranges from negative one to positive one (Ayres). When graphed, the hyperbolic tangent resembles a unimodal cdf in terms of its curvature properties and, in addition, it can be shown that the coordinates (0,0) give rise to an inflection point indicating that the derivative of equation (68) resembles a traditional bell-shaped pdf. Given the bounds on the hyperbolic tangent, it is necessary to transform the function such that the cdf is constrained to lie between zero and one. Taylor proposes the following transformation, multiply the hyperbolic tangent by one-half and then add one-half. The first transformation effectively squeezes the bounds of the function to $\pm$ 0.5, while the second shifts the entire curve upward by one-half. This transformation is illustrated in Figure 26. In mathematical terms, the transformation is

$$F(x) = \frac{1}{2} + \frac{1}{2} \tanh\,[v(x)] \quad \text{for all} \quad -\infty \leq v(x) \leq \infty \tag{69}$$

where F(x) is the cdf of x and v(x) is any function of x. For any real value taken on by the function v(x), the transformation effectively constrains F(x) to the zero-one interval. The function, v(x), gives flexibility to the transformation, permits additional modes to the pdf, and allows for the pdf to be skewed in either direction or to be symmetrical (Taylor 1981). Estimation of equation (69) requires the estimation of the parameters associated with the function, v(x). Taylor (1983) proposes the method of maximum likelihood (ML) to obtain the parameter estimates of v(x).

Table 6. Calculated Individual Kernels and Estimated Pdf for the Gaussian Kernel Example.[1]

	Observation Value			
x	1.000	1.500	3.000	Pdf[2]
0.000	0.071	0.021	0.000	0.092
0.250	0.107	0.041	0.000	0.149
0.500	0.145	0.071	0.000	0.216
0.750	0.173	0.107	0.001	0.282
1.000	0.184	0.145	0.004	0.333
1.250	0.173	0.173	0.010	0.356
1.500	0.145	0.184	0.021	0.350
1.750	0.107	0.173	0.041	0.322
2.000	0.071	0.145	0.071	0.286
2.250	0.041	0.107	0.107	0.256
2.500	0.021	0.071	0.145	0.237
2.750	0.010	0.041	0.173	0.224
3.000	0.004	0.021	0.184	0.209
3.250	0.001	0.010	0.173	0.185
3.500	0.000	0.004	0.145	0.149
3.750	0.000	0.001	0.107	0.109
4.000	0.000	0.000	0.071	0.071

1) Calculated values are $(nb)^{-1} K((x-x_i)/b)$, where K represents the Gaussian kernel (see equation (66)), b is the scaling factor, and n is the number of observations.

2) Sum of the calculated individual kernels.

Maximum Likelihood: Maximum likelihood estimation of the parameters associated with v(x) is briefly discussed. Using the quotient rule and rules of differentiation for hyperbolic functions, the pdf associated with equation (69) is

$$\frac{\partial F(x)}{\partial x} = f(x) = \frac{[2\ (e^{v(x)} + e^{-v(x)})]\ [(e^{v(x)}\frac{dv}{dx} + e^{-v(x)}\frac{dv}{dx})]}{[2(e^{v(x)} + e^{-v(x)})]^2} - \frac{[(e^{v(x)} - e^{-v(x)})]\ [2(e^{v(x)}\frac{dv}{dx} - e^{-v(x)}\frac{dv}{dx})]}{[2(e^{v(x)} + e^{-v(x)})]^2}. \tag{70}$$

This expression can be simplified and rewritten as follows

$$f(x) = \frac{4\ \frac{dv}{dx}}{2(e^{v(x)} + e^{-v(x)})^2} = (\frac{1}{2})\ (\frac{2}{e^{v(x)} + e^{-v(x)}})\ (\frac{2}{e^{v(x)} + e^{-v(x)}})\ \frac{dv}{dx}. \tag{71}$$

A further simplification can be accomplished by noting that the two terms involving e are the inverse of the hyperbolic cosine. Because the inverse of the hyperbolic cosine is the hyperbolic secant, equation (71) simplifies to (Taylor 1981)

$$f(x) = \frac{1}{2}\ \text{sech}^2\ v(x)\ \frac{dv}{dx}. \tag{72}$$

Equation (72) characterizes the pdf associated with the cdf as represented by the hyperbolic tangent function (equation (69)).

The likelihood function associated with equation (72) is

$$L(\delta) = \prod_{i=1}^{n} \frac{1}{2}\ \text{sech}^2\ v(x_i)\ \frac{dv}{dx_i} \tag{73}$$

where δ is the parameter vector (m x 1) associated with v(x) with m being the number of parameters to be estimated. Taking the natural logarithm, the log-likelihood function becomes

$$\ln L(\delta) = n \ln (\frac{1}{2}) + 2 \sum_{i=1}^{n} \ln \{\text{sech}\ [v(x_i)]\} + \sum_{i=1}^{n} \ln (\frac{dv}{dx_i}). \tag{74}$$

The maximum likelihood estimator is derived from the partial differentiation of the log-likelihood function with respect to each element of the parameter vector, δ. The resulting m partial derivatives (equations) are then set equal to zero and solved simultaneously to obtain the estimates for δ. The general form of the partial derivative of equation (74) taken with respect to the j^{th} element is

$$\frac{\partial \ln L(\delta)}{\partial \delta_j} = -2 \sum_{i=1}^{n} \left(\frac{1}{\text{sech } v(x)}\right) (\text{sech } v(x) \tanh v(x)) \left(\frac{\partial v(x)}{\partial \delta_j}\right) + \sum_{i=1}^{n} \left(\frac{1}{(dv/dx_i)}\right) \left(\frac{\partial (dv/dx_i)}{\partial \delta_j}\right) \tag{75}$$

for j = 1, ..., m. This result can be set equal to zero, simplified, and rewritten as

$$\frac{\partial \ln L(\delta)}{\partial \delta_j} = 0 \; ;$$

$$\sum_{i=1}^{n} \left(\frac{\partial (dv/dx_i)}{\partial \delta_j}\right) \left(\frac{dv}{dx_i}\right)^{-1} = \tag{76}$$

$$2 \sum_{i=1}^{n} (\tanh v(x_i)) \left(\frac{\partial v(x_i)}{\partial \delta_j}\right).$$

The j equations given by equation (76) could then be solved simultaneously to obtain the estimates for each of the elements in δ. Because solving such equations is tedious, Taylor proposes the use of numerical search procedures to solve for the estimate of the element of the vector δ (Taylor 1981). Use of an iterative numerical search routine that uses Wolfe's algorithm has been used by Taylor (1983) for solving for up to a third degree polynomial specification of v(x). Taylor's (1983) program also assesses the significance of each polynomial term in the regression equation as well, so a determination of the "best" model can be used when estimating the unconditional cdf. As with any search procedures, good starting values are important in obtaining estimates for δ. The next subsection gives one procedure to obtain starting values. Appropriate starting values increase the reliability of the parameter estimates and minimizes computational time. While ordinary least squares (OLS) estimates of equation (69) are biased and should not be used to estimate the cdf, they typically provide good starting values for the maximum likelihood estimation.

One important consideration is that the hyperbolic tangent function has trouble handling values in excess of $\pm$ 7.50. As a result, the data may need to be scaled to keep the range within the appropriate bounds. Dividing by a multiple of the sample mean or standard deviation usually provides appropriate scaling.

OLS Estimation: To estimate equation (69) using OLS an additional transformation is necessary to express v(x) as a function of F(x). To clarify, consider

$$F(x) = \frac{1}{2} + \frac{1}{2} \tanh [v(x)]. \tag{77}$$

Inverting the equation gives

$$v(z) = \tanh^{-1} [z], \quad \text{where:} \quad z = 2F(x) - 1. \tag{78}$$

Using the definition of an inverse hyperbolic tangent function equation (78) simplifies to

$$v(z) = \frac{1}{2} \ln \left(\frac{1+z}{1-z}\right), \quad z^2 < 1 \tag{79}$$

which reduces to the following upon substitution for z

$$v(x) = \frac{1}{2} \ln \left(\frac{F(x)}{1-F(x)}\right). \tag{80}$$

To obtain OLS estimates of equation (80), the following procedure is suggested. In equation (80), v(x) can be any function of x and the right hand side of equation (80) is the dependent variable. Values for the dependent variable are obtained by calculating F(x) using an empirical cdf. First, the observations are ranked from smallest to largest, a cumulative frequency is assigned to each x_i such that $F(x_{(i)}) = i/n$ where i represents the ranked observation number and n, the number of observations. Then each of the $F(x_{(i)})$'s are transformed by equation (80) to give a finite $v_{(i)}$. This $v_{(i)}$ is used as the dependent variable in the OLS regression. The last ranked observation, $x_{(n)}$, poses a potential problem because the denominator in equation (80) goes to zero. There are several practical methods for dealing with the problem associated with v_n. A slight downward adjustment of either $F(x_{(n)})$ only or for all $F(x_{(i)})$ can overcome the infinite problem. A small amount of bias is introduced by these adjustments, but a finite value for $v_{(n)}$ is attained. Another procedure is to simply drop $x_{(n)}$ from the data when using OLS. No matter which method is employed, OLS can be applied directly to the modified data set, $(x_{(i)}, v(x_{(i)}))$, to estimate the parameters of v(x).

Aquifer Application: The numerical search procedure previously discussed is used to estimate a cdf for the aquifer recharge data. Because the data ranges from 43.7 to 2003.5, which is far outside the range of $\pm$ 7.5 which is appropriate for estimation of the hyperbolic tangent transformation, the observations must to scaled. Scaling is accomplished by dividing each observation by three times the sample standard deviation. As noted above, this transformation is necessary to prevent the hyperbolic tangent from "blowing up." Several different model specifications for u(x), where x is the scaled

recharge level, are considered. Models considered are: 1) linear recharge, x, 2) squared recharge, x^2, 3) recharged cubed, x^3, 4) recharge linear and squared, x and x^2, 5) linear, squared, and cubed x, x^2, and x^3, 6) linear and cubed, x and x^3, and 7) squared and cubed, x^2 and x^3.

Taylor recommends using maximum likelihood on only those models in which the OLS parameter t-values are greater than four in absolute value. Experience suggests that the MLE significant levels are approximately one-half those of the OLS estimates used to obtain the starting values. Only those models in which all of the estimated parameters satisfy Taylor's recommended t-value of four or greater are estimated using maximum likelihood. The "best" model is chosen as the model which maximizes Schwarz's Model Selection Criteria (Judge *et al.*). A note on Schwarz criteria is appropriate here. At least two forms of the criteria have been used. Schwarz criteria is

$$SC = \ln(ML) - .5k_1 \ln T \tag{81}$$

where ML is the value for the maximum likelihood function, k_1 is the number of parameters to be estimated, and T is the number of observations. With this formulation, the criteria is to choose the model with the largest SC. Another formulation is to add rather than subtract the last term in equation (81). Under this alternative formulation, the criteria is to choose the model with the smallest SC. Taylor's (1983) program uses the formulation given in equation (81); therefore, the criteria is to choose the model with the largest SC.

The "best" HT cdf model for the aquifer data as given by the Schwarz criteria is

$$F(x) = \frac{1}{2} + \frac{1}{2}\tanh\left[-1.987 + 6.732(x_i) - 5.74(x_i^2) + 2.119(x_i^3)\right]. \tag{82}$$

$$\qquad\qquad (-6.054) \quad (4.568) \quad (-2.618) \quad (2.088)$$

$$SC = -18.714$$

where x_i is the scaled data, t-ratios are given in parentheses below the parameter estimates, and SC is the value for the Schwarz model selection criteria. For any values of x_i then, equation (82) gives the corresponding probability of being at x_i or below it. In terms of the data used for the estimation, for a given total recharge, say the mean of 635.52, the probability of being at the mean or less is 0.6046. Alternatively, the probability of being at a recharge level greater than the mean is 1 - 0.6046 = 0.3954. A graphical representation of the cdf estimated by the HT procedure is found in Figure 27. The pdf corresponding to the estimated HT cdf is illustrated in Figure 28. A note on obtaining the pdf is warranted. With scaled data, the estimated pdf is also scaled, that is, the pdf must be divided by the scaling factor.

Concluding Remarks - Univariate

Several different univariate density functions and their estimation are discussed. Because the "true" underlying function in empirical studies is seldom, if ever, known, the question remains which function is appropriate. In Figures 29 and 30, the several aquifer estimated pdfs and cdfs are plotted. These plots clearly show that differences between the functions exist. Using only nonparametric functions does not eliminate the differences; for example, compare the kernel to the hyperbolic tangent transformation in Figure 29. The choice of distribution may have an effect on decision/policy models. In the simulated aquifer data, for example, the estimated normal distribution has a higher probability for recharge levels between approximately 700 and 1400 acre feet than the other distributions. In addition, the normal distribution has a lower probability of recharge level between approximately 200 and 600 acre feet, than the other distributions. Other differences between the distributions are also evident. Such differences may affect the inference from other modeling activities. Further testing, such as Q-Q or P-P plots, are necessary to determine the difference between the estimated aquifer recharge data.

The preceding discussion and the aquifer recharge data illustrate an important point. Choice of density function is an important consideration in model building. What is missing is information on how the choice of distribution may bias economic and other models. Studies addressing this issue are necessary. One drawback of such studies is that they may be data and model specific. Nevertheless, such studies would provide valuable information to applied researchers.

MULTIVARIATE ESTIMATION PROCEDURES

Three different multivariate techniques are presented. As with the univariate estimation presentation, these procedures are not meant to be all inclusive, but rather represent a diversity of techniques available. The first technique presented relies on modeling the error term associated with an estimated equation to obtain a multivariate distribution. Although this procedure does not give a "true" multivariate distribution; it does provide a useful technique in some cases. The next multivariate technique presented is an extension of the hyperbolic tangent transformation. To completely understand these techniques, the reader should review the appropriate univariate estimation discussion. The final multivariate estimation procedure presented is for the normal case and then extended to the nonnormal case. This complex procedure for generating stochastic values from a multivariate J-Distribution over time is presented by way of explaining the procedure used for FLIPSIM (Richardson and Nixon).

Use of an Estimated Error Term

As discussed earlier, correlation between observations needs to be examined. If correlation exists, this must be taken into account. One way to account for this correlation is to estimate a

multivariate density function. Estimation of multivariate functions is more difficult than univariate functions. Presented in this subsection is a procedure to estimate a univariate density function to represent a function which depends on several variables. The procedure is illustrated in terms of a production function.

For simplicity, consider a production function which is a linear function of a single variable input

$$y_t = \alpha + \beta x_t + \epsilon_t \tag{83}$$

where y_t is yield, α and β are parameters to be estimated, x_t is the variable input, and ϵ_t is the error term for the t^{th} observation. The problem is to estimate the distribution of yields conditional on x_t. One procedure would be to estimate either a joint or conditional multivariate density function for yields depending on a given x. As noted earlier, this may be a difficult problem. Another procedure is illustrated here. This procedure relies on the assumptions of the OLS model.

One of the assumptions of the OLS model is that x's are fixed, that is they are nonstochastic. Given this assumption, yield is random only because of the stochastic error term. To obtain the distribution of yields conditional on the input level the following steps are necessary. First, equation (83) is estimated and the estimated error terms, $\hat{\epsilon}_t$, are obtained. Next, a univariate distribution of the error terms is estimated. Finally, the yield distribution is obtained from the estimated error terms and a given value for x. To illustrate the procedure to obtain a distribution of yields consider the following simple example.

Consider a model whose objective is to find the profit maximizing levels of inputs subject to stochastic yields. Mathematically the model is

$$\max_{n} E(\pi) = \sum_{y} (p\,y - c\,n)\, f(y|n)\, dy \tag{84}$$

where $E(\pi)$ is expected profit, p is price, y is yield, f(y | n) is the conditional probability of yield given an input level n, and c is the cost of the input. For simplicity, assume only three levels of y are going to be considered along with only two levels of n. For yield the levels are 50, 75, and 100 units, whereas, the input levels to be considered are 120 and 320 units. Obtaining the probability function, f(y | n), is necessary to solve equation (84). A procedure to obtain estimates for this probability density function using an estimated univariate distribution for the error term associated with an estimated production function is outlined.

The first step is to estimate a production function. Assume the estimated production function is

$$\hat{y}_t = 20.0 + 0.25\ X_t. \tag{85}$$

The second step is to estimate an univariate pdf for the estimated error terms associated with the production function estimation. This pdf becomes the basis for the stochastic component in equation (84).

Given this production function, estimated deterministic yields for the two levels of the variable input are 50 and 100 units. With three levels of yield to be considered, 50, 75, and 100, the estimated errors associated with 120 units of the input are 0, 25, and 50 (50 - 50, 75 - 50, 100 - 50). To clarify when using 120 units of the input, the only way to achieve a yield of 50 is to have no error, but to achieve a yield of 75 units, a positive error of 25 units must be realized. To achieve a yield of 100 units a positive error of 50 units must be realized. The probability associated with each yield is the probability of realizing a particular error value.

Similarly, when 320 units of input are used the errors are -50, -25, and 0. Assuming a symmetric univariate distribution is fitted to the error terms (for example a normal error distribution), then the probabilities associated with the absolute value of each estimated error are identical. Let the estimated probabilities associated with the absolute value of each error be: 1) probability of an error of zero is .35 , 2) probability of an error of 25 units is .4, and 3) the probability of an error of 50 units is .25. In Table 7 the probability of each yield level is summarized conditional on a given input level. The probabilities clearly differ by input level. For example, a 35% probability is realized for a yield of 50 units when the input level is 120 units, whereas, with an input level of 320 units this probability changes to 25%.

Continuing the above example, for a price of \$2.00/unit and an input cost of \$0.20/unit, the expected profit becomes

$$\begin{aligned}
&\textit{for } n = 120 \\
&E(\pi) = .35[2(50) - .2(120)] + \\
&\qquad .40[2(75) - .2(120)] + \\
&\qquad .25[2(100) - .2(120)] \quad \textit{and} \\
&\textit{for } n = 320 \\
&E(\pi) = .35[2(50) - .2(320)] + \\
&\qquad .40[2(75) - .2(320)] + \\
&\qquad .35[2(100) - .2(320)].
\end{aligned} \tag{86}$$

Table 7. Example of Conditional Probabilities Based on Estimation of the Error Term Distribution.

Model Yield	Estimated Error Input Level[1]		Probability of Yield Given an Input Level[1]	
	120	320	120	320
50	0	-50	.35	.25
75	25	-25	.40	.40
100	50	0	.25	.35

1) Based on an estimated yield of 50 units when the input level is 120 and 100 units when the input level is 320.

Expected profit for an input level of 120 units is $121, whereas, for 320 units expected profit is $91. This procedure allows for "conditional" probabilities to be estimated using only univariate pdf estimation techniques. As such it has a computational advantage over other multivariate techniques. An additional advantage is that the procedure can be used with more than one independent variable in equation (83). Further, the basis of the procedure is the assumptions necessary to use OLS. Disadvantages are the techniques is applicable in only certain settings and a "true" conditional pdf is not estimated. Further, disadvantages are that only conditional pdf's are obtained and not joint pdfs. The same procedure can be used to estimate probabilities in which the observations are correlated. For the case of first order autocorrelation, equation (83) is revised such that y_t is the observed variable at time t, and x_t is the observed variable observed at y_{t-1}, that is lagged one time period.

Use of Hyperbolic Tangent to Obtain a Multivariate Distribution

To estimate conditional cdfs using Taylor's hyperbolic tangent transformation, only slight modifications of the procedure outlined for univariate cdfs are necessary (see earlier discussion). A cdf is termed "conditional" if its value is dependent on the value of some other independent variable. That is, there is a systematic relationship between the pdfs for different values of the independent variable. Taylor (1984) estimated cdfs for cotton and corn yield conditional on the rate of nitrogen application as an example of the use of the hyperbolic tangent for empirical estimation of conditional cdfs. In keeping with the previous notation, the relevant hyperbolic tangent transformation is:

$$F(x|y) = \frac{1}{2} + \frac{1}{2} \tanh [g(x, y)] \qquad (87)$$

where y is an independent variable or a set of independent variables. Taylor notes that including y in g(x, y) allows for a systematic relationship between the pdfs associated with different values of y (Taylor 1983). Interaction terms may also be specified between x and y to allow for changes in the basic form of the pdf for different y values.

The derivation of the MLE for the conditional cdf follows that of the univariate case. The log likelihood function is formed using the pdf associated with the cdf of equation (87). Differentiating the log likelihood function with respect to each element of the parameter

vector, δ, gives m first order conditions. Simultaneous solution of these equations gives parameter estimates. The log likelihood function is given as

$$\ln L(\delta) = n \ln \left(\frac{1}{2}\right) + 2 \sum_{i=1}^{n} \ln \{\operatorname{sech} [g(x,y)]\} + \sum_{i=1}^{n} \ln \left(\frac{\partial g(x,y)}{\partial x_i}\right) \tag{88}$$

with first order conditions

$$\frac{\partial \ln L(\delta)}{\partial \delta_j} = 0 \ ;$$

$$\sum_{i=2}^{n} \left[\frac{\partial (\partial g(x,y)/\partial x_i)}{\partial \delta_j}\right] \left[\frac{\partial g(x,y)}{\partial x_i}\right]^{-1} \tag{89}$$

$$= 2 \sum_{i=1}^{n} [\tanh\ g(x_i,y)] \left[\frac{\partial g(x_i,y)}{\partial \delta_j}\right] \quad \text{for } j = 1, \ldots, m.$$

As with the case of the unconditional cdf estimation, the conditional MLE cdf estimation requires good starting values for each element in δ. The numerical search procedure proposed by Taylor is capable of handling the conditional cdf estimation. OLS estimates are considered good starting values.

The procedure to find starting values for the unconditional cdf estimation can be used to obtain starting values for the parameters associated with the x's. A starting value of zero can be used for any coefficient associated with y. This procedure normally works, but the HT may "blow up" using this procedure. In this case, better starting values are necessary.

To apply OLS to determine the starting values for ML estimation, a further transformation of equation (87) is needed as was the case for the unconditional cdf. This transformation is

$$g(x,y) = \frac{1}{2} \ln \left[\frac{F(x|y)}{1 - F(x|y)}\right] \tag{90}$$

Again, a ranking procedure must be employed on the x's from smallest to largest and a cumulative frequency applied of the form $F(x_{iz}|y_{iz}) = i/n$, where z is the zth value of y. Then using equation (90) a transformed data set of the form (g_{iz}, x_{iz}, y_{iz}) can be attained. OLS can be applied directly to the modified data set to obtain the starting values for the MLE procedure. Once starting values are obtained, equation (89) is solved to obtain the parameter estimates of the conditional cdf.

Correlation of Random Variables in Simulation

The procedure for "appropriately" correlating random variables in a simulation model was demonstrated for agricultural models in 1971 by Clements, Mapp, and Eidman. They demonstrated the procedure for a multivariate normal distribution and suggested that it could be "easily" expanded to the nonnormal case. This challenge was achieved several years later by two independent research teams, Richardson and Condra (1978, 1981) and King. Both the multivariate normal case and the multivariate nonnormal cases are presented here in a simulation framework. The two sections are not independent and should be read in order. The concluding section presents a description of how this procedure is used in a whole farm simulation model, FLIPSIM.

Multivariate Normal: In simulation models, it is often advisable to correlate the random variables. Law and Kelton and Reutlinger refer to correlation of random variables as a variance reduction technique. To assume independence for random variables that are negatively (positively) correlated will over (under) state the variance in the system (Law and Kelton).

Simulation models developed using econometric equations generally assume the error terms for the individual equations are normally distributed to facilitate the use of various statistical tests. As a result, these models generally assume the error terms are normally distributed for simulation. For a four-variable case the model may look like

$$\begin{aligned} y_1 &= a_1 + b_1 x_1 + e_1 \\ y_2 &= a_2 + b_2 x_2 + e_2 \\ y_3 &= a_3 + b_3 x_3 + e_3 \\ y_4 &= a_4 + b_4 x_4 + e_4 \end{aligned} \qquad (91)$$

where y_1 through y_4 are four dependent variables explained by the four non-stochastic variables x_1, x_2, x_3, x_4, and the e_i's represent the unexplained portion (error terms) of each equation.

A deterministic simulation of the four-equation model would involve setting all of the e_i's to zero and calculating y_i's with alternative x's. A stochastic simulation of the model assuming independent normally distributed error terms results in the following formulation

$$\begin{aligned} y_1 &= a_1 + b_1 x_1 + (\hat{\gamma}_1\ SND_1) \\ y_2 &= a_2 + b_2 x_2 + (\hat{\gamma}_2\ SND_2) \\ y_3 &= a_3 + b_3 x_3 + (\hat{\gamma}_3\ SND_3) \\ y_4 &= a_4 + b_4 x_4 + (\hat{\gamma}_4\ SND_4) \end{aligned} \qquad (92)$$

where $\hat{\gamma}_i$ is the estimated standard deviation for the e_i's and SND_i represents a randomly generated standard normal deviate with mean zero and standard deviation one (N(0, 1)). The SND's are independent, therefore, the y_i's are independent. A large sample of y_i's must be generated using a different set of SND_i's for each realization or iteration.

For correlated error terms equation (92) must be modified. To correlate the error terms in the four variable model it is necessary to first calculate the covariance matrix for the error terms in equation (91). This is easily done using SAS or another statistical package, but one must be careful to not change the temporal order of the error terms. The second step is to take the square root of (factor) the covariance matrix (see Clements, Mapp, and Eidman). (The Choleski decomposition in SAS provides a procedure for factoring a covariance matrix.) The factored matrix is analogous to using the simultaneously determined standard deviation for the system in place of the single standard deviates in equation (92). Rewriting equation (92) in matrix notation for stochastic simulation yields:

$$Y = A + BX + \sum^{-1/2} S \qquad (93)$$

where Y is a 4 x 1 vector of dependent variables, A is a 4 x 1 vector of intercepts, B is a 4 x 4 diagonal matrix of slope coefficients, X is a 4 x 1 vector of exploratory variables x_i, $\Sigma^{-1/2}$ is a 4 x 4 square root of the upper right triangle covariance matrix, and S is a 4 x 1 vector of independent standard normal deviates, N(0, 1). The model in equation (93) is simulated a large number of times, say, 100 iterations by generating 100 sets of SND's for S and solving for y's. The simulated y_i values will be correlated approximately the same as the $\hat{e}_i$'s in equation (91). Clements, Mapp, and Eidmon refer to this as "appropriately correlated" random variables.

Multivariate Nonnormal: In the case of models where the error terms are not normally distributed the errors can be correlated by using the following procedure.

- Calculate the error terms for the stochastic variables. This involves using the error terms from estimated equations such as in equation (91) or from another process.
- Calculate the correlation matrix (P) for the error terms and factor the upper triangle correlation matrix using the square root method such that

$$R = P^{-1/2}. \qquad (94)$$

- Generate an independent standard normal deviate (D) for each equation in the model, if there are four random variables such as the model in equation (92) four independent standard normal deviates are necessary.
- Correlate the independent standard normal deviates using

$$C = R\ D. \tag{95}$$

- Convert the correlated standard normal deviates in C to correlated uniform deviates using a Z-table look up function or a standard normal integration function (ERFF is one such function)

$$U = Z_{table}\ C. \tag{96}$$

- Use the correlated uniform random numbers in U and the cdf generating functions for each of the random variables to simulate correlated values for the y's. Because the U's and C's are correlated they can be used in any type of pdf; for the four equation example, e_1 can be normally distributed (so use c_1), e_2 can be empirical (so use u_2), e_3 can be exponential (so use u_3), while e_4 can be distributed gamma (so use u_4). The resulting y's will be appropriately correlated as they were over the estimation period.

Richardson and Condra (1978) showed that this procedure is a generalized form of the multivariate normal distribution in equation (93). This result occurs because the covariance matrix (Σ) can be expressed as the product of a diagonal standard deviation matrix (γ) and the correlation matrix (P):

$$\Sigma = \gamma P \tag{97}$$

Rewriting equation (93) in terms of γ and P yields

$$Y = A + B\ X + \gamma_{ii}\ P\ S \tag{98}$$

which explicitly recognizes the intuitive notation that it is the correlation coefficients that correlate error terms not the standard deviations in the covariance matrix.

FLIPSIM Method: A brief description of the FLIPSIM model is required so that the reader will understand how to generate empirical distribution functions for the stochastic variables within the model. FLIPSIM is an annual whole farm level simulation model that analyzes the impacts of farm program and tax policies on a representative farm. The model simulates the production activities for a farm throughout the year. Representative farms are created from the interaction of a panel of

farmers from a particular area. The panel members are asked to provide descriptive, financial and production data to be used as input into FLIPSIM. The reader is referred to "Description of FLIPSIM V: A General Firm Level Policy Simulation Model" by Richardson and Nixon for further documentation of the FLIPSIM model.

Empirical probability distribution functions are utilized to incorporate price and yield risk during the farm simulations. Farmers are asked to provide 10 years of historical price and yield data for each crop. This data is then used to generate a multivariate empirical probability distribution from which the simulation model will draw random values and in effect, simulate price and yield risk. Richardson cites five conditions that facilitate or require using the empirical distribution function:

- When there is little or no data,
- When the random event has finite tails,
- When the minimum or maximum must be observed,
- If the data are lumpy, and
- If you observe discrete values but the event is continuous.

In farm level analyses, there is typically very little data available at the farm level. Also, data is normally lumpy (e.g., yields may be bimodal) and discrete values are observed when the distribution is in fact continuous.

The multivariate J-Distribution for continuous empirical distribution is one of the procedures used in FLIPSIM to simulate stochastic prices and yields for crops. The model allows for a maximum of 20 crops so 40 random variables (yields and prices for 20 crops) are generated by the model. The model recursively simulates up to 10 years using exogenous annual yields and prices as the means of their respective distributions. Because these means can trend up or down over the planning horizon and thus make a model non-covariance stationary, a procedure was developed to specify the empirical distributions. The procedure guarantees that the random variable in years 1 and 10 will have the same relative variability (coefficient of variation) even though the mean may have doubled. (In contrast, a multivariate normal distribution will reduce the risk over the planning horizon if the means increase).

The steps to specify the multivariate empirical probability distribution using 10 or more observations are as follows:

- Estimate the error terms for the random variables by detrending each of the y_i variables

$$\hat{y}_t = \hat{a} + \hat{b}\, t + \hat{\epsilon}_t. \tag{99}$$

- Calculate the correlation matrix (P) using the unsorted $\hat{\epsilon}_t$ values.
- Calculate the factored upper right triangle correlation matrix (R) by factoring P

$$R = P^{-1/2}. \tag{100}$$

- Convert the error terms for each of the equations represented by equation (99) to percentage changes from the predicted values or

$$E_t = \frac{\hat{\epsilon}_t - \hat{y}_t}{\hat{y}_t} \tag{101}$$

- Sort the percent deviate error terms, E_t, in equation (101) from low to high and create a J-Distribution (see earlier section). The sorted E_t are placed in the X^i vector. Because there are 40 potentially random variables there are 40 different X^i vectors in the X matrix:

$$x^i(1),\ x^i(2),\ ...,\ x^i(10),\ x^i(11),\ x^i(12) \tag{102}$$

This complete the steps to develop a covariance stationary multivariate J-Distribution for a continuous emperial distribution.

During simulation, FLIPSIM combines the components for the multivariate distribution (R and X) with annual mean yields and prices (y_{ti}) and independent standard normal deviates (D) to generate multivariate empirical yields and prices. The steps followed for each year of the planning horizon to generate stochastic yields and prices are:

- Generate 40 independent standard normal deviates (D_{40x1}) using GAUSS (See Appendix B for a listing of GAUSS).
- Correlate the independent standard normal deviates

$$C = R_{40x40}\ D_{40x1}. \tag{103}$$

- Convert the correlated standard normal deviates in C to correlated uniform deviates using a table look up function for the Z table.

$$U_{40x1} = Z_{table} C_{40x1} \tag{104}$$

- Use the correlated standard normal deviates (U) and the sorted percent deviates error terms E_t in X to calculate stochastic percent deviates (T) by following the steps for the J-Distribution:

  Let m = 12, u = 1/12(μ_i) and I = K + 1, then

$$T_i = x^i_{(2)} \qquad if\ \mu_i \le \frac{1}{m} \tag{105}$$

$$T_i = \frac{\left(\mu_i - \frac{k}{m}\right)}{\frac{1}{m}} x^i_{(l)} - x^i_{(k)} + x^i_{(k)} \qquad if\ \frac{K}{M} < \mu_i \le \frac{I}{m} \tag{106}$$

$$T_i = x^i_{(m-1)} \qquad if\ \mu_i \ge \frac{11}{12}. \tag{107}$$

- Estimate the stochastic yields and prices (Y) using the stochastic percent deviates (T) and the means (M) for the current year t:

$$Y_{it} = M_{it}\ (1 + T_i). \tag{108}$$

The resulting Y_{it} values are appropriately correlated within each year t based on the correlation coefficients in P. Tests of stochastic yields and prices reveal that the means of the simulated values are statistically equal to the means in the M_t and the correlation coefficients for yields and prices are similar to those in P. Additionally, the coefficients of variation are constant over the planning horizon regardless of the trend on yields and prices. Because the J-Distribution is used the actual minimum and maximum values are observed in simulation with about a 10 percent frequency, recall only 10 observations are used.

USE OF PROBABILITY DENSITY FUNCTIONS IN MODELING

The primary objective in estimating most density functions is to use the function in a modeling effort. This section discusses two of the more common uses of density functions in modeling. The first section discusses the use of the density function in optimization models, while the second addresses the use of such density functions in simulation models.

Optimization Modeling

Stochastic profit maximization in its simplistic form can be expressed as

$$\max_{w} \pi = \int_{0}^{\infty} [\, p\, y(w \mid x) - c(w) \,]\, f(x)\, dx \qquad (109)$$

where p is output price, y is the production process which is dependent on input w and random event x, c(w) is the cost function, and f(x) is the pdf associated with x. In simple models where the integral of f(w) can be taken, the maximization of equation 109 is straight forward. Problems in using equation 109 occur in most applied models, because of their size, data availability, difficulty in evaluating the pdf, and objective of the research. In most applied models of the form given in equation (109), a discretizing of the variables in the model occurs to provide approximation of the continuous solution.

Discretizing a variable is simply taking the range of the variable and dividing it into m intervals. These m intervals are used to represent the continuous process. Although not required, common practice is to make the m intervals equal in size. In fact, in many models it may be advantageous to have unequal interval sizes. One may wish to have large intervals away from the optimal area, and smaller intervals around the optimal. Such a discretizing scheme may increase the accuracy of the model. Usually a single point within the interval is used to represent the interval. Commonly used points are either the interval midpoint or endpoint. What is of interest here is how to obtain the probabilities associated with these intervals. Larger m's provide an increase in accuracy of the model at the expense of increased computational costs. Further, as m goes to infinity, the continuous case is obtained.

An important consideration in calculating the probabilities associated with the intervals is the probabilities of the individual intervals must sum to one. If the probabilities do not sum to one, results from the model will be erroneous. Consider the two pdfs, normal and gamma. The lower bound on the normal is negative infinity, whereas, for the gamma the lower bound is zero. For the normal distribution, the lower bound for the first interval must be changed to negative infinity.

Whereas for the gamma the lower bound is zero. Such differences between density functions must be taken into account when calculating probabilities associated with intervals.

The probability associated with each interval is obtained using a pdf as follows

$$P_1 = \int_{I_0}^{I_1} f(x)\,dx \quad for\ i = 1,$$
$$P_i = \int_{I_i}^{I_{i+1}} f(x)\,dx \ for\ i < I_T,\ and \tag{110}$$
$$P_T = \int_{I_{I_T-1}}^{I_T} f(x)\,dx \ for\ i = I_T$$

where P_i is the probability associated with the ith interval, f(x) the pdf, I_0 the lower bound associated with the pdf f(x), I_i is the bounds of the intervals, and I_T is the upper bound associated with pdf f(x). Usually, I_0 is either zero or negative infinity, whereas, I_T is positive infinity. The use of equation (110) is complicated by the fact that integration is involved. For many functions, numerical integration is necessary as no closed form for the integral exists.

Numerical integration of a single variable function is not difficult, but does require some time. Many software packages exist that can numerically integrate a function. Some can even symbolically integrate some functions. Programming numerical integration techniques in languages such as FORTRAN or C is not difficult. One such approximation for numerical integration is Simpson's method. This method is piecewise interpolation method (Sedgewick) given by

$$s = \sum_{1 \le j \le N} \left[\left(\frac{x_{j+1}}{6}\right)(f(x_j) + 4f\left(\frac{x_j + x_{j+1}}{2}\right) + f(x_{j+1}))\right] \tag{111}$$

where s is the integral amount, $f(x_j)$ is the function to be integrated, x_j are endpoint values on x, associated with the various intervals, and N the total number of intervals or iterations to be considered. It needs to be noted that the x_j's in equation (111) are not the x_i's discussed earlier. Two intervals are being considered. First the intervals that are considered for the random event, the x_i's. These are as discussed previously. The x_j's are intervals used in the numerical integration within each interval of the random variable. To clarify, suppose you want to find the probability of being in the range of 10 to 15 for some pdf. Numerical integration divides this range into small subintervals of say .005 [this case N = 1000, that is (15 -10) /1000 =.005]. Using this interval width of .005, equation (111) is evaluated 1000 times to find s, the probability of being in the range 10 to 15. In other words, numerical integral simply divides the range to be integrated into very small subintervals

and sums the areas under the curve calculated from the function itself. Equation (111) is then applied to each interval of the random variable being considered. Obviously, infinity can not be used in equation (111), but it is replaced by a sufficiently large number. This short digression on numerical integration is to illustrate that the procedure is not as hard as it sounds. The need to integrate a pdf should not preclude the use of a particular pdf. Interested readers should consult the multitude of literature concerning this subject.

Probabilities associated with intervals are obtained in a slightly different manner when using a cdf. Using a cdf, probabilities associated with intervals are obtained as follows

$$\begin{aligned} P_1 &= F(x_1), \\ P_i &= F(x_i) - F(x_{i-1}) \quad for \quad 1 < i < I_T, \\ P_{I_T} &= 1.0 - F(x_{I_{T-1}}), \end{aligned} \tag{112}$$

where P_i is the probability associated with the interval, $F(x_i)$ is the cdf, and $x_{I_{T-1}}$ is the upper bound of the next to last interval to be considered. Equation (112) forces the probabilities associated with the intervals to sum to one. The two tricks used are associated with the first and last intervals as with the pdf. The first interval's probability goes from the lower bound of the distribution to the x value at the boundary between the first and second intervals. The second trick is involves using one as the upper bound for the cdf. Probability of the last interval is found by subjecting the probability of being in any of the intervals other than the last interval from one.

Aquifer Example: Probabilities associated with various intervals of the aquifer data are found using several of the previously estimated density functions. Density functions used are the normal, gamma, exponential, empirical, hyperbolic tangent and histogram. Probabilities associated with histogram are found using the interval size discussed later. Numerical integration was used to find the probabilities associated with the normal and gamma, whereas, the estimated cdfs associated with the remaining three distributions where used. Equation (55) is used for the empirical distribution. For the normal distribution, the probabilities could also have been found using a standard normal probability table. In this example, the range of recharge levels is divided into 10 equally spaced intervals ranging from zero to 2003.5 acre feet. Interval size is 200.35 acre feet [(2003.5 - 0.0)/10]. The first and last intervals where changed appropriately to insure the probabilities sum to one. For the first interval either negative infinity or zero was used as the lower bound depending on the distribution. Finally for the last interval, positive infinity was used as the upper bound.

Probabilities associated with each interval for the various distributions are given in Table 8. Differences exists between the various distributions. These differences are most noticeable in the left-

Table 8. Probability of Aquifer Recharge Ranges for Various Density Functional Forms.

Interval	Probability Distribution					
	Normal	Gamma[1]	Exponential	Empirical	Hyperbolic	Histogram
	Percent					
0 to 200.35	15.15	11.17	27.04	18.19	10.75	18.18
200.35 to 400.7	13.62	22.83	19.73	12.72	21.74	12.72
400.7 to 601.05	18.04	21.65	14.40	25.11	24.48	23.63
601.05 to 801.4	18.36	16.34	10.50	14.27	16.53	14.55
801.4 to 1001.75	15.61	11.04	7.66	15.07	9.53	14.55
1001.75 to 1202.1	10.21	6.99	5.59	6.39	5.90	7.30
1202.1 to 402.45	5.57	4.23	4.08	2.36	4.21	1.82
1402.45 to 1602.8	2.34	2.49	2.98	3.20	3.17	3.64
1602.8 to 1803.15	.81	1.43	2.17	1.45	2.13	1.82
1803.15 to 2003.5	.29	.81	5.86	1.25	1.56	1.82

1) Using the moment matching technique to obtain parameter estimates for the gamma distribution.

hand tail of the distributions. As expected, the exponential distributions deviates the most from the other three functions. The gamma and the hyperbolic tangent function are similar, whereas, the histogram and empirical are similar.

Application in Simulation

One application of density functions in research is in Monte Carlo simulation. A mathematical representation of the system being modeled is developed from primary and secondary data. Stochastic variables affecting the system are included by substituting probability distributions into the model for these stochastic variables. During simulation, values for these stochastic variables are selected at random from their pdfs and used in the model. The model is simulated a large number of times (iterations) to develop a sufficient sample of the simulated output variables. Stochastic simulation is used to empirically estimate probability distributions for key output variables in the system that cannot be (a) readily observed or measured, (b) solved for analytically, or (c) both.

A Monte Carlo simulation model has one or more random variables. Parameters to define the pdfs for these random variables can be estimated using the techniques described in the first sections of this report. The more accurately the estimated pdfs represent the stochastic variables in the model, the more accurately the model will estimate the target (or desired) output variable's pdf.

The purpose of this section is to present the results of simulating the aquifer recharge pdf using the estimated alternative distributions. Formulas used to simulate the different distributions are described briefly in the text and the FORTRAN computer code for simulating these distributions is presented in Appendix B. The statistical results from simulating each of the assumed distributions are presented in tabular and graphical form so the reader can examine how each assumed distribution fits the aquifer recharge data.

Normal Distribution: The normal distribution is totally defined by its mean and variance. The mean of the aquifer recharge data is 635.52 and the standard deviation is 422.55. One of the problems with simulating a normal distribution is that its minimum is defined by negative infinity. The statistics in Table 9 summarize the results of simulating the aquifer recharge data assuming the normal distribution is not truncated and that it is truncated at zero. Both normal distributions were simulated for 100, 200, 500, and 1,000 iterations to demonstrate the affects of sample size on the simulated results.

The minimum recharge values are -248 to -787 for the untruncated normal pdf as the sample size increases from 100 to 1,000 (Table 9). About 6 percent of the simulated values are less than zero for sample sizes of 500 and less. The sample mean for the three distributions with 200 to 1,000 iterations are about equal to the population mean (635.52 vs. 632.69, 634.71, and 636.09). The standard deviations for

all four of the simulated normal series are less than the population value. The Fortran programming code to simulate the untruncated random values is:

```
NAR = 31415
DO 100 I = 1, NOITER
CALL GAUSE (GAS, NAR)
VAL(I) = 635.52 + 422.55*GAS
100     CONTINUE
```

The variable NAR is the Gaussion random number seed and is used by the subroutine GAUSE which generates standard normal deviates. A seed value must be used that is "good" in the sense that it produces a non-reporting sequence of numbers. The seed value of 31415 has been found, through extensive use, to be a "good value." The words GAS and NAR pass values to the Gause subroutine to generate and retrieve the random variable. The fourth line of the code use the mean (635.52), standard deviation (422.55), and the random generated standard normal deviate (GAS) to generate the random value. See the computer code in TEST-DIST in Appendix B for the code used to generate the samples reported in Table 9.

The normal distribution can be truncated at zero by adding two statements as follows:

```
        NAR = 31415
        DO 100 I = 1, NOITER
90      CONTINUE
        CALL GAUSE (GAS, NAR)
        VAL(I) = 635.52 + 422.55*GAS
        IF (VAL(I).LE.0.0), GO TO 90
100     CONTINUE
```

The results of truncating the aquifer recharge data at zero and simulating it for 100, 200, 500, and 1,000 iterations are summarized in Table 9 and Figure 3.1. As expected, the minimum values become positive, the mean has been increased, and the standard deviation reduced by truncating the distribution at zero. Assuming the data are normally distributed would result in never simulating the maximum observed value of 2003.5 and observing six percent of the values less than zero, unless a truncated normal was used. The computer code to generate the truncated normal sample is included in the TEST-DIST program in Appendix B.

Gamma Distribution: Three estimates of the alpha and beta parameters for the Gamma distribution from Table 3 were used to simulate the data. The simulation results for 500 interactions are summarized in Table 10 and Figure 32. All three methods resulted in means that are less than 635.52, although the Sinha MLE approximation resulted in the "best" mean. The standard deviation for Sinha MLE approximation (421.16) was about the same as the original data. The maximum simulated values (2268, 2450, and 2281) for three of the samples were in excess of the 2004 observed from the parent distribution.

Table 9. Statistics and Interval Observations for Simulated Aquifer Recharge, Assuming Normality for Alternative Sample Sizes.

	+- Normal				Truncated Normal			
	Interation							
Interval	100	200	500	1,000	100	200	500	1,000
	Percent							
< 0	6	5.5	5.8	2.4	0	0	0	0
0-200	7	6.5	8.8	2.7	9	7	8.8	9.1
200-400	11	15.5	14.2	13.5	11	15.5	14.8	14
400-600	20	20.5	18.4	19.3	22	21.5	19.8	20.3
600-800	23	21	16.6	18.2	24	22	18.2	19.2
800-1000	11	9.5	15.2	15	11	10.5	16.2	15.7
1000-1200	15	12.5	12	11.3	15	13.5	12.4	12.1
1200-1400	5	7	7.2	6	6	8	7.6	6.6
1400-1600	2	2	1.8	2.4	2	2	2.2	2.5
1600-1800	0	0	0	0	0	0	0	0.3
1800 >	0	0	0	1	0	0	0	0.1
	Statistics							
Mean	644.46	632.69	634.71	636.09	687.56	687.61	690.2	690.6
Std. Dev.	397.19	401.26	412.02	408.31	355.9	360.43	362.46	365.99
Minimum	-248.93	-572.39	-572.39	-787.78	1.66	1.66	1.66	0.22
Maximum	1580.68	1580.68	1580.68	1862.75	1580.68	1580.68	1580.68	1862.76
N	100	200	500	1000	100	200	500	1000

The FORTRAN code used to simulate the Gamma distributions is included in Appendix B. The procedure used in TEST-DIST comes from Naylor and required converting the alpha and beta values from Table 3 prior to simulation (see the comment lines).

Beta Distribution: Using the Beta parameters in Table 4, estimated by the moment estimation procedure, the aquifer recharge data was simulated for four different sample sizes (Table 11 and Figure 33). All four of the samples produced means that were far below the original mean of 635.52 and standard deviations far below the original value of 422.55. The minimum simulated values were less than 4.0 and the maximum did not exceed 1,825 even for sample size of 1,000.

The equations used to simulate the Beta distribution are included in Appendix B as part of TEST-DIST. The procedure was suggested by Law and Kelton.

Exponential Distribution: Four sample sizes were simulated for the exponential distribution (Table 12 and Figure 34). Larger sample sizes resulted in a mean that was closer to the original value (646 vs. 635). The exponential distribution resulted in a much larger standard deviation (632 to 771) than the original data.

The FORTRAN code to simulate random numbers for an exponential distribution is included in TEST-DIST (Appendix B). As the code indicates, the exponential distribution is very simple and probably explains why it is often selected. The results in Table 12 suggest, however, that the exponential distribution should be tested prior to making it a part of any simulation model.

Cumulative Empirical Distribution: The aquifer recharge data were simulated as a continuous empirical distribution, assuming 55 intervals and a mean of 635.52 (Table 13 and Figure 35). The results of the simulation suggest that the continuous empirical distribution would not reproduce the actual minimum and maximum values but would do a reasonable job of reproducing the mean at higher sample sizes. The procedure, however, generated a smaller standard deviation (371 to 381) than the original data (422). This is due, in part, to not achieving extreme values similar to the original distribution.

The FORTRAN program used to develop the continuous empirical probability distribution samples is named CDIST and is included in Appendix B. CDIST is capable of simulating from univariate or multivariate continuous empirical distributions.

J-Distribution: The aquifer recharge data was simulated using the J-Distribution for four sample sizes (Table 14 and Figure 36). The mean used to simulate the J-Distribution was 635.52. The simulations resulted in means ranging from 632 to 643 and standard deviations of 406 to 423. Both the means and standard deviations are closer to the original distribution than observed for the continuous empirical distribution (Tables 13 and 14). Sample sizes 500 and 1,000 produced superior results based on the means

Table 10. Statistics and Interval Observations for Simulated Aquifer Recharge, Assuming a Gamma Distribution with a Sample Size of 500.

	Wilk	Sinha	Matching
Interval		Percent	
< 0	0	0	0
0-200	14.4	12.6	14.6
200-400	26.2	24.6	26.4
400-600	20	19.6	20
600-800	16.4	17.2	16
800-1000	10.2	10.8	10.4
1000-1200	4.6	5.4	4.6
1200-1400	3.8	7	3.6
1400-1600	2.2	0.8	2.2
1600-1800	1.2	1	1.2
1800 >	1	1	1
		Statistics	
Mean	571.28	616.97	574.41
Std. Dev.	389.91	421.16	392.11
Minimum	6.98	7.54	7.02
Maximum	2268.75	2450.2	2281.19
N	500	500	500
ALPHA	2.26	2.09	2.25
BETA	280.95	303.42	282.49

Table 11. Statistics and Interval Observations for Simulated Aquifer Recharge, Assuming Beta Distribution for Alternative Sample Sizes.

	Interation			
Interval	100	200	500	1000
		Percent		
< 0	0	0	0	0
0-200	20	27.5	27.6	28.4
200-400	22	24.5	23.6	20.3
400-600	21	17.5	16.6	17.4
600-800	17	12.5	12	12.2
800-1000	10	8.5	8.8	8.8
1000-1200	4	3.5	5.8	6.8
1200-1400	4	3.5	3.8	4.1
1400-1600	2	2	1.6	1.5
1600-1800	0	0.5	0.2	0.4
1800 >	0	0	0	0.1
		Statistics		
Mean	528.95	473.99	483.95	497.69
Std. Dev.	342.43	363.11	366.39	380.47
Minimum	3.73	3.73	0.97	0.97
Maximum	1494.97	1722.37	1722.37	1824.58
N	100	200	500	1000
p	1.5	1.5	1.5	1.5
q	3.36	3.36	3.36	3.36
a	0	0	0	0
b	2003.5	2003.5	2003.5	2003.5

Table 12. Statistics and Interval Observations for Simulated Aquifer Recharge, Assuming Exponential Distribution for Alternative Sample Sizes.

	Interation			
Interval	100	200	500	1000
		Percent		
< 0	0	0	0	0
0-200	28	25.5	23.8	26.3
200-400	11	20	16	18.2
400-600	15	13	16.4	14.1
600-800	12	10	12.6	12.1
800-1000	8	6.5	7.8	8.6
1000-1200	5	6	6	5.9
1200-1400	2	3.5	3.8	3.4
1400-1600	4	5	4.2	3.1
1600-1800	2	3	2.6	2.6
1800 >	13	7.5	6.8	5.7
		Statistics		
Mean	767.55	680.34	693.25	646.13
Std. Dev.	770.98	663.1	658.24	632.44
Minimum	2.04	2.04	0.01	0.01
Maximum	3570.4	3570.4	3640.93	5055.47
N	100	200	500	1000

Table 13. Statistics and Interval Observations for Simulated Aquifer Recharge, Assuming Continuous Empirical Distribution for Alternative Sample Size.

	Interation			
Interval	100	200	500	1000
		Percent		
< 0	0	0	0	0
0-200	15	14	16.6	16.2
200-400	9	13.5	12.4	11.7
400-600	33	32	38	27.8
600-800	11	11.5	13.8	14.2
800-1000	18	12.5	15.6	14.5
1000-1200	8	8	8	7.8
1200-1400	1	2	2.8	2.4
1400-1600	3	4	3.6	3.3
1600-1800	1	1.5	1.4	1.8
1800 >	1	0.5	0.2	0.3
		Statistics		
Mean	632.85	621.32	626.15	624.9
Std. Dev.	371.37	380.54	381.92	381.57
Minimum	137.31	55.07	55.07	45.8
Maximum	1801.96	1801.96	1801.96	1874.45
N	100	200	500	1000

Table 14. Statistics and Interval Observations for Aquifer Recharge, Assuming a J-Distribution for Alternative Sample Sizes.

	Interation			
Interval	100	200	500	1000
	Percent			
< 0	0	0	0	0
0-200	16	15.5	18.2	17.7
200-400	10	14	11.8	10.8
400-600	31	30	24.4	26.9
600-800	11	11.5	13.6	14
800-1000	16	11.5	15.2	14.2
1000-1200	6	7	6.4	6.3
1200-1400	4	2.5	3.2	3.2
1400-1600	4	6	4.2	3.4
1600-1800	0	0	1.4	1.2
1800 >	2	2	1.6	2.3
	Statistics			
Mean	643.41	632.97	640	639.36
Std. Dev.	406.84	416.99	421.79	423.25
Minimum	44.62	43.72	43.72	43.72
Maximum	2003.48	2003.48	2003.48	2003.48
N	100	200	500	1000

and standard deviations. The results of the simulation reproduced the minimum and maximum values in the original data, as the J-Distribution was designed to do.

The FORTRAN code for the program to simulate the J-Distribution is included in JDIST in Appendix B. The code is capable of simulating both a univariate and a multivariate J-Distribution.

Hyperbolic Tangent Distribution: The hyperbolic tangent F(x) distribution estimated for the aquifer recharge data in (82) was solved in a simulation mode using FORTRAN program SOLVE in Appendix B. The results of the simulations are summarized in Table 15 and Figure 37. The mean and standard deviations (635.99 and 419.20, respectively) are almost identical to the original data. The minimum of 5.4 is slightly less than the original minimum, but the maximum (2546.26) is quite a bit larger than the actual data (2003).

Simulation Summary

The simulation results demonstrate that some distributions are better suited to simulate the aquifer recharge distribution than others. These results suggest that each distribution should be tested with different specifications prior to selecting the final one for inclusion in a simulation model. A graph of the simulated distributions for N equals 500 is presented in Figure 16.

Table 15. Statistics and Interval Observations for Simulated Aquifer Recharge, Assuming Hyperbolic Tangent Distribution for Alternative Sample Size.

	Interation			
Interval	100	200	500	1000
		Percent		
< 0	0	0	0	0
0-200	16	14.5	11.6	14
200-400	20	20.5	23	34
400-600	26	22	27.2	23.2
600-800	8	17.5	13.8	13.8
800-1000	12	11	9.4	4.9
1000-1200	2	2	3.8	4.8
1200-1400	5	3.5	3.6	3.2
1400-1600	5	4	3.2	1.4
1600-1800	3	3.5	3	0.6
1800 >	3	1.5	1.4	0.1
		Statistics		
Mean	639.6	627.4	617.2	635.99
Std. Dev.	479.9	437.7	427.2	419.2
Minimum	6.8	6.8	5.4	5.45
Maximum	2055.8	2055.8	2546.3	2546.27
N	100	200	500	1000

REFERENCES

Axum, Technical Graphics and Data Analysis. TriMetrex, Inc. 444 NE Ravenna Boulevard, Suite 210, Seattle, WA 98115 ←

Ayres, F. *Differential and Integral Calculus*. *Schaum's Outline Series*. McGraw-Hill Book Company, New York. Second Edition, 1978.

Bain, L. J. *Statistical Analysis of Reliability and Life-Testing Models Theory and Methods*. Marcel Dekker, Inc. New York. 1978.

Beckman, R. J., and G. L. Tietjen. "Maximum Likelihood Estimation for the Beta Distribution." *J. of Statistical Computer Simulation*. 7(1978):253-258.

Boyd, D. W., and M. J. Steele. "Lower Bounds for Nonparametric Density Estimation Rates." *The Annals of Statistics* 6(1978):932-34.

Brownlee, J. *Pearson Karl: Tracts for Computers, N o IX, Tables of Logarithms of the Complete Gamma Function of x, from x=1 to 50.9 by Intervals of 0.01*. Cambridge University Press, 1923.

Bullock, K. D. "Mean Square Error Properties of Density Estimates." *The Annals of Statistics*, 3(1975):1025-1030.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical Methods for Data Analysis*. Duxbury Press, Boston, 1983. 395 pp.

Clements, A. M., Jr., H. P. Mapp, Jr., and V. R. Eidmon. "A Procedure for Correlating Events in Farm Firm Simulation Models." Technical Bulletin T-131. Oklahoma Agricultural Experiment Station. 1971.

Day, R. "Probability Distributions of Field Crops." *J. of Farm Econ*. 47(1958):733-740.

Devroye, L., and C. S. Penrod. "The Consistency of Automatic Kernal Density Estimates." *The Annals of Statistics* 12(1984):1231-1249.

Durbin, J. *Distribution Theory for Tests Based on the Sample Distribution Function*. Society of Industrial and Applied Mathematics, Philadelphia, 1973.

El harrack, A. "Estimation of Parametric and Nonparametric Probability Density Functions for U.S. Annual Wheat Prices." Unpublished M.S. Professional Paper. Texas A&M University. 1992.

Freund, J. E., R. E. Walpole. *Mathematical Statistics*. Prentice-Hall, Inc. Englewood Cliffs. Fourth Edition, 1987.

Fryer, M. J. "A Review of Some Non-parametric Methods of Density Estimation." *J. Inst. Maths Applic*. (1977):335-354.

Granger, C. W. J., and P. Newbold. *Forecasting Economic Time Series*. Academic Press, Inc. New York. 1977.

Grice, J. V. and L. J. Bain. "Inferences Concerning the Mean of the Gamma Distribution." *Journal of the American Statistical Association*, 75(1980):929-33.

Hastings, N. A. J., and J. B. Peacock. *Statistical Distributions*. John Wiley and Sons. New York. 1975.

Hogg, R. V., and A. T. Craig. *Introduction to Mathematical Statistics*. Macmillan Publishing Co. Inc. New York. Fourth Edition, 1978.

Johnson, N. L. and S. Kotz. *Continuous Univariate Statistics - 2*. Houghton Mifflin. Boston. 1970.

Judge, G., R. Hill, W. Griffiths, H. Lütkepohl, T. Lee. *Introduction to the Theory and Practice of Econometrics*. John Wiley and Sons. New York. 1982.

Kendall, M. G., and A. Stuart. *The Advanced Theory of Statistics*. Hafner. New York. Third Edition, 1969.

Kenkel, P. L., J. C. Buzby, and J. R. Skees. "A Comparison of Candidate Probability Distributions for Historical Yield Distributions." Paper Presented at S. Ag. Econ. Assoc. Meeting. Feb. 1991.

King, R. P. "Operational Techniques for Applied Decision Analysis Under Uncertainty." Department of Agricultural Economics, Michigan State University, unpublished Doctoral Thesis. 1979.

Kmenta, J. *Elements of Econometrics*. Macmillan Publishing Company. New York. 1986.

Kraft, C. H. and C. V. Eeden. *A Nonparametric Introduction to Statistics*. The Macmillan Company, New Yourk, 1968.

Law, A. M., and W. D. Kelton. *Simulation Modeling & Analysis*. McGraw-Hill, Inc. New York. Second Edition, 1991.

Mjelde, J. W., B. L. Dixon, and S. T. Sonka. "Estimating the Value of Sequential Updating Solutions for Intrayear Crop Management." *West. J. of Ag. Econ.* 14(1989):1-8.

Naylor, T. H. *Computer Simulation Experiments With Models of Economic Systems*. New York: John Wiley & Sons, Inc., 1971.

Nelson, C. H. "The Influence of Distributional Assumptions on the Calculation of Crop Insurance Premia." *N. Central J. of Ag. Econ.* 12(1990):71-78.

Nelson, C. H., and P. V. Preckel. "The Conditional Beta Distribution As a Stochastic Production Function." *Amer. J. of Ag. Econ.* 71(1989):370-378.

Ott, L. *An Introduction to Statistical Methods and Data Analysis*. PWS-Kent Publishing Co. Boston. 1988.

Pearson, E. S. *Pearson Karl: Tracts for Computers,N o VIII, Tables of Logarithms of the Complete Gamma Function,for Arguments 2 to 1200, i.e. beyond Legendre's Range*. Cambridge University Press. 1922.

Pindyck, R. S. and D. L. Rubinfeld. *Econometric Models and Economic Forecasts*. McGraw-Hill, Inc. New York. 1991.

Reutlinger, S. "Techniques for Project Appraisal Under Conditions of Uncertainty." World Bank Staff Occasional Paper, No. 10. Baltimore: The Johns Hopkins University Press. 1970.

Richardson, J. W. and G. D. Condra. "Farm Size Evaluation in the El Paso Valley: A Survival/Success Approach." *Amer. J. of Ag. Econ.* 63(1981):430-37.

Richardson, J.W. and G. D. Condra. "A General Procedure for Correlating Events in Simulation Models." Texas Agricultural Experiment Station. Department of Agricultural Economics (mimeo). 1978.

Richardson, J. W., C. J. Nixon. *Description of FLIPSIM V: A General Firm Level Policy Simulation Model*. Agricultural and Food Policy Center, Department of Agricultural Economics, Texas A&M University, July 1986.

Richardson, J. W. *Notes on Applied Simulation in Agriculture*, Unpublished, 1991.

Rodriguez, A., and R. G. Taylor. "Stochastic Modeling of Short-Term Cattle Operations." *Amer. J. of Ag. Econ.* 70(1988):123-32.

Roeder, K. "Density Estimation with Confidence Sets Exemplified by Superclusters and Voids in the Galaxies." *J. of the Amer. Statistical Association*. 85(1990):617-27.

Rosenblatt, H. M. "Remarks on Some Nonparametric Estimates of a Density Function." *Ann. Math. Statist.* (1956):832-837.

Scott, D. W. "On Optimal and Data-Based Histograms." *Biometrika*. 66(1979):605-610.

Sedgewick, R. *Algoriths in C*. Addison-Wesley Publishing Company, Inc. 1992

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York. 1986.

Sinha, S. K. *Reliability and Live Testing*. Wiley Eastern Limited, New Delhi, 1986.

Staniswalis, J. G. "Local Bandwidth Selection for Kernal Estimates." *J. of the Amer. Statistical Association*. 84(1989):284-88.

Stephens, M. A., "EDF Statistics for Goodness of Fit and Some Comparisons." *Journal of the American Statistical Association*, 69(1974):730-37

Tapia, R. A., and J. R. Thompson. *Nonparametric Probability Density Estimation*. Johns Hopkins University Press. Baltimore, 1987.

Taylor, R. "A Simple Method for Estimating Empirical Probability Density Functions." Staff Paper 81-1, Montana State University, Bozeman, MT. 1981.

Taylor, R. "A Computer Program for Maximum Likelihood Estimation of a Hyperbolic Tangent/Cubic Approximation of a Probability Distribution Function." Staff Paper 83-11, Montana State University, Bozeman, MT. 1983.

Taylor, R. "A Flexible Method for Empirically Estimating Probability Functions." *West. J. of Ag. Econ.* 9(1984):66-76.

Taylor, R. "A Practical Procedure for Fitting Multivariate Nonnormal Probability Density Functions." Unpublished Manuscript. 1987.

Terrell, G. R. "The Maximum Smoothing Principle in Density Estimation." *J. of the Amer. Statistical Association*. 85(1990):470-76.

Van Tassel, L. W., J. W. Richardson, and J. R. Conner. "Empirical Distribution and Production Analysis: A Documentation Using Meterological Data." The University of Tennessee, Agricultural Experiment Station, Bulletin 671. September 1989.

Vinod H.D., and A. Ullah. "Flexible Production Function Estimation by Nonparametric Kernel Estimators." Edited by Rhodes G.F. Jr. and Fomby T.B. *Advances in Econometrics*. Jai Press, Greenwich, CT. 1988.

Wertz, W. *Statistical Density Estimation: A Survey*. Vandenhoeck & Ruprecht, Gottingen. 1978:30-50.

Wilk, M. B., R. Gnanadesikan and M. J. Huyett. "Estimation of Parameters of the Gamma Distribution Using Order Statistics." *Biometrika*, 49(1962):525-35.

FIGURES

Figure 1. Use of a Cumulative Density Function to Obtain the Probability Associated with a Range.

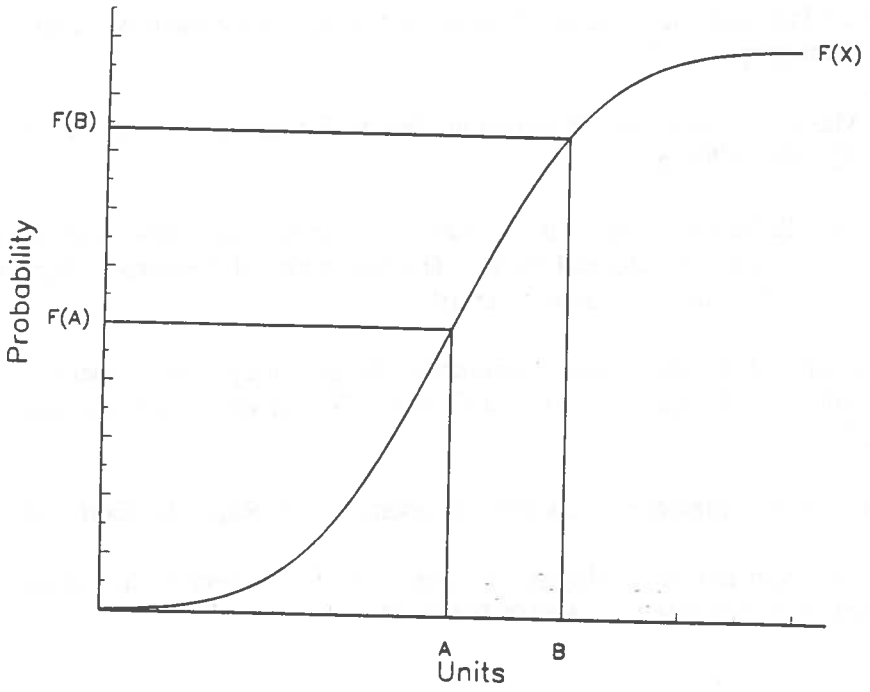


Figure 2. Use of a Probability Density Function to Obtain the Probability Associated with a Range.

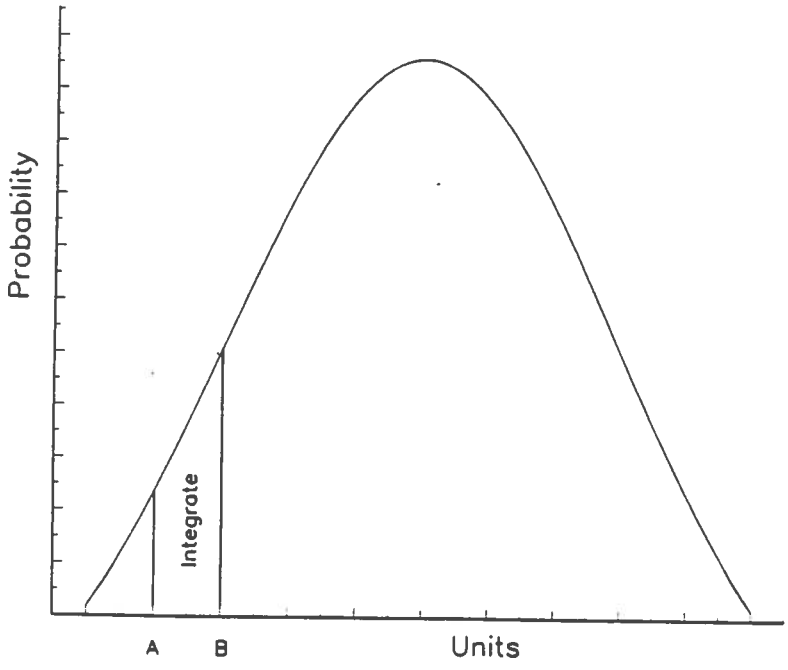


Figure 3. Dispersion Around a Mean versus a Trend.

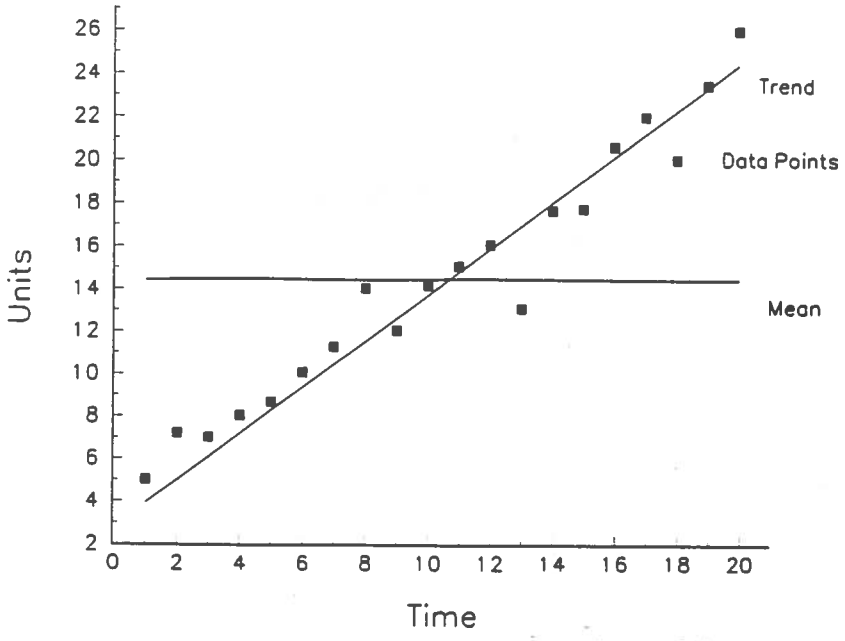


Figure 4. Definition of a Q–Q Plot.

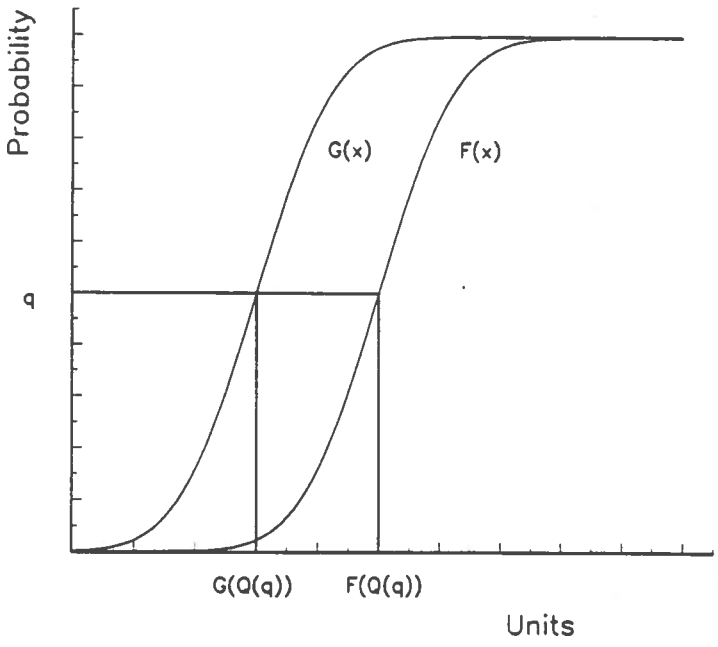


Figure 5. Example of a Q – Q Plot.

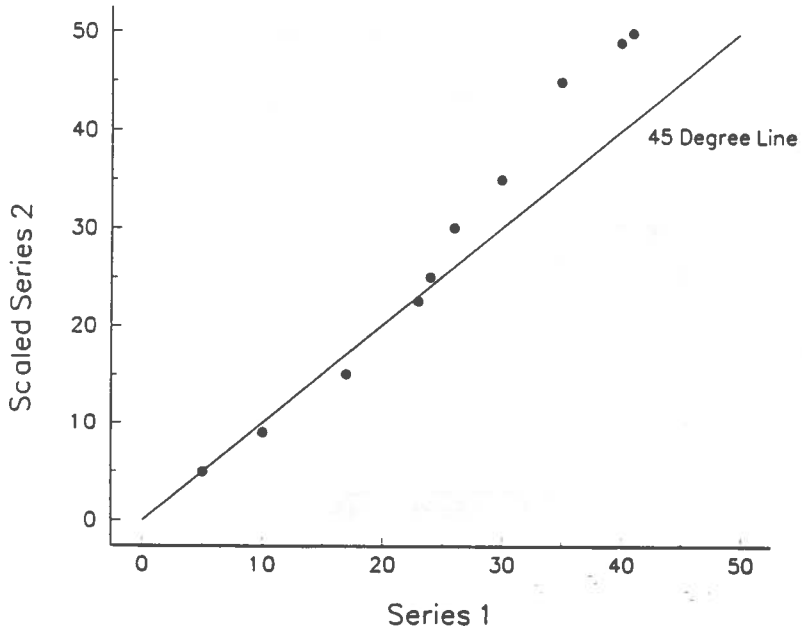


Figure 6. Definition of a P – P Plot.

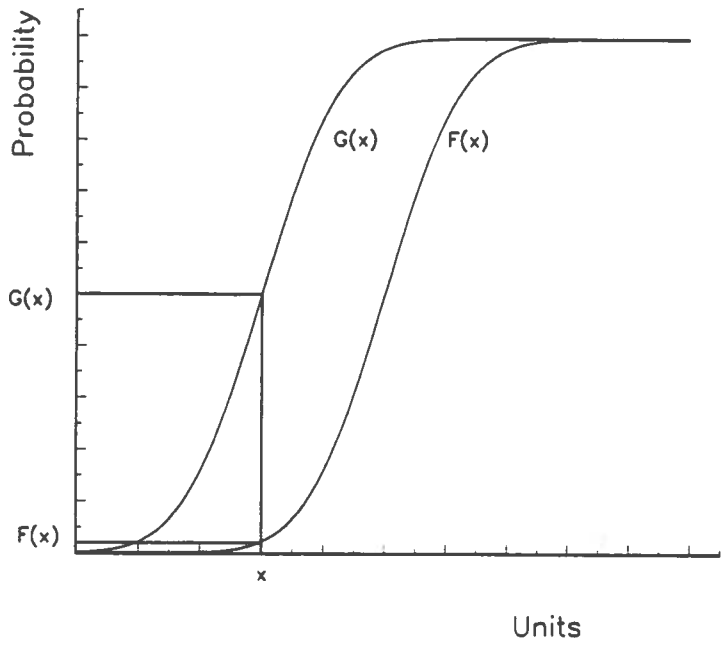


Figure 7. Example of a P – P Plot.

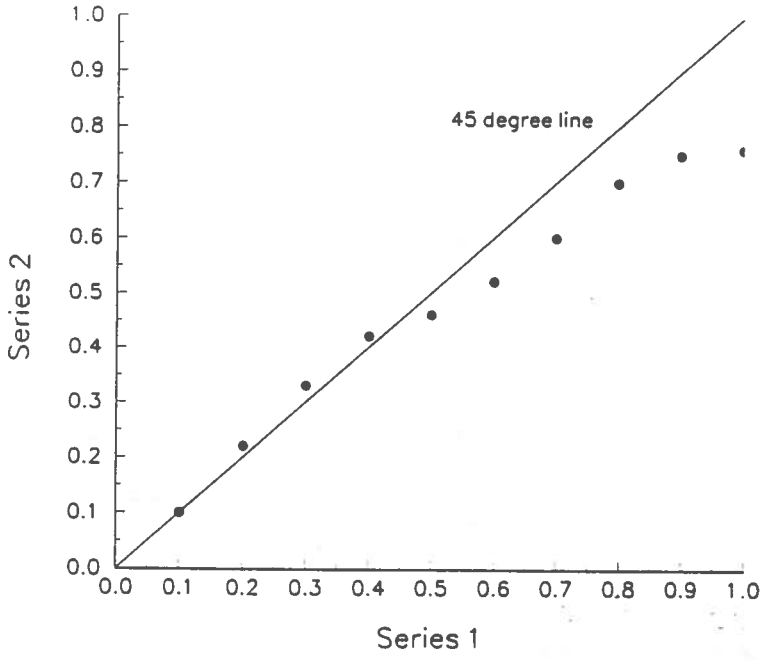


Figure 8. Simulated Aquifer Recharge Levels.

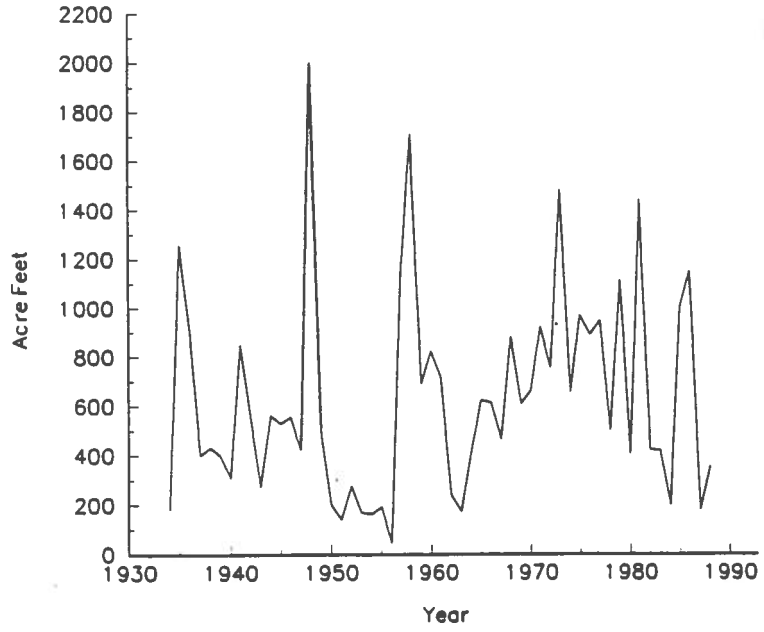


Figure 9. Stem Leaf and Boxplot of Simulated Aquifer Data.

```
Stem Leaf                    #        Boxplot
  20 0                       1           0
  19
  18
  17 1                       1           |
  16                                     |
  15                                     |
  14 59                      2           |
  13                                     |
  12 6                       1           |
  11 245                     3           |
  10 0                       1           |
   9 1357                    4           |
   8 2589                    4           +------+
   7 26                      2           |      |
   6 122669                  6           |  +   |
   5 013666                  6           *------*
   4 001122237               9           |      |
   3 16                      2           +------+
   2 00478                   5           |
   1 4677889                 7           |
   0 4                       1           |
     ----+----+----+----+
Multiply Stem.Leaf by 10**+2
```

Figure 10. Standard Normal Probability Density Function.

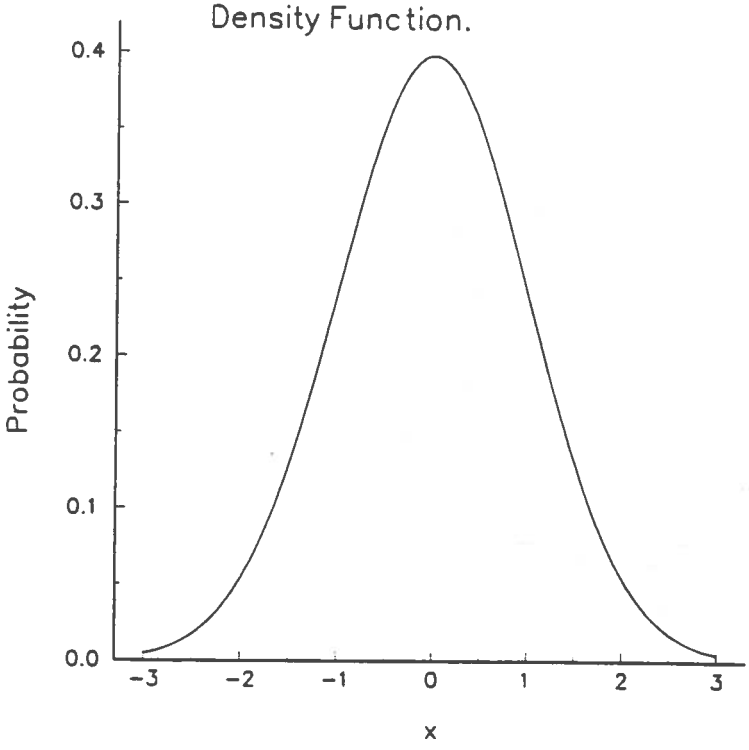


Figure 11. Estimated Normal PDF for the Simulated Aquifer Data.

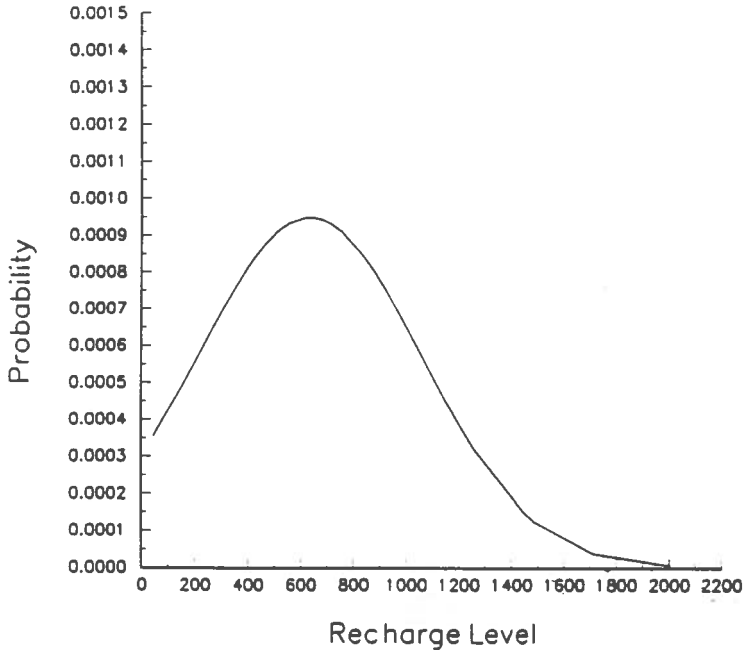


Figure 12. Different Shapes for the Gamma PDF Depending on the Values for Alpha and Beta.

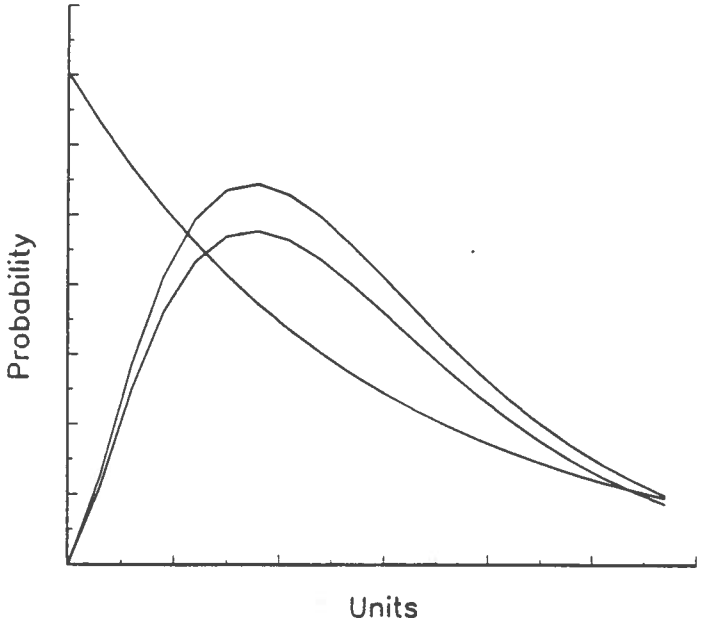


Figure 13. Three Estimated Gamma PDFs for the Simulated Aquifer Data.

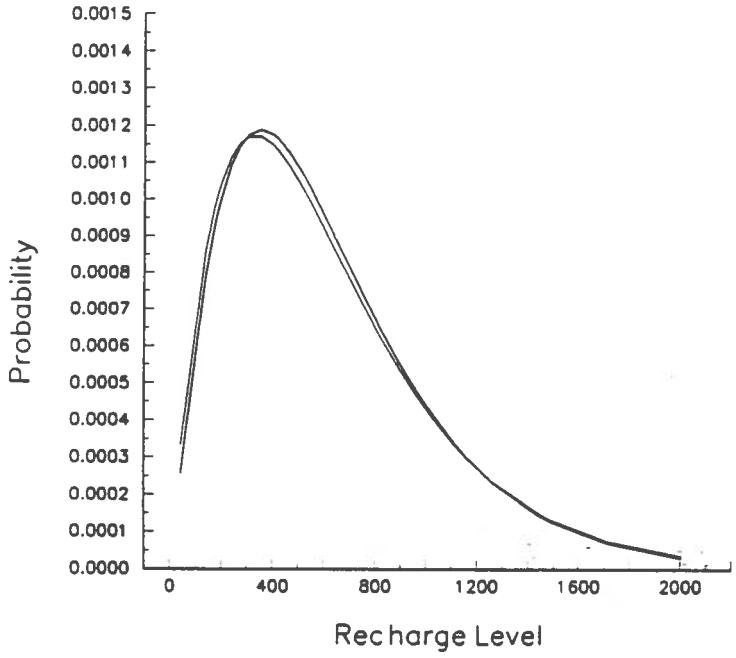


Figure 14. Example of an Exponential PDF.

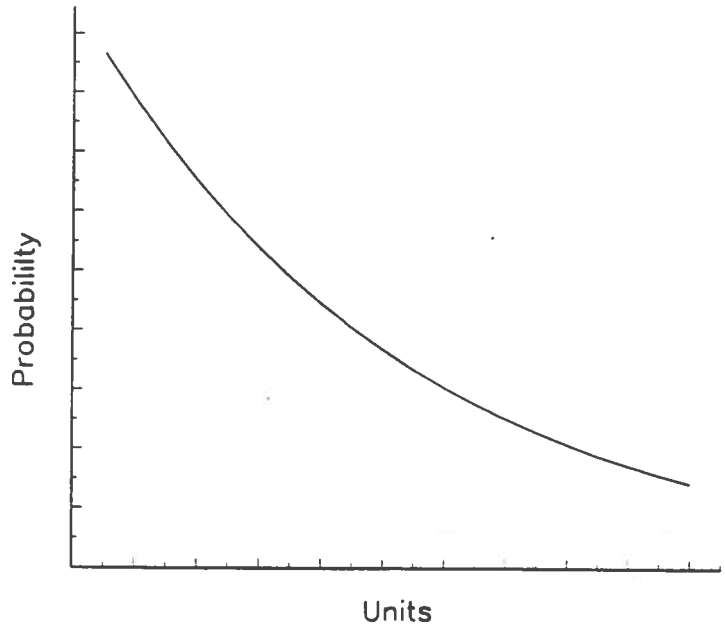


Figure 15. Estimated Exponential PDF for the Simulated Aquifer Data.

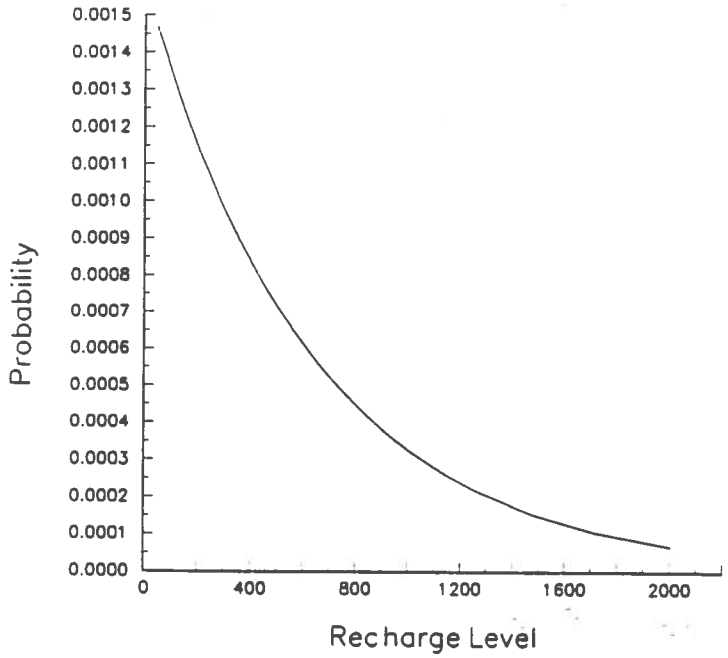


Figure 16. Estimated Exponential CDF for the Simulated Aquifer Data.

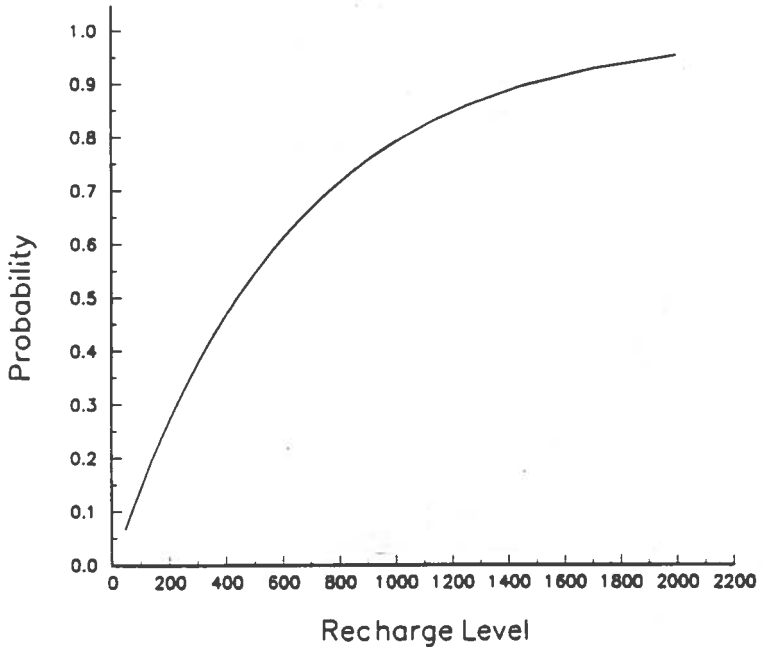


Figure 17. Examples of Different Beta Distributions Shapes Depending on the Parameter Values.

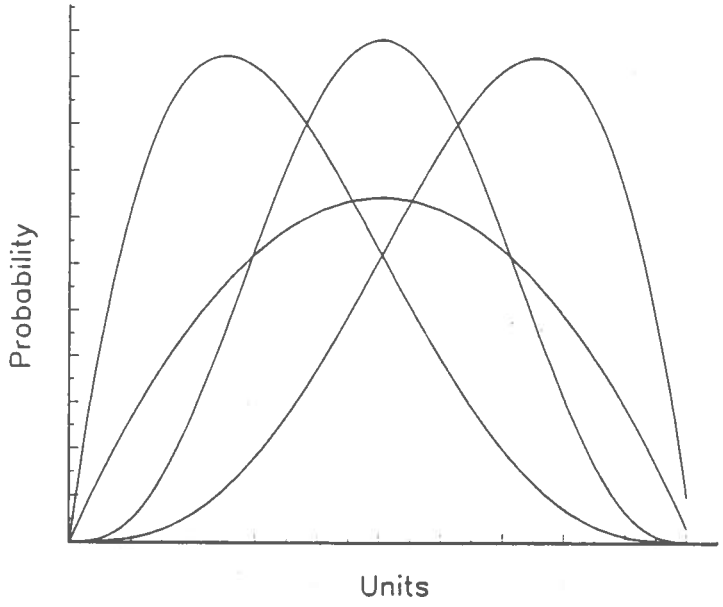


Figure 18. Estimated Beta Distributions for the Simulated Aquifer Data.

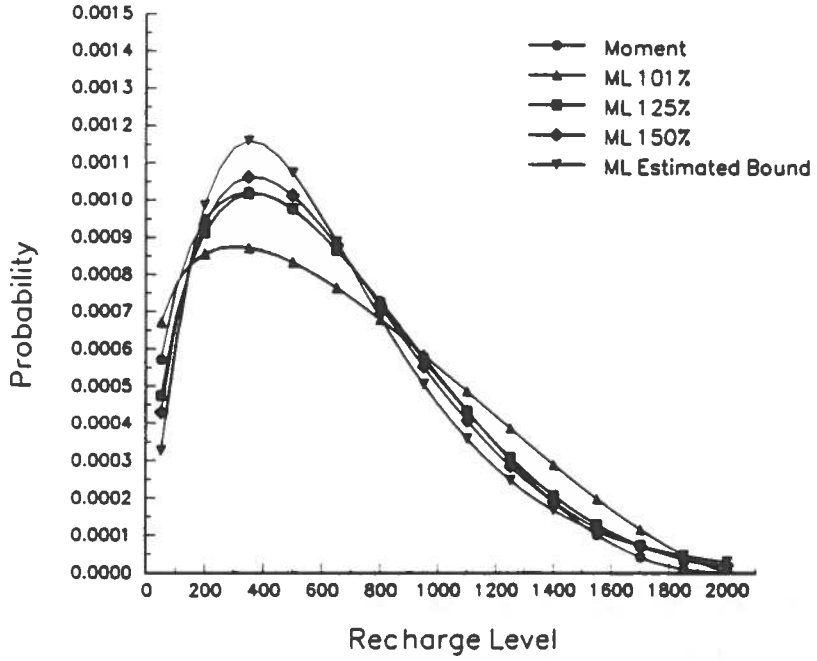


Figure 19. Estimated Histogram Using Sturge's Rule for the Simulated Aquifer Data.

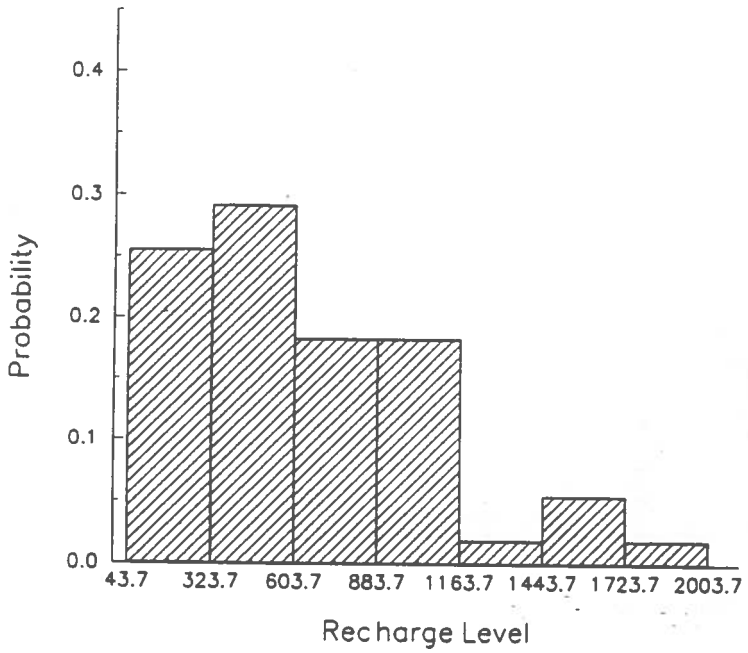


Figure 20. Estimated Histogram with Seven Intervals Starting at Zero for the Simulated Aquifer Data.

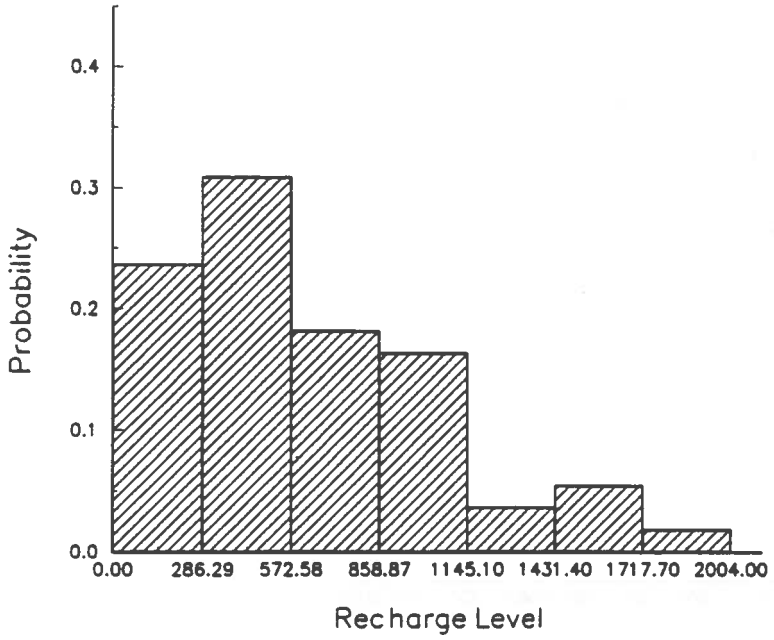


Figure 21. Estimated Histogram Using Scott's Rule for the Simulated Aquifer Data.

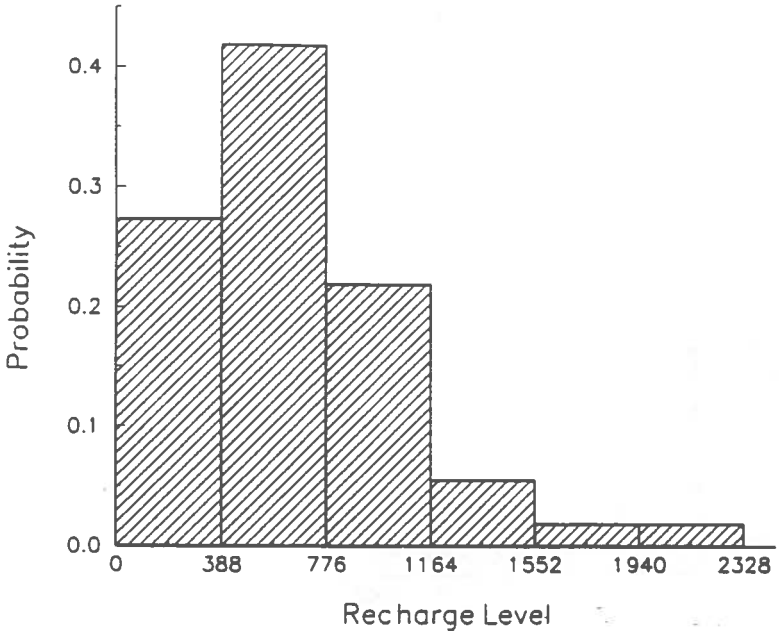


Figure 22. Estimated Empirical CDFs for Simulated Aquifer Data.

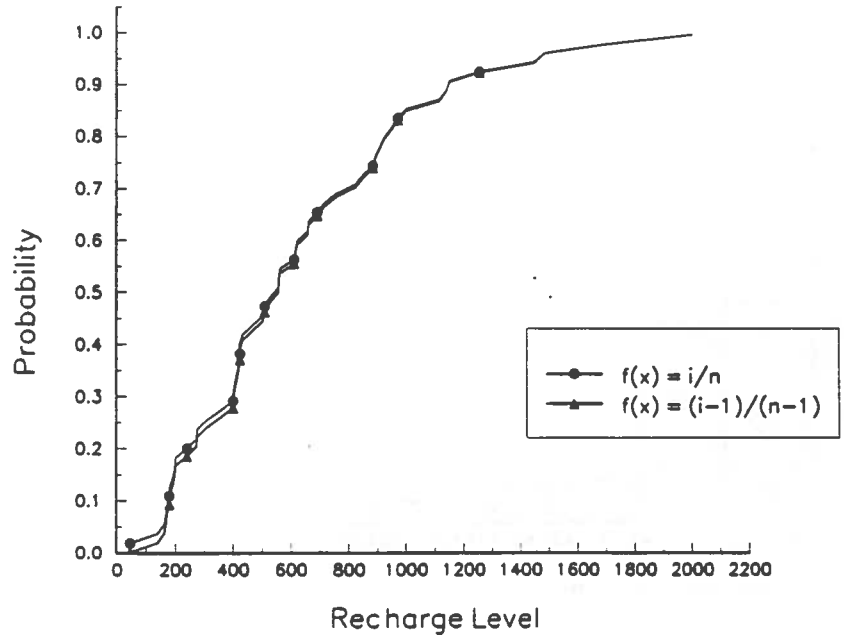


Figure 23. Effect of Bandwidth (b) on Kernel PDFs.

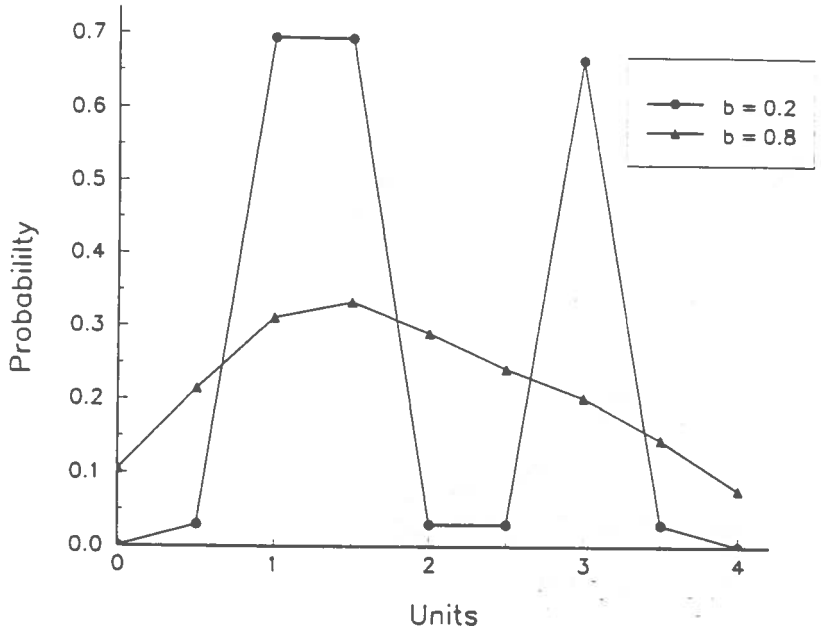


Figure 24. Example of the Development of a Kernel PDF.

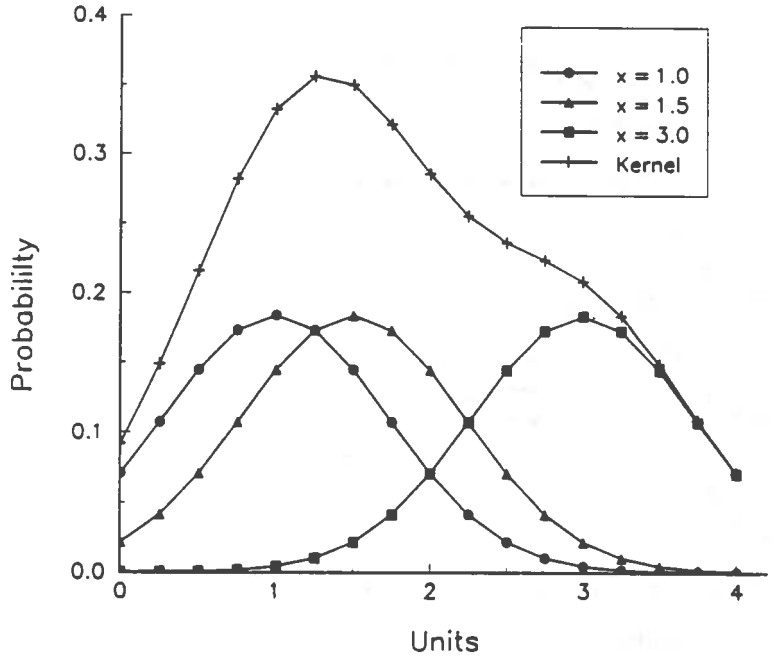


Figure 25. Estimated Kernel PDFs Using the Simulated Aquifer Data.

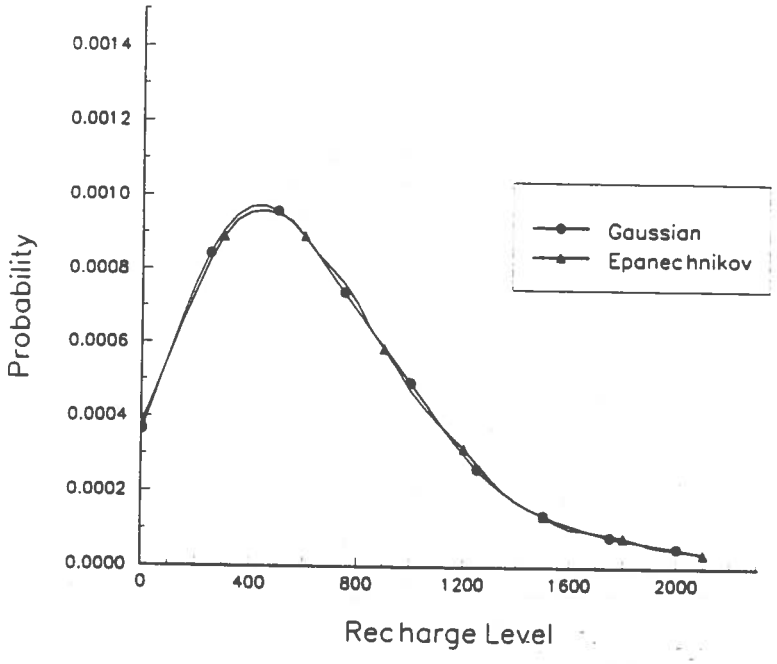


Figure 26. Taylor's Hyperbolic Tangent Transformation.

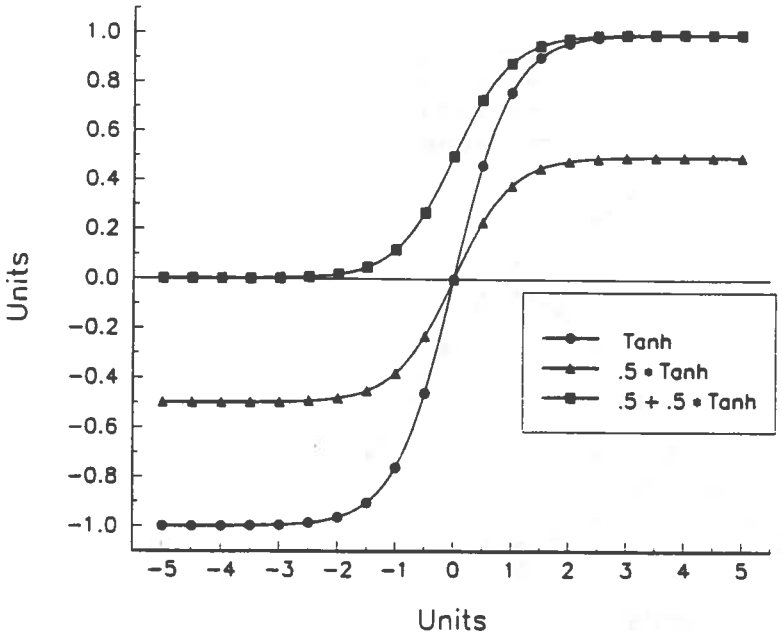


Figure 27. Estimated CDF Using Taylor's Hyperbolic Tangent Transformation for the Simulated Aquifer Data.

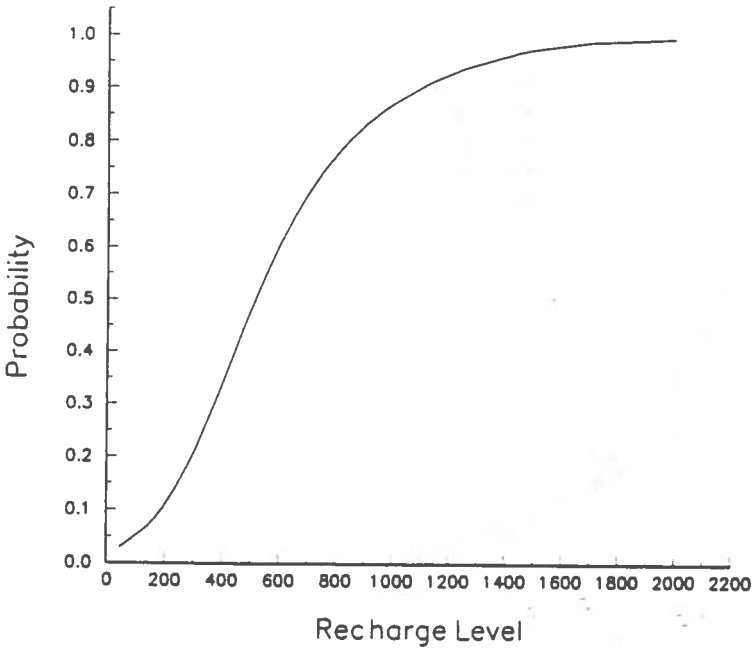


Figure 28. Estimated PDF Using Taylor's Hyperbolic Tangent Transformation for the Simulated Aquifer Data.

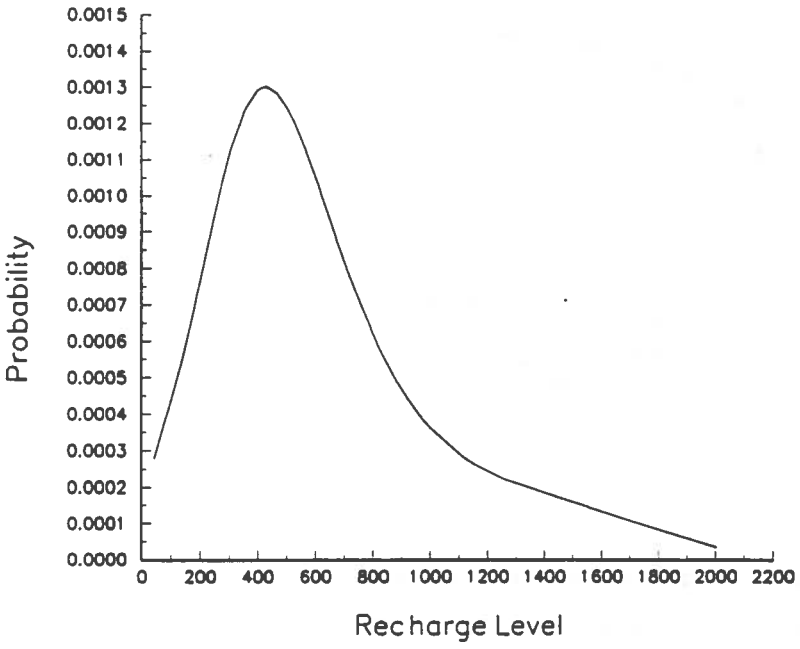


Figure 29. Plots of the Various Estimated PDFs for the Simulated Aquifer Data.

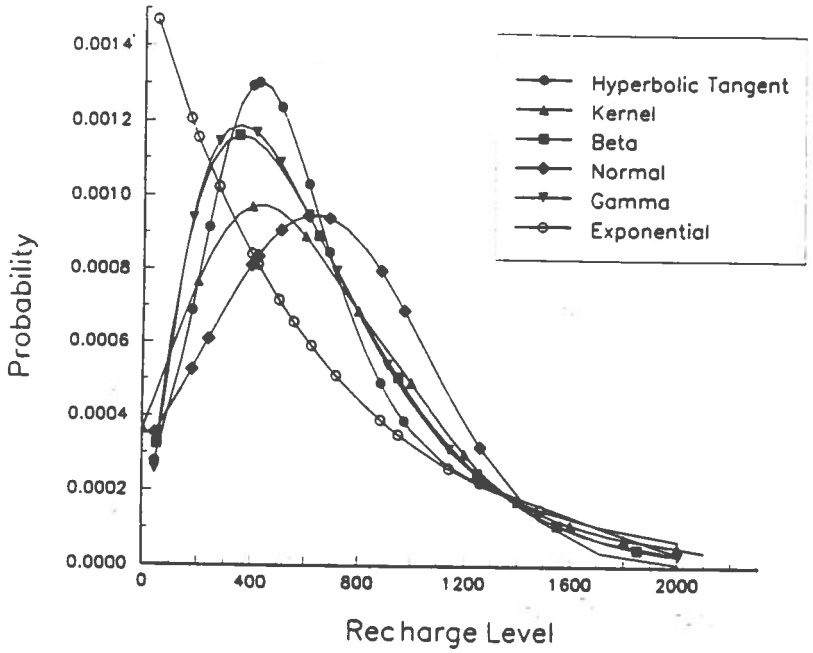


Figure 30. Plots of the Various Estimated CDFs for the Simulated Aquifer Data.

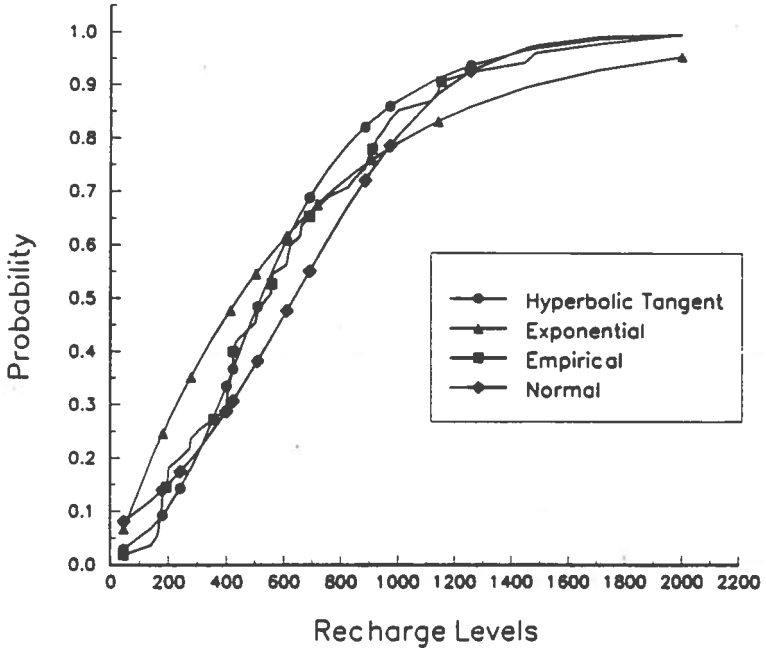


Figure 31. Monte Carlo Simulated CDFs for Aquifer Recharge Using the Normal Distribution.

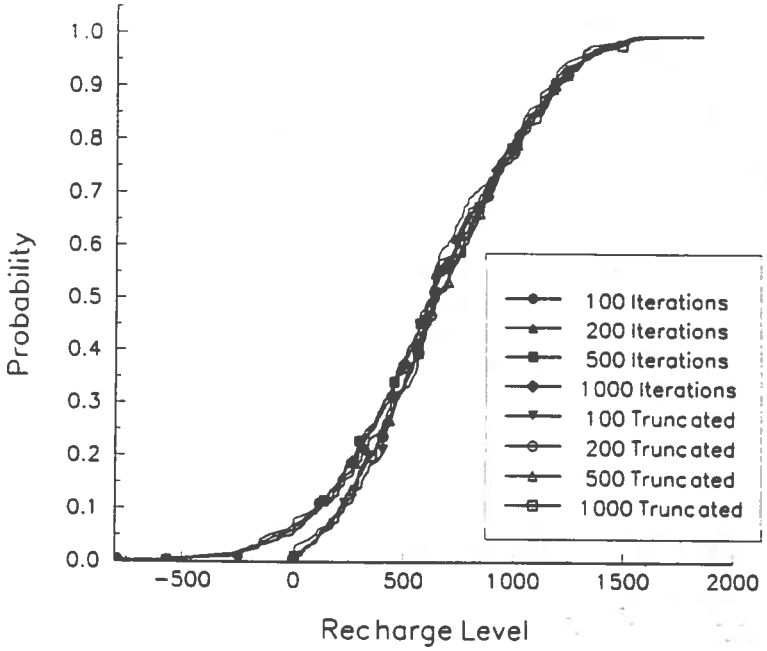


Figure 31a. Monte Carlo Simulated CDFs for Aquifer Recharge Using the Normal Distribution.

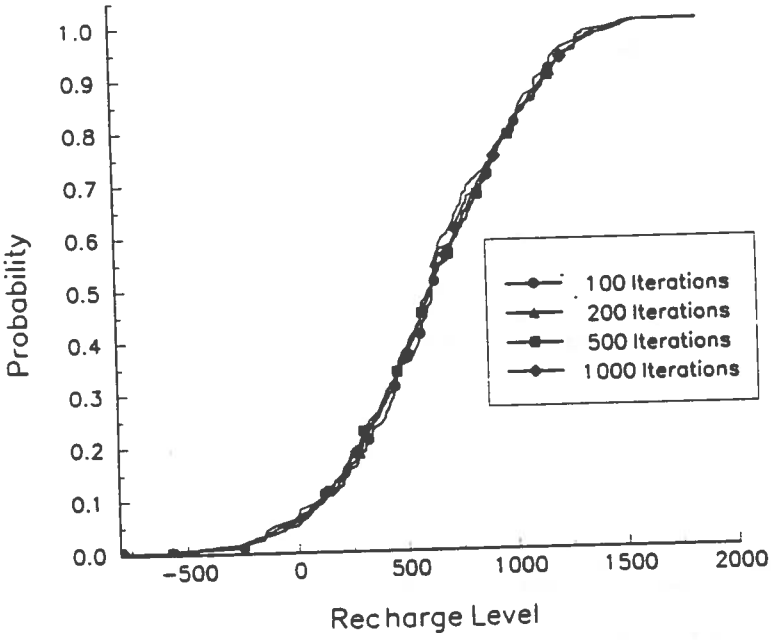


Figure 31 b. Monte Carlo Simulated CDFs for Aquifer Recharge Using the Truncated Normal Distribution.

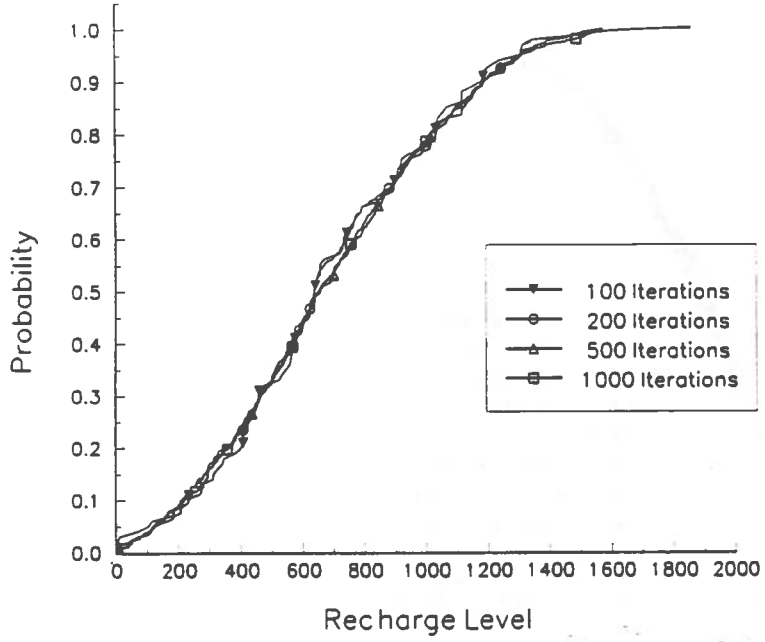


Figure 32. Monte Carlo Simulated CDFs for Aquifer Recharge Using the Gamma Distribution for 500 Iterations.

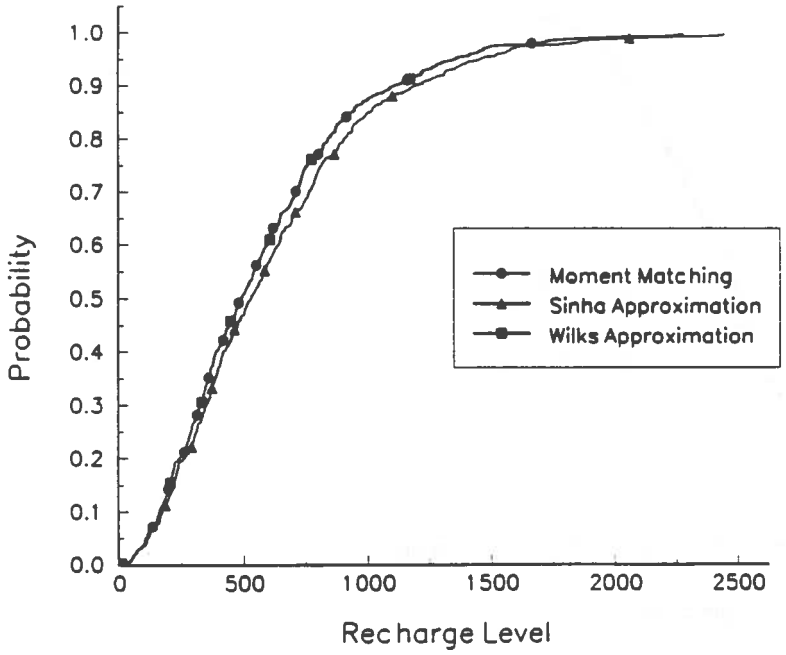


Figure 33. Monte Carlo Simulated CDFs for Aquifer Recharge Using the Beta Distribution.

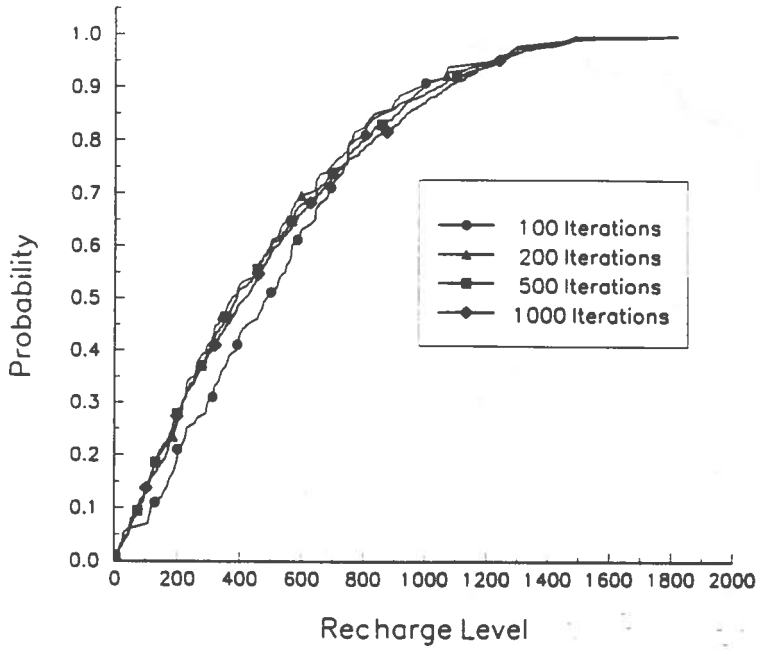


Figure 34. Monte Carlo Simulated CDFs for Aquifer Recharge Using the Exponential Distribution.

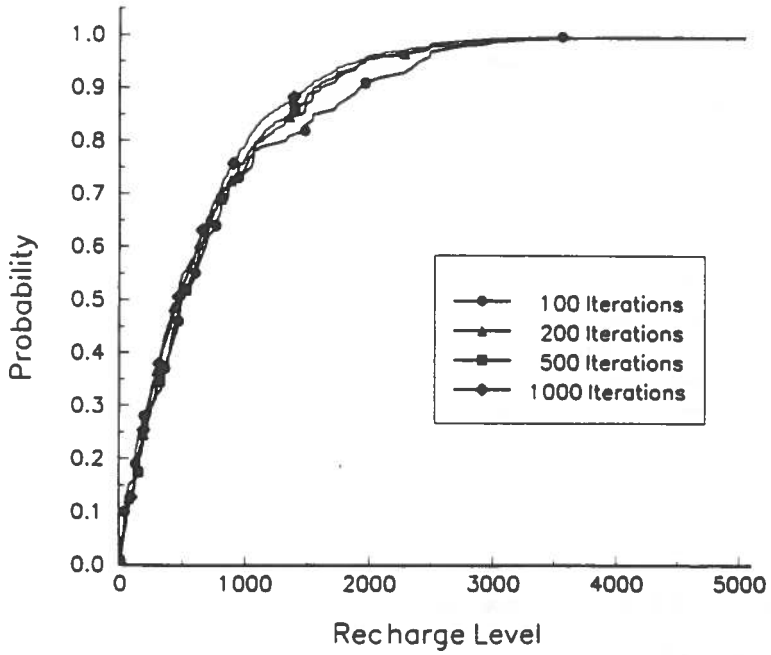


Figure 35. Monte Carlo Simulated CDFs for Aquifer Recharge Using the Cumulative Distribution.

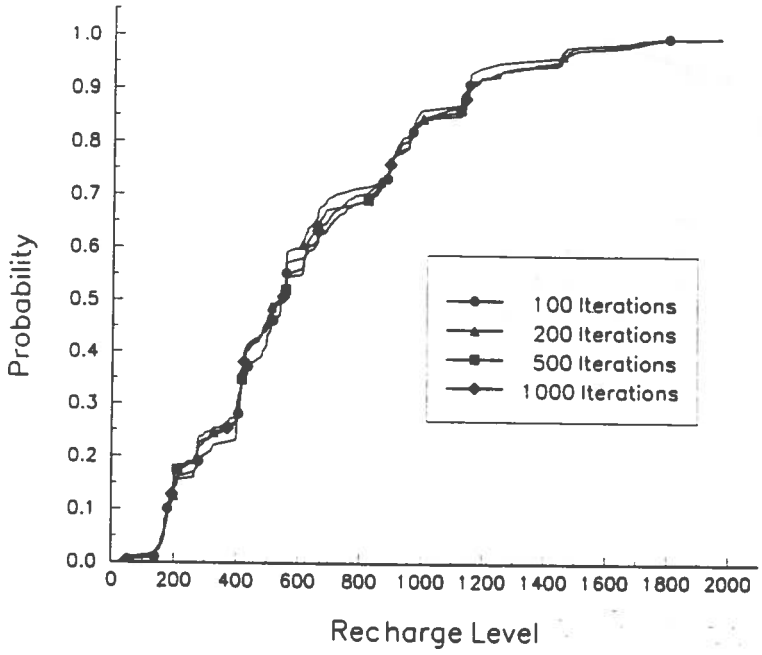


Figure 36. Monte Carlo Simulated CDFs for Aquifer Recharge Using the J–Distribution.

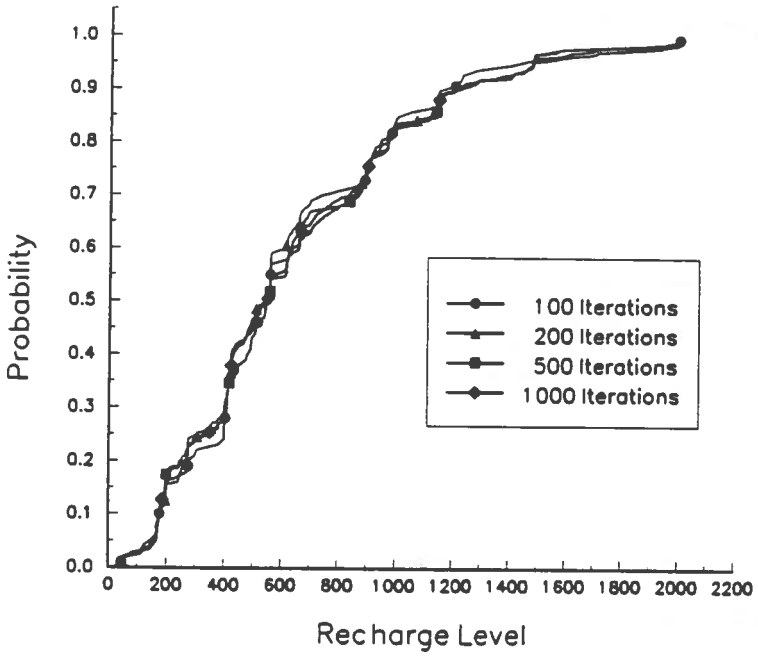


Figure 37. Monte Carlo Simulated CDFs for Aquifer Recharge Using the Hyperbolic Tangent Distribution.

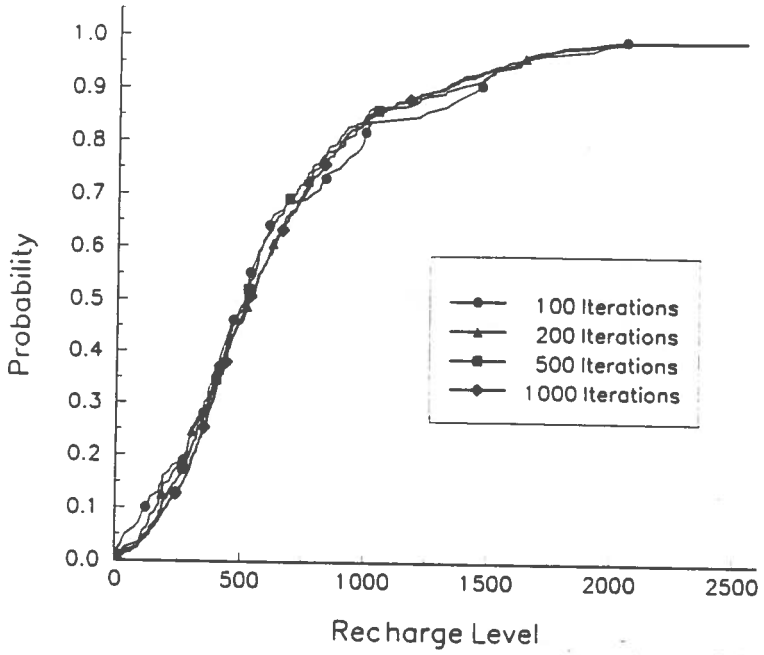


Figure 38. Monte Carlo Simulated CDFs for Various Distributions Using 500 Iterations.

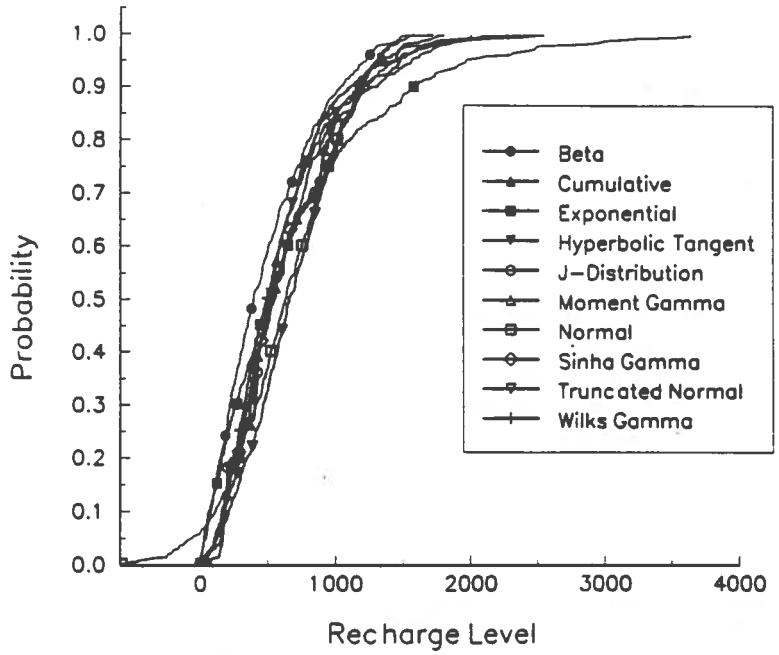


Figure 38a. Monte Carlo Simulated CDFs for Various Distributions Using 500 Iterations.

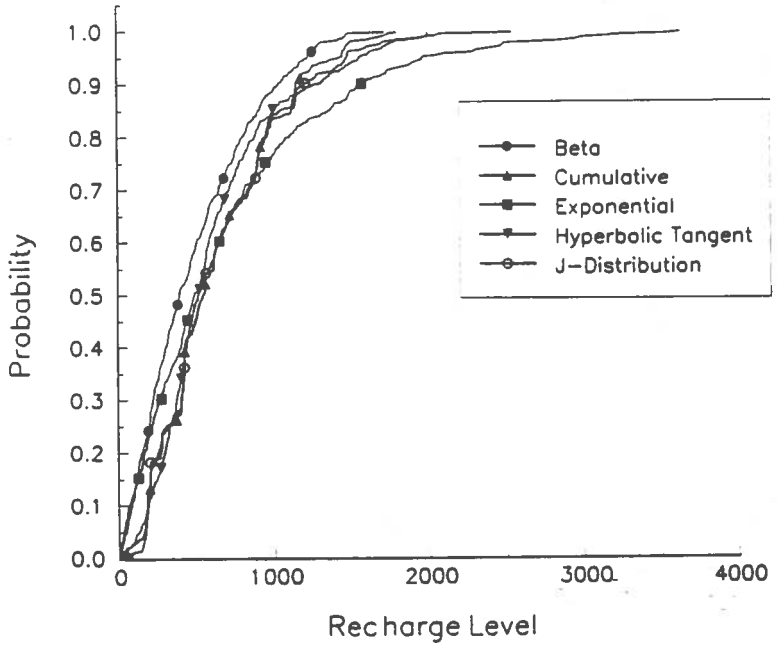
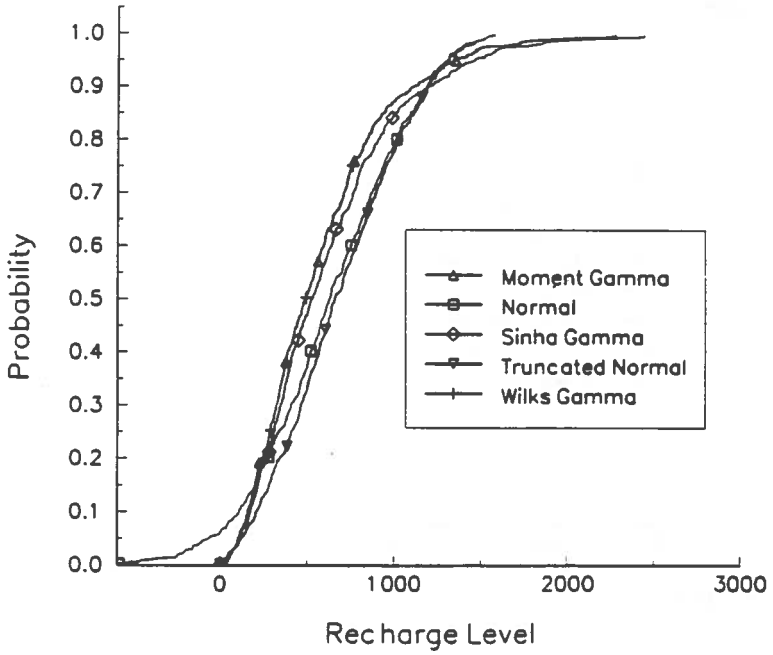


Figure 38b. Monte Carlo Simulated CDFs for Various Distributions Using 500 Iterations.



APPENDIX A - SIMULATED AQUIFER DATA

Year	Recharge Level	Year	Recharge Level
1934	179.6	1962	239.4
1935	1258.2	1963	170.7
1936	909.6	1964	413.2
1937	400.7	1965	623.5
1938	432.7	1966	615.2
1939	399.0	1967	466.5
1940	308.8	1968	884.7
1941	850.7	1969	610.5
1942	557.8	1970	661.6
1943	273.1	1971	925.3
1944	560.9	1972	756.4
1945	527.8	1973	1486.5
1946	556.1	1974	658.5
1947	442.6	1975	973.0
1948	2003.5	1976	894.1
1949	508.1	1977	952.0
1950	200.2	1978	502.5
1951	139.9	1979	1117.8
1952	275.5	1980	406.4
1953	167.6	1981	1448.3
1954	162.1	1982	422.4
1955	192.0	1983	420.1
1956	43.7	1984	197.7
1957	1142.6	1985	1003.3
1958	1711.2	1986	1153.8
1959	690.4	1987	178.3
1960	824.8	1988	355.5
1961	717.1		

```
C--   TEST-DIST A PROGRAM TO DEMONSTRATE THE NORMAL,  BETA,  ETC
C--   DISTRIBUTIONS.
C--   DIMENSIONS ARE SET FOR 1000 ITERATIONS
C--                    DEVELOPED BY JWR 7/18/93
C--
C---------------------------------------------------------------------
      DIMENSION RESULT(1000,2)
C--   COMMON FOR THE GAUSE SUBROUTINE
      COMMON /JWR3/ RNAR(128),RKK
      REAL*8 RKK
C---------------------------------------------------------------------
C--
C--   TEST.INP   INPUT FILE
      I042 = 12
      OPEN (UNIT=I042,FILE='TEST.INP',STATUS='UNKNOWN',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')

C--   RESULT.OUT OUTPUT FILE FOR
      I045 = 15
      OPEN (UNIT=I045,FILE='RESULT.OUT',STATUS='SCRATCH')
      CLOSE (I045)
      OPEN (UNIT=I045,FILE='RESULT.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')

1     FORMAT (3I4)
2     FORMAT (2I2,6X,10F7.0)
3     FORMAT(20I4)
4     FORMAT(2A4,I2,10F10.3)
5     FORMAT(T4,2A4,I4,10F10.3)
9     FORMAT('EXPONENTIAL PDF ',T16,2I4,40F10.4)
10    FORMAT('NORMAL PDF ',T16,2I4,40F10.4)
11    FORMAT('GAMMA PDF ',T16,2I4,40F10.4)
12    FORMAT(T2,2I4,40F7.3)
13    FORMAT(T7,'(c) Copyright 1993 Texas Agricultural Experiment Statio
     1n',//)
14    FORMAT(/,T7,'TEST-DIST  VERSION 1.1   RELEASE DATE JULY 18, 1993')
15    FORMAT('BETA PDF ',T16,2I4,40F10.4)
16    FORMAT('NORMAL PDF ',T16,2I4,40F10.4)
17    FORMAT('NAYLOR"S NORMAL PDF ',T16,2I4,40F10.4)

C--   WRITE THE NAME AND COPYRIGHT TO THE SCREEN
      WRITE(*,14)
      WRITE(*,13)

C---------------------------------
      READ(I042,3) NOITER,IVAR,NOBS
      WRITE(I045,1) IVAR , NOITER,NOBS
      WRITE(*,1) IVAR, NOITER, NOBS

      DO 1000 ISET = 1,99

      READ(I042,4,END=1100) ANAME1,ANAME2,IDIST,VAR1,VAR2
      IF(VAR1.EQ.0.0.AND.VAR2.EQ.0.0) GO TO 1100
      WRITE(*,5) ANAME1,ANAME2,IDIST,VAR1,VAR2
      IF(IDIST.EQ.7) THEN
      READ(I042,4,END=1100) ANAME3,ANAME4,IDIS2,VAR3,VAR4
      WRITE(*,5) ANAME3,ANAME4,IDIS2,VAR3,VAR4
      END IF

C--   SEED FOR RANDOM GENERATOR
      NAR = 31415
      I = 1
C---------------------------------------------------------------------------
C--   GENERATE RANDOM NUMBER NORMAL DISTRIBUTION
      IF(IDIST.EQ.1) THEN
      AMEAN = VAR1
      STDDEV= VAR2
      DO 400 L=1,NOITER
      CALL GAUSE(GAUSES,NAR)
      RESULT(L,I) = AMEAN + (STDDEV * GAUSES)
      WRITE(I045,10) I,L,RESULT(L,I)
400   CONTINUE                    ! PHYSICAL END OF THE NORMAL DISTRIBUTION
      GO TO 1000
      END IF

C---------------------------------------------------------------------------
C--   GENERATE TRUNCATED NORMAL DISTRIBUTION
C--   DISCARD RANDOM VALUES THAT ARE LESS THAN ATEST VALUE
      IF(IDIST.EQ.5) THEN
      AMEAN = VAR1
      STDDEV= VAR2
      ATEST = 0.0         ! MINIMUM CUTOFF FOR A RANDOM NORMAL VALUE
      DO 450 L=1,NOITER
410   CONTINUE
      CALL GAUSE(GAUSES,NAR)
      RESULT(L,I) = AMEAN + (STDDEV * GAUSES)
      IF(RESULT(L,I).LT.ATEST) GO TO 410
      WRITE(I045,16) I,L,RESULT(L,I)
```

```
      GO TO 1000
      END IF

C-----------------------------------------------------------------------
C--   GENERATE RANDOM NUMBER NORMAL DISTRIBUTION SUGGESTED BY NAYLOR
C--   IT HAS A NEGATIVE SKEW IF THE SAMPLE SIZE IS LESS THAN 500,
C--   THE MEAN IS ALMOST 0.0 AND THE VARIANCE IS ABOUT 1.0 FOR LARGER
C--   SAMPLES.  I WOULD PREFER GAUSE
      IF(IDIST.EQ.4) THEN
      AMEAN = VAR1
      STDDEV= VAR2
      DO 395 L=1,NOITER
      SUM = 0.0
      DO 390 I=1,12
      ANO1 = RANG(NAR)
390   SUM = SUM + ANO1
      RESULT(L,1) = (SUM-6.0)
C      RESULT(L,I) = AMEAN + (STDDEV * GAUSES)
      WRITE(IO45,17) I,L,RESULT(L,1)
395   CONTINUE                    ! PHYSICAL END OF THE NORMAL DISTRIBUTION
      GO TO 1000
      END IF

C-----------------------------------------------------------------------
C--   GAMMA DISTRIBUTION SUGGESTED BY NAYLOR
      IF(IDIST.EQ.2) THEN
C--   CONVERT THE ALPHA & BETA FROM MELDJE'S NOTATION AND FORMAT TO
C--   NAYLOR'S FORMULA
C--   MJELDE ASSUMES  BETA-HAT = VARX/MEANX
C--              AND  ALPHA-HAT= (MEANX)**2/VARX.
      AK = VAR1
      IK = AINT(AK)
      ABETA  = 1.0/VAR2
      DO 500  L=1,NOITER
      SUM = 1.0
      DO 490 I=1,IK
      ANO1 = RANG(NAR)
      SUM = SUM * ANO1
490   CONTINUE
      RESULT(L,1) = -LOG(SUM)/ABETA
      WRITE(IO45,11) I,L,RESULT(L,1)
500   CONTINUE
      GO TO 1000
      END IF

C-----------------------------------------------------------------------
C--   EXPONENTIAL DISTRIBUTION
      IF(IDIST.EQ.3) THEN
      AMEAN = VAR1
      DO 600 L=1,NOITER
      ANO = RANG(NAR)
      RESULT(L,I) = (-AMEAN) * LOG(ANO)
      WRITE(IO45,9) I,L,RESULT(L,I)
600   CONTINUE
      GO TO 1000
      END IF

C-----------------------------------------------------------------------
C--   BETA DISTRIBUTION SUGGESTED BY LAW & KELTON P. 492
      IF(IDIST.EQ.7) THEN
      IK1 = AINT(VAR1)
      IK2 = AINT(VAR2)
      ABETA1  = 1.0
      ABETA2  = 1.0
      DO 700  L=1,NOITER
      SUM1 = 1.0
      SUM2 = 1.0
      DO 680 I=1,IK1              !GENERATE Y1 ~ GAMMA(VAR1,1)
      ANO1 = RANG(NAR)
      SUM1 = SUM1 * ANO1
680   CONTINUE
      DO 690 I=1,IK2              !GENERATE Y2 ~ GAMMA(VAR2,1)
      ANO2 = RANG(NAR)
      SUM2 = SUM2 * ANO2
690   CONTINUE
      ALPHA1 = -LOG(SUM1)/ABETA1
      ALPHA2 = -LOG(SUM2)/ABETA2          ! GENERATE X ~ BETA(Y1,Y2)
      BETA = ALPHA1/(ALPHA1+ALPHA2)       ! X = Y1/(Y1+Y2)
      RESULT(L,1) = VAR3 + (VAR4 - VAR3) * BETA
      WRITE(IO45,15) I,L,RESULT(L,1)
700   CONTINUE
      GO TO 1000
      END IF

1000  CONTINUE
1100  CONTINUE
      STOP
      END
```

```
C-----------------------------------------------------------------------
      SUBROUTINE GAUSE(GAUSF,NAR)
      COMMON /JWR3/ RNAR(128),RKK
      REAL*8 RKK
C--   MAY 19, 1990   LAW & KELTON PAGE 491 SECOND EDITION
C--   GENERATES RANDOM NORMAL(0,1) DEVIATE BY POLAR METHOD
 10   VA    = 2.0*RANG(NAR) - 1.0
      VB    = 2.0*RANG(NAR) - 1.0
      SUMSQ = VA**2 + VB**2
      IF ( (SUMSQ.GT.1.0) .OR. (SUMSQ.EQ.0.0) ) GOTO 10
      GAUSF = VA*(SQRT(((-2.0)*(ALOG(SUMSQ)))/SUMSQ))
      RETURN
      END
C-----------------------------------------------------------------------
      FUNCTION RANG(NAR)
      COMMON /JWR3/ RNAR(128),RKK
      REAL*8 RKK
      REAL*8 MDX,MDX128,MULT,INC
C--   LINEAR CONGRUENTIAL GENERATOR WITH SHUFFLING.  USES ALGORITHM B
C--   OF KNUTH, VOLUME 2, P. 32 FOR SHUFFLING TO INCREASE PERIOD
C--   FROM 32768 TO AN UNKNOWN, LARGER PERIOD.  KNUTH LIKES IT.
C--   GENERATES PSEUDO-RANDOM NUMBERS, UNIFORMLY DISTRIBUTED ON (0,1)
C--   ACTUAL RANGE OF VALUES IS (1/65536, 65535/65536)
C--   REAL*8 ARITHMETIC IS USED FOR PORTABILITY.
      MDX = 65536.0D0
      MDX128 = 512.0D0
      MULT = 25173.0D0
      INC = 13849.0D0
      IF (NAR.EQ.0) GO TO 60
C--   INITIALIZE SHUFFLER BY FILLING 128 CUBBYHOLES WITH RANDOM SEEDS-
      RKK     = DBLE(FLOAT(NAR))
      RKK     = DMOD(RKK,MDX)
      IF (RKK.LE.0.0) RKK = 12345.0
      NAR     = 0
      DO 50 J=1,128
      RKK     = RKK*MULT + INC
      RKK     = DMOD(RKK,MDX)
50    RNAR(J) = RKK
C--   OBTAIN RANDOM CUBBYHOLE NUMBER
60    RKK     = RKK*MULT + INC
      RKK     = DMOD(RKK,MDX)
      J       = 1 + IDINT(RKK/MDX128)
C--   PULL OLD RANDOM SEED OUT OF CUBBYHOLE FOR USE
      RANG    = RNAR(J)/MDX
C--   RESTOCK THE DEPLETED CUBBYHOLE WITH NEW RANDOM SEED
      RKK     = RKK*MULT + INC
      RKK     = DMOD(RKK,MDX)
      RNAR(J) = RKK
      RETURN
      END

C-----------------------------------------------------------------------
      SAMPLE OF THE INPUT DATA FOR TEST

1000   1   1
NORMAL    1  0.00      1.0000       MEAN AND STD. DEV.   STANDARD NORMAL
NORMAL    1  635.52    422.55       MEAN AND STD. DEV.   NORMAL
TNORMAL   5  635.52    422.55       MEAN AND STD. DEV.   TRUNCATED NORMAL

EXPONENT  3  635.42                 MEAN (BETA)

MM-GAMMA  2  2.26      280.95       ALPHA AND BETA
S-GAMMA   2  2.09      303.42       ALPHA AND BETA
W-GAMMA   2  2.25      282.49       ALPHA AND BETA

BETA      7  1.500     3.36         P AND Q
SEC BETA  7  0.000     2003.5       SEC. CARD FOR BETA HAS A AND B
BETA      7  1.500     3.36         P AND Q
SEC BETA  7  0.000     2023.50      SEC. CARD FOR BETA HAS A AND B
BETA      7  1.500     3.36         P AND Q
SEC BETA  7  0.000     2504.38      SEC. CARD FOR BETA HAS A AND B
BETA      7  1.500     3.36         P AND Q
SEC BETA  7  0.000     3005.25      SEC. CARD FOR BETA HAS A AND B
BETA      7  1.500     3.36         P AND Q
SEC BETA  7  0.000     11336.53     SEC. CARD FOR BETA HAS A AND B
BETA      7  1.500     3.36         P AND Q
SEC BETA  7  0.000     11336.53     SEC. CARD FOR BETA HAS A AND B
```

```
C--   CDIST A PROGRAM TO DEMONSTRATE THE UNIVARIATE AND MULTIVATIATE
C--   CONTINUOUS EMPRIICAL C - DISTRIBUTION.
C--   DIMENSIONS ARE SET FOR 1000 ITERATIONS AND 40 VARIABLES
C--                    DEVELOPED BY JWR 7/7/93
C-------------------------------------------------------------------------
C--     CODE THE DATA BY COLUMN USING A 2I4 FORMAT FOR CARD ONE
C--     1-4 IS NO. OF VARIABLES,
C--     5-8 IS NO. OF ITERATIONS,
C--     8-12 IS NO. OF POINTS IN PDF.
C--     FOLLOWED BY A BRIEF NAME FOR THE DATASET, (SEE EXAMPLE)
C--     AND A I4,6X,10F7.0 FORMAT FOR SUBSEQUENT CARDS.
C--     MEANS ARE FIRST SET OF CARDS,
C--     FACTORED CORRELATION MATRIX FOLLOWS,
C--     DEVIATAES AS A % OF MEAN OR PREDICTED VALUES ARE LAST.
C--     NOTE: ENTER ONLY THE UPPER RIGHT TRIANGLE OF THE FACTORED MATRIX.
C-------------------------------------------------------------------------
C--     THIS IS A SAMPLE DATASET FOR TWO INDEPENDENT VARIABLES
C--  12345678901234567890123456789012345678901234567890123456789012345678901234567890
C--
C--  0002 100  10  THIS IS A TEST DATA SET FOR 2 INDEPENDENT VARIABLES
C--  0001MEAN 1  38.85
C--  0002MEAN 2  63.44
C--  0001MATRIX 1.0000 0.0000
C--  0002MATRIX 1.0000
C--  0001DEVT1 -0.2780-0.1904 0.2855 0.0333 0.2106 0.2973 0.0274 0.0870-0.4727-0.0041
C--  0002DEVT2 -0.4823 0.0522 0.0695 0.0148 0.4006 0.2936 0.2210 0.1651-0.5452-0.1944
C-------------------------------------------------------------------------
      DIMENSION R(80,40),DEV(82,40),XBAR(82,40),CV(82),SV(82),ISND(40),
     1CSND(40),CUNIF(40),RDEV(40),DEVX(82),RESULT(100,40)
      REAL ISND
C--   COMMON FOR THE GAUSE SUBROUTINE
      COMMON /JWR3/ RNAR(128),RKK
      REAL*8 RKK
C--   COMMON FOR THE ZTABLE FUNCTION
      COMMON /JWRZT/ Z(390,2),ZTABLE(390)
      EXTERNAL ZDATA
C-------------------------------------------------------------------------
C--   CDIST.INP: THIS IS WHERE THE INPUT DATA RESIDE
      IN = 10
      OPEN (UNIT=IN,FILE='CDIST.INP',STATUS='UNKNOWN',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
C--   ISND.OUT OUTPUT FILE
      IO42 = 12
      OPEN (UNIT=IO42,FILE='ISND.OUT',STATUS='SCRATCH')
      CLOSE (IO42)
      OPEN (UNIT=IO42,FILE='ISND.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
C--   CSND.OUT OUTPUT FILE FOR
      IO43 = 13
      OPEN (UNIT=IO43,FILE='CSND.OUT',STATUS='SCRATCH')
      CLOSE (IO43)
      OPEN (UNIT=IO43,FILE='CSND.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
C--   CUNIF.OUT OUTPUT FILE FOR
      IO44 = 14
      OPEN (UNIT=IO44,FILE='CUNIF.OUT',STATUS='SCRATCH')
      CLOSE (IO44)
      OPEN (UNIT=IO44,FILE='CUNIF.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
C--   RESULT.OUT OUTPUT FILE FOR
      IO45 = 15
      OPEN (UNIT=IO45,FILE='RESULT.OUT',STATUS='SCRATCH')
      CLOSE (IO45)
      OPEN (UNIT=IO45,FILE='RESULT.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
C--   RDEV.OUT OUTPUT FILE FOR
      IO46 = 16
      OPEN (UNIT=IO46,FILE='RDEV.OUT',STATUS='SCRATCH')
      CLOSE (IO46)
      OPEN (UNIT=IO46,FILE='RDEV.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
C--   CDIST.DBG OUTPUT FILE FOR
      IO47 = 17
      OPEN (UNIT=IO47,FILE='CDIST.DBG',STATUS='SCRATCH')
      CLOSE (IO47)
      OPEN (UNIT=IO47,FILE='CDIST.DBG',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
1     FORMAT (3I4)
2     FORMAT (2I2,6X,10F7.0)
3     FORMAT(20I4)
4     FORMAT(/,'DEV(I,K) FOR EACH K')
5     FORMAT(T2,I4,40F7.4)
6     FORMAT(/,'UNIFORM CDF VALUES',/,100(I4,T6,F7.4))
11    FORMAT('CDIST.OUT',T16,2I4,40F10.4)
12    FORMAT(T2,2I4,40F7.3)
13    FORMAT(T7,'(c) Copyright 1993 Texas Agricultural Experiment Statio
     1n',//)
14    FORMAT(/,T7,'C-DIST  VERSION 1.1   RELEASE DATE JULY 7, 1993')
15    FORMAT (F13.0)
```

```
 16    FORMAT (I4,F13.3)
C--    WRITE THE NAME AND COPYRIGHT TO THE SCREEN
       WRITE(*,14)
       WRITE(*,13)
C--    READ NO. OF VARIABLES & THE NO. OF ITERATIONS & NO. OF PTS IN PDF
       READ(IN,1) IVAR,NOITER,NOBS
C--    SET UP THE DATA FOR STAT
       WRITE(1045,1) IVAR , NOITER
       WRITE(*,*) 'NO. OF VARIABLES IS ',IVAR
       WRITE(*,*) 'NO. OF ITERATIONS IS ', NOITER
C--    READ THE ANNUAL MEANS BY VARIABLE
C--    FORMAT IS 2I2 FOLLOWED BY 6X AND UP TO 10 YEARS OF MEANS
       DO 110 K=1,IVAR
       READ (IN,2) IC,IR,(XBAR(J,K),J=1,10)
       WRITE(1047,12) IC,IR,(XBAR(J,K),J=1,10)
110    CONTINUE
       WRITE(1047,*) ' END OF THE MEANS'

       DO 120 I=1,40
       DO 120 J=1,40
120    R(I,J) = 0.0
C--    READ R MATRIX INTO COLUMNS 1-40
C--    FACTORED MATRIX CORRELATION MATRIX FOR THE VARIABLES READ AS THE
C--    UPPER RIGHT TRIANGLE MATRIX WITH THE ROW NUMBER ON THE CARDS
       IF(IVAR.GE.30) THEN
       DO 130 K=1,IVAR
       WRITE(1047,*) '  K EQUALS',K
       IF(K.GE.31) THEN
       READ (IN,2) IC,IR,(R(K,J),J=K,IVAR)
       WRITE(1047,12) IC,IR,(R(K,J),J=K,IVAR)
       GO TO 130
       END IF
       IF(K.GE.21.AND.K.LE.30) THEN
       IEND = K + 9
       ISTART = IEND + 1
       READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
       WRITE(1047,12) IC,IR,(R(K,J),J=K,IEND)
       READ(IN,2) IC,IR,(R(K,J),J=ISTART,IVAR)
       WRITE(1047,12) IC,IR,(R(K,J),J=ISTART,IVAR)
       GO TO 130
       END IF
       IF(K.GE.11.AND.K.LE.20) THEN
       IEND = K + 9
       ISTRT1= IEND + 1
       IEND2 = ISTRT1 + 9
       ISTRT2 = IEND2 + 1
       READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
       WRITE(1047,12) IC,IR,(R(K,J),J=K,IEND)
       READ(IN,2) IC,IR,(R(K,J),J=ISTRT1,IEND2)
       WRITE(1047,12) IC,IR,(R(K,J),J=ISTRT1,IEND2)
       READ(IN,2) IC,IR,(R(K,J),J=ISTRT2,IVAR )
       WRITE(1047,12) IC,IR,(R(K,J),J=ISTRT2,IVAR )
       GO TO 130
       END IF
       IF(K.GE.1.AND.K.LE.10) THEN
       IEND = K + 9
       ISTRT1= IEND + 1
       IEND2 = ISTRT1 + 9
       ISTRT2 = IEND2 + 1
       IEND3  = ISTRT2 + 9
       ISTRT3 = IEND3 + 1
       READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
       WRITE(1047,12) IC,IR,(R(K,J),J=K,IEND)
       READ(IN,2) IC,IR,(R(K,J),J=ISTRT1,IEND2)
       WRITE(1047,12) IC,IR,(R(K,J),J=ISTRT1,IEND2)
       READ(IN,2) IC,IR,(R(K,J),J=ISTRT2,IEND3)
       WRITE(1047,12) IC,IR,(R(K,J),J=ISTRT2,IEND3)
       READ(IN,2) IC,IR,(R(K,J),J=ISTRT3,IVAR )
       WRITE(1047,12) IC,IR,(R(K,J),J=ISTRT3,IVAR )
       END IF
130    CONTINUE
       END IF   ! FOR IVAR EQUAL 40

       IF(IVAR.LE.10) THEN
          DO 131 K=1,IVAR
          WRITE(1047,*) '  K EQUALS',K
          READ (IN,2) IC,IR,(R(K,J),J=K,IVAR)
          WRITE(1047,12) IC,IR,(R(K,J),J=K,IVAR)
131       CONTINUE
       END IF    ! FOR IVAR EQUAL 1 TO 10

       IF(IVAR.GT.10.AND.IVAR.LE.20) THEN
          DO 132 K=1,IVAR
          WRITE(1047,*) '  K EQUALS',K
          IF(K.GE.11) THEN
          READ (IN,2) IC,IR,(R(K,J),J=K,IVAR)
          WRITE(1047,12) IC,IR,(R(K,J),J=K,IVAR)
          GO TO 132
```

```
          IF(K.GE.1.AND.K.LE.10) THEN
          IEND = K + 9
          ISTART = IEND + 1
          READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
          WRITE(IO47,12) IC,IR,(R(K,J),J=K,IEND)
          READ(IN,2) IC,IR,(R(K,J),J=ISTART,IVAR)
          WRITE(IO47,12) IC,IR,(R(K,J),J=ISTART,IVAR)
          GO TO 132
          END IF
 132      CONTINUE
       END IF      ! FOR IVAR EQUAL 11 TO 20

       IF(IVAR.GT.20.AND.IVAR.LE.30) THEN
          DO 134 K=1,IVAR
          WRITE(IO47,*) '  K EQUALS',K
          IF(K.GE.21) THEN
          READ (IN,2) IC,IR,(R(K,J),J=K,IVAR)
          WRITE(IO47,12) IC,IR,(R(K,J),J=K,IVAR)
          GO TO 134
          END IF
          IF(K.GE.11.AND.K.LE.20) THEN
          IEND = K + 9
          ISTART = IEND + 1
          READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
          WRITE(IO47,12) IC,IR,(R(K,J),J=K,IEND)
          READ(IN,2) IC,IR,(R(K,J),J=ISTART,IVAR)
          WRITE(IO47,12) IC,IR,(R(K,J),J=ISTART,IVAR)
          GO TO 134
          END IF
          IF(K.GE.1.AND.K.LE.10) THEN
          IEND = K + 9
          ISTRT1= IEND + 1
          IEND2 = ISTRT1 + 9
          ISTRT2 = IEND2 + 1
          READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
          WRITE(IO47,12) IC,IR,(R(K,J),J=K,IEND)
          READ(IN,2) IC,IR,(R(K,J),J=ISTRT1,IEND2)
          WRITE(IO47,12) IC,IR,(R(K,J),J=ISTRT1,IEND2)
          READ(IN,2) IC,IR,(R(K,J),J=ISTRT2,IVAR )
          WRITE(IO47,12) IC,IR,(R(K,J),J=ISTRT2,IVAR )
          GO TO 134
          END IF
 134      CONTINUE
       END IF    ! FOR IVAR EQUAL 1 TO 30

       WRITE(IO47,*) 'END OF CORRELATION MATRIX'

C--    READ % DEVIATES INTO DEV(.)
C--    EMPIRICAL PDF FOR STOCHASTIC VARIABLES, ASSUME 10 POINTS IN PDF
C--    AND THE DEVIATES ARE A % OF THE MEAN OR PREDICTED TREND VALUES
       DO 160 K=1,IVAR
       DO 160 I=1,NOBS           ! FOR THE SINGLE COLUMN ONLY
       READ (IN,15) DEV(I,K)     ! READ DEVIATES AS A SINGLE COLUMN PER VARIABLE
       WRITE(IO47,16) DEV(I,K)
C       READ(IN,2) IC,IR,(DEV(I,K),I=1,NOBS)
C       WRITE(IO47,12) IC,IR,(DEV(I,K),I=1,NOBS)
160    CONTINUE
C--------------------------------

C--    DEVELOP THE Z-TABLE FROM ZDATA BLOCK DATA FILE
       DO 170 I=1,390
       Z(I,2) = ZTABLE(I)
       Z(I,1) = I * 0.01 - 0.01
170    CONTINUE
C--------------------------------

C--    DEVO. THE EMPIRICAL PDF'S FOR THE J-DISTRIBUTION
C--    SORT THE DEVIATES AND FIX THE END VALUES
       DO 270 J=1,IVAR
       ICOL = J
C--    MOVE DEVIATES TO ANOTHER ARRAY FOR STORAGE
       DO 210 I=1,NOBS
210    CV(I) = DEV(I,ICOL)
       AMAX = -99999999.9
       DO 220 I=1,NOBS
       IF(AMAX.LT.CV(I)) AMAX = CV(I)
220    CONTINUE
       IF(AMAX.NE.CV(NOBS)) THEN
       DO 250 K=1,NOBS
       SV(K) = AMAX
       DO 230 I=1,NOBS
       IF(SV(K).GE.CV(I)) SV(K) = CV(I)
230    CONTINUE
       IF(K.EQ.NOBS) GO TO 250
       DO 240 I=1,NOBS
       IF(SV(K).NE.CV(I)) GO TO 240
       IF(AMAX.GE.0.0) CV(I) = AMAX * 10.0
       IF(AMAX.LT.0.0) CV(I) = AMAX *(-10.0)
```

```
240   CONTINUE
250   CONTINUE
      ELSE   ! DEVIATES ARE SORTED ALREADY.
      DO 255 I=1,NOBS
255   SV(I) = CV(I)
      END IF
      DO 260 I=1,NOBS
260   DEV(I,ICOL) = SV(I)
270   CONTINUE
      WRITE(IO47,4)
      DO 271 K=1,IVAR
      WRITE(IO47,5) K,(DEV(I,K),I=1,NOBS)
271   CONTINUE
C--   FILL IN THE PROBABILITY
      DELTA = 1.0 / FLOAT(NOBS-1)
      DEVX(1) = 0.0
      IEND = NOBS + 1
      DO 275 I=2,NOBS
275   DEVX(I) = DEVX(I-1) + DELTA
      WRITE(IO47,6) (I,DEVX(I),I=1,NOBS)
C--   RANDOM NUMBER GENERATOR SEED FOR GAUSE
      NAR = 31415
      I = 1
C--   GENERATE RANDOM NUMBER FOR THE J - DISTRIBUTION EMPIRICAL PDF
      DO 400 ITER=1,NOITER
      WRITE(IO47,*) 'ITERATION',ITER
      DO 280 KK=1,IVAR          ! GENERATE IVAR RANDON STD. NORMAL DEVIATES
      CALL GAUSE(GAUSES,NAR)
      ISND(KK) = GAUSES         ! INDEPENDENT STD. NORMAL DEVIATES
      CSND(KK) = 0.0
280   CONTINUE
      DO 300 KK=1,IVAR          ! CORRELATE THE STD. NORMAL DEVIATES
      CSND(KK) = 0.0
      DO 290 LL=1,IVAR
      IF(KK.GT.LL) GO TO 290
      CSND(KK) = CSND(KK) + R(KK,LL) * ISND(LL)
290   CONTINUE
300   CONTINUE

C--   TRANSFORM CORRELATED DEVIATES TO A UNIFORM
      DO 310 K=1,IVAR
      E1   =   CSND(K)
310   CUNIF(K)  =  ERFZ(E1)     ! CONVERT CORR. DEVIATES TO UNIFORM RANDOM NOS.
C--                             ! USING THE Z-TABLE LOOKUP FUNCTION
C--   INTERPOLATE THE EMPIRICAL PDFS
      DO 330 K=1,IVAR
      RANDNO = CUNIF(K)
      DO 320 IR=2,NOBS          ! DETER INTERVAL THAT THE RANDOM NO. IS IN
      IJ=IR-1
      IF(RANDNO.GT.DEVX(IJ).AND.RANDNO.LE.DEVX(IR)) GO TO 325
320   CONTINUE
325   RDEV(K) = (( DEV(IR,K)-DEV(IJ,K) )*( (RANDNO-DEVX(IJ)) / DELTA) )
     1        + DEV(IJ,K)       ! INTERPOLATE TO GET THE RANDOM % DEV FROM MEAN
330   CONTINUE
C--   CALCULATE FINAL VALUE FOR VARIABLES USING MEAN AND THE RANDON % DEVIATE
      DO 340 K=1,IVAR
      RESULT(I,K) = XBAR(I,K) * ( 1.0 + RDEV(K) )
340   CONTINUE
      WRITE(IO42,11) I,ITER,(ISND(K),K=1,IVAR)
      WRITE(IO43,11) I,ITER,(CSND(K),K=1,IVAR)
      WRITE(IO44,11) I,ITER,(CUNIF(K),K=1,IVAR)
      WRITE(IO45,11) I,ITER,(RESULT(I,K),K=1,IVAR)
      WRITE(IO46,11) I,ITER,(RDEV(K),K=1,IVAR)
400   CONTINUE                  ! PHYSICAL END OF THE ITERATION LOOP
      STOP
      END
C-------------------------------------------------------------------------
C     USER MUST PROVIDE GAUSE AND RANG AT THIS POINT, SEE TEST.

C-------------------------------------------------------------------------
C--   ZTABLE SECTION
C--   FUNCTION  ***********************************************************
      FUNCTION ERFZ(X)
      COMMON /JWRZT/ Z(390,2),ZTABLE(390)
      Y = ABS(X)
C--   PREPROCESS THE DATA TO REDUCE TEST LOOPS
      IF(Y.LE.1.0) GO TO 90
      IF(Y.GT.1.0 .AND. Y.LT.2.0) GO TO 150
      IF(Y.GE.2.0 .AND. Y.LT.3.0) GO TO 250
      IF(Y.GE.3.0) GO TO 350
90    CONTINUE
C--   Y IS LESS THAN 1.0
      DO 100 I=2,101
      J = I-1
      IF(Y.GE.Z(J,1) .AND. Y.LT.Z(I,1)) THEN
      ERFZ = Z(J,2) + ( (Z(I,1)-Y)/(Z(I,1)-Z(J,1)) ) * (Z(I,2)-Z(J,2))
C--   ERFZ = Z(J,2)
      GO TO 500
      END IF
```

```
100   CONTINUE
C--   Y IS BETWEEN 1.0 AND 2.0
150   CONTINUE
      DO 200 I=100,201
      J = I-1
      IF(Y.GE.Z(J,1) .AND. Y.LT.Z(I,1)) THEN
      ERFZ = Z(J,2) + ( (Z(I,1)-Y)/(Z(I,1)-Z(J,1)) ) * (Z(I,2)-Z(J,2))
C--   ERFZ = Z(J,2)
      GO TO 500
      END IF
200   CONTINUE
C--   Y IS BETWEEN 2.0 AND 3.0
250   CONTINUE
      DO 300 I=200,301
      J = I-1
      IF(Y.GE.Z(J,1) .AND. Y.LT.Z(I,1)) THEN
      ERFZ = Z(J,2) + ( (Z(I,1)-Y)/(Z(I,1)-Z(J,1)) ) * (Z(I,2)-Z(J,2))
C--   ERFZ = Z(J,2)
      GO TO 500
      END IF
300   CONTINUE
C--   Y IS GREATER THAN 3.0
350   CONTINUE
      DO 400 I=300,390
      J = I-1
      IF(Y.GE.Z(J,1) .AND. Y.LT.Z(I,1)) THEN
      ERFZ = Z(J,2) + ( (Z(I,1)-Y)/(Z(I,1)-Z(J,1)) ) * (Z(I,2)-Z(J,2))
C--   ERFZ = Z(J,2)
      GO TO 500
      END IF
400   CONTINUE
      ERFZ = 1.0
500   CONTINUE
      IF(X.GT.0.0D0) THEN
      ERFZ = ERFZ + 0.50D0
      ELSE
      ERFZ = 0.5D0 - ERFZ
      END IF
      RETURN
      END

C-----------------------------------------------------------------------
      BLOCK DATA  ZDATA
C--   ZTABLE LOOK-UP TABLE DATA
      COMMON /JWRZT/ Z(390,2),ZTABLE(390)
      DATA ZTABLE /
     1 0.0000,0.0040,0.0080,0.0120,0.0160,0.0199,0.0239,0.0279,0.0319,
     1 0.0359,0.0398,0.0438,0.0478,0.0517,0.0557,0.0596,0.0636,0.0675,
     1 0.0714,0.0754,0.0793,0.0832,0.0871,0.0910,0.0948,0.0987,0.1026,
     1 0.1064,0.1103,0.1141,0.1179,0.1217,0.1255,0.1293,0.1331,0.1368,
     1 0.1406,0.1443,0.1480,0.1517,0.1554,0.1591,0.1628,0.1664,0.1700,
     1 0.1736,0.1772,0.1808,0.1844,0.1879,0.1915,0.1950,0.1985,0.2019,
     1 0.2054,0.2088,0.2123,0.2157,0.2190,0.2224,0.2258,0.2291,0.2324,
     1 0.2357,0.2389,0.2422,0.2454,0.2486,0.2518,0.2549,0.2580,0.2612,
     1 0.2642,0.2673,0.2704,0.2734,0.2764,0.2794,0.2823,0.2852,0.2881,
     1 0.2910,0.2939,0.2967,0.2996,0.3023,0.3051,0.2079,0.3106,0.3133,
     1 0.3159,0.3186,0.3212,0.3238,0.3264,0.3289,0.3315,0.3340,0.3365,
     1 0.3389,0.3413,0.3438,0.3461,0.3485,0.3508,0.3531,0.3554,0.3577,
     1 0.3599,0.3621,0.3643,0.3665,0.3686,0.3708,0.3729,0.3749,0.3770,
     1 0.3790,0.3810,0.3830,0.3849,0.3869,0.3888,0.3907,0.3925,0.3944,
     1 0.3962,0.3980,0.3997,0.4015,0.4032,0.4049,0.4066,0.4082,0.4099,
     1 0.4115,0.4131,0.4147,0.4162,0.4177,0.4192,0.4207,0.4222,0.4236,
     1 0.4251,0.4265,0.4279,0.4292,0.4306,0.4319,0.4332,0.4345,0.4357,
     1 0.4370,0.4382,0.4394,0.4406,0.4418,0.4430,0.4441,0.4452,0.4463,
     1 0.4474,0.4485,0.4495,0.4505,0.4515,0.4525,0.4535,0.4545,0.4554,
     1 0.4564,0.4573,0.4582,0.4591,0.4599,0.4608,0.4616,0.4625,0.4633,
     1 0.4641,0.4649,0.4656,0.4664,0.4671,0.4678,0.4686,0.4693,0.4700,
     1 0.4706,0.4713,0.4719,0.4726,0.4732,0.4738,0.4744,0.4750,0.4756,
     1 0.4762,0.4767,0.4773,0.4778,0.4783,0.4788,0.4793,0.4798,0.4803,
     1 0.4808,0.4812,0.4817,0.4821,0.4826,0.4830,0.4834,0.4838,0.4842,
     1 0.4846,0.4850,0.4854,0.4857,0.4861,0.4865,0.4868,0.4871,0.4875,
     1 0.4878,0.4881,0.4884,0.4887,0.4890,0.4893,0.4896,0.4898,0.4901,
     1 0.4904,0.4906,0.4909,0.4911,0.4913,0.4916,0.4918,0.4920,0.4922,
     1 0.4925,0.4927,0.4929,0.4931,0.4932,0.4934,0.4936,0.4938,0.4940,
     1 0.4941,0.4943,0.4945,0.4946,0.4948,0.4949,0.4951,0.4952,0.4953,
     1 0.4955,0.4956,0.4957,0.4959,0.4960,0.4961,0.4962,0.4963,0.4964,
     1 0.4965,0.4966,0.4967,0.4968,0.4969,0.4970,0.4971,0.4972,0.4973,
     1 0.4974,0.4974,0.4975,0.4976,0.4977,0.4977,0.4978,0.4979,0.4980,
     1 0.4980,0.4981,0.4981,0.4982,0.4983,0.4983,0.4984,0.4984,0.4985,
     1 0.4985,0.4986,0.4986,0.4987,0.4987,0.4987,0.4988,0.4988,0.4989,
     1 0.4989,0.4989,0.4990,0.4990,0.4990,0.4991,0.4991,0.4991,0.4992,
     1 0.4992,0.4992,0.4992,0.4993,0.4993,0.4993,0.4993,0.4994,0.4994,
     1 0.4994,0.4994,0.4994,0.4995,0.4995,0.4995,0.4995,0.4995,0.4996,
     1 0.4996,0.4996,0.4996,0.4996,0.4996,0.4996,0.4997,0.4997,0.4997,
     1 0.4997,0.4997,0.4997,0.4997,0.4997,0.4997,0.4998,0.4998,0.4998,
     1 0.4998,0.4998,0.4998,0.4998,0.4998,0.4998,0.4998,0.4998,0.4998,
     1 0.4998,0.4999,0.4999,0.4999,0.4999,0.4999,0.4999,0.4999,0.4999,
     1 0.4999,0.4999,0.4999,0.4999,0.4999,0.4999,0.4999,0.4999,0.4999,
```

```
     * 0.5000,0.5000,0.5000/
      END
C-----------------------------------------------------------------------
    SAMPLE DATA FOR CDIST

   11000  55    EDWARDS AQUIFER DATA WITH 55 POINTS ON THE PDF AS A % OF THE MEAN.
0001MEAN 1 635.52
27 1MATRIX 1.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
     -0.7174
      0.9798
      0.4313
      0.3695
     -0.3191
     -0.3722
     -0.5141
      0.3386
     -0.1223
     -0.5703
     -0.1174
     -0.1695
     -0.1250
     -0.3350
     -0.7194
     -0.2005
     -0.6850
     -0.7799
     -0.5665
     -0.7363
     -0.7449
     -0.6979
     -0.9312
      0.7979
      1.6927
      0.0864
      0.2979
      0.1284
     -0.6233
     -0.7314
     -0.3498
     -0.0189
     -0.0320
     -0.2659
      0.3921
     -0.0394
      0.0411
      0.4560
      0.1902
      1.3391
      0.0362
      0.5311
      0.4069
      0.4980
     -0.2093
      0.7589
     -0.3605
      1.2790
     -0.3353
     -0.3390
     -0.6899
      0.5787
      0.8156
      2.1526
     -0.4406
C-----------------------------------------------------------------------
CDIST CAN BE RUN WITH MULTIPLE RANDOM VARIABLES, SUCH AS:
     NOTE THE DEVIATES ARE ENTERED AS 10 INTERVALS AND ARE ACROSS THE LINE

0002 100   TEST DATA FOR 2 INDEPENDENT VARIABLES
0001VAR# 1  38.85
0002VAR# 2  63.44
0001MATRIX 1.0000 0.0000
0002MATRIX 1.0000
0001DEVT  -0.2780-0.1904 0.2855 0.0333 0.2106 0.2973 0.0274 0.0870-0.4727-0.0041
0002DEVT  -0.4823 0.0522 0.0695 0.0148 0.4006 0.2936 0.2210 0.1651-0.5452-0.1944
```

```
C--   JDIST A PROGRAM TO DEMONSTRATE THE UNIVARIATE AND MULTIVATIATE
C--   J - DISTRIBUTION.
C--   DIMENSIONS ARE SET FOR 1000 ITERATIONS AND 20 VARIABLES
C--                        DEVELOPED BY JWR 7/1/93
C-----------------------------------------------------------------------
C--     THIS IS A SAMPLE DATASET FOR TWO INDEPENDENT VARIABLES
C--     CODE THE DATA BY COLUMN USING A 3I4 FORMAT FOR CARD ONE
C--     FOLLOWED BY A BRIEF NAME FOR THE DATASET, (SEE EXAMPLE)
C--     AND A I4,6X,10F7.0 FORMAT FOR SUBSEQUENT CARDS.
C--     ENTER ONLY THE UPPER RIGHT TRIANGLE OF THE FACTORED MATRIX.
C-----------------------------------------------------------------------
C--  123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
C--  2 VARIABLES;  100 ITERATIONS;   AND 10 PTS ON THE PDF.
C--
C--  0002 100  10   THIS IS A TEST DATA SET FOR 2 INDEPENDENT VARIABLES
C--  0001MEAN 1  38.85
C--  0002MEAN 2  63.44
C--  0001MATRIX 1.0000 0.0000
C--  0002MATRIX 1.0000
C--  0001DEVT1 -0.2780-0.1904 0.2855 0.0333 0.2106 0.2973 0.0274 0.0870-0.4727-0.0041
C--  0002DEVT2 -0.4823 0.0522 0.0695 0.0148 0.4006 0.2936 0.2210 0.1651-0.5452-0.1944
C-----------------------------------------------------------------------
      DIMENSION R(20,20),DEV(82,20),XBAR(82,20),CV(82),SV(82),ISND(80),
     1CSND(20),CUNIF(20),RDEV(20),DEVX(82),RESULT(2,20)
      REAL ISND
C--   COMMON FOR THE GAUSE SUBROUTINE
      COMMON /JWR3/ RNAR(128),RKK
      REAL*8 RKK
C--   COMMON FOR THE ZTABLE FUNCTION
      COMMON /JWRZT/ Z(390,2),ZTABLE(390)
      EXTERNAL ZDATA
C-----------------------------------------------------------------------
C--
C--   FPDF.INP: THIS IS WHERE THE INPUT DATA RESIDE
      IN = 10
      OPEN (UNIT=IN,FILE='JDIST.INP',STATUS='UNKNOWN',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
C--
C--   ISND.OUT OUTPUT FILE
      IO42 = 12
      OPEN (UNIT=IO42,FILE='ISND.OUT',STATUS='SCRATCH')
      CLOSE (IO42)
      OPEN (UNIT=IO42,FILE='ISND.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')

C--   CSND.OUT OUTPUT FILE FOR
      IO43 = 13
      OPEN (UNIT=IO43,FILE='CSND.OUT',STATUS='SCRATCH')
      CLOSE (IO43)
      OPEN (UNIT=IO43,FILE='CSND.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')

C--   CUNIF.OUT OUTPUT FILE FOR
      IO44 = 14
      OPEN (UNIT=IO44,FILE='CUNIF.OUT',STATUS='SCRATCH')
      CLOSE (IO44)
      OPEN (UNIT=IO44,FILE='CUNIF.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')

C--   RESULT.OUT OUTPUT FILE FOR
      IO45 = 15
      OPEN (UNIT=IO45,FILE='RESULT.OUT',STATUS='SCRATCH')
      CLOSE (IO45)
      OPEN (UNIT=IO45,FILE='RESULT.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')

C--   RDEV.OUT OUTPUT FILE FOR
      IO46 = 16
      OPEN (UNIT=IO46,FILE='RDEV.OUT',STATUS='SCRATCH')
      CLOSE (IO46)
      OPEN (UNIT=IO46,FILE='RDEV.OUT',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')

C--   JDIST.DBG OUTPUT FILE FOR
      IO47 = 17
      OPEN (UNIT=IO47,FILE='JDIST.DBG',STATUS='SCRATCH')
      CLOSE (IO47)
      OPEN (UNIT=IO47,FILE='JDIST.DBG',STATUS='NEW',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')

1     FORMAT (20I4)
2     FORMAT (2I2,6X,10F7.0)
3     FORMAT(20I4)
4     FORMAT(/,'DEV(I,K) FOR EACH K')
5     FORMAT(T2,I4,60F7.4)
6     FORMAT(/,'UNIFORM CDF VALUES',/,T6,20F7.4)
11    FORMAT('JDIST.OUT',T16,2I4,40F10.4)
12    FORMAT(T2,2I4,40F7.3)
13    FORMAT(T7,'(c) Copyright [illegible]
```

```
     1n',//)
 14   FORMAT(/,T7,'J-DIST  VERSION 1.1   RELEASE DATE JULY 5, 1993')
 15   FORMAT (F13.0)
 16   FORMAT (I4,F13.0)
C--   WRITE THE NAME AND COPYRIGHT TO THE SCREEN
      WRITE(*,14)
      WRITE(*,13)
C--   READ NO. OF VARIABLES & THE NO. OF ITERATIONS
      READ(IN,1) IVAR,NOITER,NOBS
C--   SET UP THE DATA FOR STAT
      WRITE(IO45,1) IVAR , NOITER,NOBS
      WRITE(*,*) 'NO. OF VARIABLES IS ',IVAR
      WRITE(*,*) 'NO. OF ITERATIONS IS ', NOITER
      WRITE(*,*) 'NO. OF OBSERVATIONS IN THE PDF IS ', NOBS
C--   READ THE ANNUAL MEANS BY VARIABLE
C--   FORMAT IS 2I2 FOLLOWED BY 6X AND UP TO 10 YEARS OF MEANS
      DO 110 K=1,IVAR
      READ (IN,2) IC,IR,(XBAR(J,K),J=1,10)
      WRITE(IO47,12) IC,IR,(XBAR(J,K),J=1,10)
110   CONTINUE
      WRITE(IO47,*) ' END OF THE MEANS'
      DO 120 I=1,IVAR
      DO 120 J=1,IVAR
120   R(I,J) = 0.0
C--   READ R MATRIX INTO COLUMNS 1-40
C--   FACTORED MATRIX CORRELATION MATRIX FOR THE VARIABLES READ AS THE
C--   UPPER RIGHT TRIANGLE MATRIX WITH THE ROW NUMBER ON THE CARDS
      IF(IVAR.GE.30) THEN
      DO 130 K=1,IVAR
      WRITE(IO47,*) '  K EQUALS',K
      IF(K.GE.31) THEN
      READ (IN,2) IC,IR,(R(K,J),J=K,IVAR)
      WRITE(IO47,12) IC,IR,(R(K,J),J=K,IVAR)
      GO TO 130
      END IF
      IF(K.GE.21.AND.K.LE.30) THEN
      IEND = K + 9
      ISTART = IEND + 1
      READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
      WRITE(IO47,12) IC,IR,(R(K,J),J=K,IEND)
      READ(IN,2) IC,IR,(R(K,J),J=ISTART,IVAR)
      WRITE(IO47,12) IC,IR,(R(K,J),J=ISTART,IVAR)
      GO TO 130
      END IF
      IF(K.GE.11.AND.K.LE.20) THEN
      IEND = K + 9
      ISTRT1= IEND + 1
      IEND2 = ISTRT1 + 9
      ISTRT2 = IEND2 + 1
      READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
      WRITE(IO47,12) IC,IR,(R(K,J),J=K,IEND)
      READ(IN,2) IC,IR,(R(K,J),J=ISTRT1,IEND2)
      WRITE(IO47,12) IC,IR,(R(K,J),J=ISTRT1,IEND2)
      READ(IN,2) IC,IR,(R(K,J),J=ISTRT2,IVAR )
      WRITE(IO47,12) IC,IR,(R(K,J),J=ISTRT2,IVAR )
      GO TO 130
      END IF
      IF(K.GE.1.AND.K.LE.10) THEN
      IEND = K + 9
      ISTRT1= IEND + 1
      IEND2 = ISTRT1 + 9
      ISTRT2 = IEND2 + 1
      IEND3  = ISTRT2 + 9
      ISTRT3 = IEND3 + 1
      READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
      WRITE(IO47,12) IC,IR,(R(K,J),J=K,IEND)
      READ(IN,2) IC,IR,(R(K,J),J=ISTRT1,IEND2)
      WRITE(IO47,12) IC,IR,(R(K,J),J=ISTRT1,IEND2)
      READ(IN,2) IC,IR,(R(K,J),J=ISTRT2,IEND3)
      WRITE(IO47,12) IC,IR,(R(K,J),J=ISTRT2,IEND3)
      READ(IN,2) IC,IR,(R(K,J),J=ISTRT3,IVAR )
      WRITE(IO47,12) IC,IR,(R(K,J),J=ISTRT3,IVAR )
      END IF
130   CONTINUE
      END IF   ! FOR IVAR EQUAL 40

      IF(IVAR.LE.10) THEN
         DO 131 K=1,IVAR
         WRITE(IO47,*) '  K EQUALS',K
         READ (IN,2) IC,IR,(R(K,J),J=K,IVAR)
         WRITE(IO47,12) IC,IR,(R(K,J),J=K,IVAR)
131      CONTINUE
      END IF    ! FOR IVAR EQUAL 1 TO 10

      IF(IVAR.GT.10.AND.IVAR.LE.20) THEN
         DO 132 K=1,IVAR
         WRITE(IO47,*) '  K EQUALS',K
         IF(K.GE.11) THEN
```

```
          WRITE(1047,12) IC,IR,(R(K,J),J=K,IVAR)
          GO TO 132
          END IF
          IF(K.GE.1.AND.K.LE.10) THEN
          IEND = K + 9
          ISTART = IEND + 1
          READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
          WRITE(1047,12) IC,IR,(R(K,J),J=K,IEND)
          READ(IN,2) IC,IR,(R(K,J),J=ISTART,IVAR)
          WRITE(1047,12) IC,IR,(R(K,J),J=ISTART,IVAR)
          GO TO 132
          END IF
 132      CONTINUE
       END IF       ! FOR IVAR EQUAL 11 TO 20

       IF(IVAR.GT.20.AND.IVAR.LE.30) THEN
          DO 134 K=1,IVAR
          WRITE(1047,*) '  K EQUALS',K
          IF(K.GE.21) THEN
          READ (IN,2) IC,IR,(R(K,J),J=K,IVAR)
          WRITE(1047,12) IC,IR,(R(K,J),J=K,IVAR)
          GO TO 134
          END IF
          IF(K.GE.11.AND.K.LE.20) THEN
          IEND = K + 9
          ISTART = IEND + 1
          READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
          WRITE(1047,12) IC,IR,(R(K,J),J=K,IEND)
          READ(IN,2) IC,IR,(R(K,J),J=ISTART,IVAR)
          WRITE(1047,12) IC,IR,(R(K,J),J=ISTART,IVAR)
          GO TO 134
          END IF
          IF(K.GE.1.AND.K.LE.10) THEN
          IEND = K + 9
          ISTRT1= IEND + 1
          IEND2 = ISTRT1 + 9
          ISTRT2 = IEND2 + 1
          READ(IN,2) IC,IR,(R(K,J),J=K,IEND)
          WRITE(1047,12) IC,IR,(R(K,J),J=K,IEND)
          READ(IN,2) IC,IR,(R(K,J),J=ISTRT1,IEND2)
          WRITE(1047,12) IC,IR,(R(K,J),J=ISTRT1,IEND2)
          READ(IN,2) IC,IR,(R(K,J),J=ISTRT2,IVAR )
          WRITE(1047,12) IC,IR,(R(K,J),J=ISTRT2,IVAR )
          GO TO 134
          END IF
 134      CONTINUE
       END IF    ! FOR IVAR EQUAL 1 TO 30

       WRITE(1047,*) 'END OF CORRELATION MATRIX'

C--    READ % DEVIATES INTO DEV(.)
C--    EMPIRICAL PDF FOR STOCHASTIC VARIABLES, ASSUME 10 POINTS IN PDF
C--    AND THE DEVIATES ARE A % OF THE MEAN OR PREDICTED TREND VALUES
       DO 160 K=1,IVAR
       DO 160 I=1,NOBS           ! ONE COLUMN WRITE
       READ(IN,15) DEV(I,K)      ! ONE COLUMN WRITE
       WRITE(1047,16) DEV(I,K)  ! ONE COLUMN WRITE
C       READ(IN,2) IC,IR,(DEV(I,K),I=1,10)
C       WRITE(1047,12) IC,IR,(DEV(I,K),I=1,10)
160    CONTINUE
C--------------------------------

C--    DEVELOP THE Z-TABLE FROM ZDATA BLOCK DATA FILE
       DO 170 I=1,390
       Z(I,2) = ZTABLE(I)
       Z(I,1) = I * 0.01 - 0.01
170    CONTINUE
C--------------------------------
       NOBS1 = NOBS + 1
       NOBS2 = NOBS + 2
C--    DEVO. THE EMPIRICAL PDF'S FOR THE J-DISTRIBUTION
C--    SORT THE DEVIATES AND FIX THE END VALUES
       DO 270 J=1,IVAR
       ICOL = J
C--    MOVE DEVIATES TO ANOTHER ARRAY FOR STORAGE
       DO 210 I=1,NOBS
210    CV(I+1) = DEV(I,ICOL)
       AMAX = -99999999.9
       DO 220 I=2,NOBS1
       IF(AMAX.LT.CV(I)) AMAX = CV(I)
220    CONTINUE
       DO 250 K=2,NOBS1
       SV(K) = AMAX
       DO 230 I=2,NOBS1
       IF(SV(K).GE.CV(I)) SV(K) = CV(I)
230    CONTINUE
       IF(K.EQ.NOBS1) GO TO 250
       DO 240 I=2,NOBS1
```

```
      IF(AMAX.GE.0.0) CV(I) = AMAX * 10.0
      IF(AMAX.LT.0.0) CV(I) = AMAX *(-10.0)
      GO TO 250
  240 CONTINUE
  250 CONTINUE
C--   ADD THE END POINTS
      SV(1) = SV(2) - 1.0
      SV(NOBS2)= SV(NOBS1)+1.0
      DO 260 I=1,NOBS2
  260 DEV(I,ICOL) = SV(I)
  270 CONTINUE
      WRITE(IO47,4)
      DO 271 K=1,IVAR
      DO 271 I=1,NOBS2          ! ONE COLUMN OF DEV PTS.
      WRITE(IO47,5) I,DEV(I,K)  ! ONE COLUMN OF DEV PTS.
C      WRITE(IO47,5) K,(DEV(I,K),I=1,NOBS2)
  271 CONTINUE
C--   FILL IN THE PROBABILITY
      DELTA = 1.0 / FLOAT(NOBS1)
      DEVX(1) = 0.0
      DO 275 I=2,NOBS2
  275 DEVX(I) = DEVX(I-1) + DELTA
      WRITE(IO47,6) (DEVX(I),I=1,NOBS2)
C--   RANDOM NUMBER GENERATOR SEED FOR GAUSE
      NAR = 31415
      I = 1
C--   GENERATE RANDOM NUMBER FOR THE J - DISTRIBUTION EMPIRICAL PDF
      DO 400 ITER=1,NOITER

      WRITE(IO47,*) 'ITERATION',ITER
      DO 280 KK=1,IVAR          ! GENERATE IVAR RANDON STD. NORMAL DEVIATES
      CALL GAUSE(GAUSES,NAR)
      ISND(KK) = GAUSES         ! INDEPENDENT STD. NORMAL DEVIATES
      CSND(KK) = 0.0
  280 CONTINUE

      DO 300 KK=1,IVAR          ! CORRELATE THE STD. NORMAL DEVIATES
      CSND(KK) = 0.0
      DO 290 LL=1,IVAR
      IF(KK.GT.LL) GO TO 290
      CSND(KK) = CSND(KK) + R(KK,LL) * ISND(LL)
  290 CONTINUE
  300 CONTINUE

C--   TRANSFORM CORRELATED DEVIATES TO A UNIFORM
      DO 310 K=1,IVAR
      E1   =   CSND(K)
  310 CUNIF(K)  =  ERFZ(E1)     ! CONVERT CORR. DEVIATES TO UNIFORM RANDOM NOS.
C--                             ! USING THE Z-TABLE LOOKUP FUNCTION
C--   INTERPOLATE THE EMPIRICAL PDFS
      DO 330 K=1,IVAR
      RANDNO = CUNIF(K)
      DO 320 IR=2,NOBS2         ! DETER INTERVAL THAT THE RANDOM NO. IS IN
      IJ=IR-1
      IF(RANDNO.GT.DEVX(IJ).AND.RANDNO.LE.DEVX(IR)) GO TO 325
  320 CONTINUE
  325 RDEV(K) = (( DEV(IR,K)-DEV(IJ,K) )*( (RANDNO-DEVX(IJ)) / DELTA) )
     1          + DEV(IJ,K)    ! INTERPOLATE TO GET THE RANDOM % DEV FROM MEAN
C--   TEST THE RANDOM DEVIATE AS % OF MEAN FOR END POINT FOLD BACK
      TESTMI = DEV(2 ,K)
      TESTMX = DEV(NOBS1,K)
      IF(RDEV(K).LT.TESTMI)   RDEV(K) = TESTMI
      IF(RDEV(K).GT.TESTMX)   RDEV(K) = TESTMX
  330 CONTINUE
C--   CALCULATE FINAL VALUE FOR VARIABLES USING MEAN AND THE RANDON % DEVIATE
      DO 340 K=1,IVAR
      RESULT(1,K) = XBAR(I,K) * ( 1.0 + RDEV(K) )
  340 CONTINUE
      WRITE(IO42,11) I,ITER,(ISND(K),K=1,IVAR)
      WRITE(IO43,11) I,ITER,(CSND(K),K=1,IVAR)
      WRITE(IO44,11) I,ITER,(CUNIF(K),K=1,IVAR)
      WRITE(IO45,11) I,ITER,(RESULT(1,K),K=1,IVAR)
      WRITE(IO46,11) I,ITER,(RDEV(K),K=1,IVAR)
  400 CONTINUE                  ! PHYSICAL END OF THE ITERATION LOOP
      STOP
      END


C---------------------------------------------------------------------
USER MUST PROVIDE SUBROUTINE GAUSE, FUNCTION RANG, SUBROUTINE ERFZ AND
BLOCK DATA FOR ZTABLE.
C---------------------------------------------------------------------
```

LISTING OF THE INPUT DATA FOR JDIST

```
   11000  55     EDWARDS AQUIFER DATA WITH 55 POINTS ON THE PDF AS A % OF THE MEAN.
0001MEAN 1 635.520
27 1MATRIX 1.0000
     -0.7174
      0.9798
      0.4313
     -0.3695
     -0.3191
     -0.3722
     -0.5141
      0.3386
     -0.1223
     -0.5703
     -0.1174
     -0.1695
     -0.1250
     -0.3350
     -0.7194
     -0.2005
     -0.6850
     -0.7799
     -0.5665
     -0.7363
     -0.7449
     -0.6979
     -0.9312
      0.7979
      1.6927
      0.0864
      0.2979
      0.1284
     -0.6233
     -0.7314
     -0.3498
     -0.0189
     -0.0320
     -0.2659
      0.3921
     -0.0394
      0.0411
      0.4560
      0.1902
      1.3391
      0.0362
      0.5311
      0.4069
      0.4980
     -0.2093
      0.7589
     -0.3605
      1.2790
     -0.3353
     -0.3390
     -0.6899
      0.5787
      0.8156
      2.1526
     -0.4406
```

```
C--   SOLVE ..     FRAN PROGRAM WRITTEN TO SIMULATE RANDOM VALUES FROM
C--   A HYPERBOLIC TANGENT FUNCTION.
C--                                DEVELOPED BY JWR 7-11-1993
C-------------------------------------------------------------------------------
C--   SOLVE ANY FUNCTION BY CUTTING THE DIFFERENCE TO ZERO BY HALF
C--   ENTER THE FUNCTION 4 TIMES IN THE PROGRAM TO GIVE IT 2 STARTING
C--   VALUES AND THE FINAL FUNCTION TO SOLVE
C-------------------------------------------------------------------------------
      COMMON /JWR3/ RNAR(128),RKK
      REAL*8 RKK
      OPEN (UNIT=11,FILE='SOLVE.OUT',STATUS='UNKNOWN',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
      OPEN (UNIT=12,FILE='RESULT.OUT',STATUS='UNKNOWN',
     1FORM='FORMATTED',ACCESS='SEQUENTIAL')
 1    FORMAT(10X,I4,10F10.0)
 2    FORMAT(/,' ANO  = ',T15,F10.5,/,
     1       ' ATEST  = ',T15,F10.5,/,
     1       ' XVALUE = ',T15,F10.5,/,
     2       ' YVALUE = ' ,T15,F10.5,/,
     3       ' NO ITERATIONS EQUALS',I4)
 3    FORMAT(/,' ANO  = ',T15,F10.5,/,
     1        ' TEST VALUE IS    ',T15,F10.4,/,
     1       ' MIN TOLERANCE IS ',T15,F10.4,/,
     2       ' MAX TOLERANCE IS ',T15,F10.4,/,
     3       ' MAX ITERATIONS IS',I4)
 4    FORMAT('  ITER      ANO      ATEST      YVALUE      XVALUE      OUT')
 5    FORMAT(3X,I4,10F10.4)
 6    FORMAT(I4,T10,6F10.4)
 10   FORMAT('HYPERBOLIC TANGENT ',T16,2I4,40F10.4)
 11   FORMAT (3I4)
C--   IDEBUG IS 1 TO PRINT INTERMEDIATE RESULTS TO SCREEN
      IDEBUG = 1
      IDEBUG = 0
      NAR = 31415
      NOBS = 1
      NOITER = 1000
      IVAR = 1
      WRITE(12,11) IVAR , NOITER,NOBS
      WRITE(11,4)
C--   ITERATION LOOP TO GENERATE NOITER RANDOM VALUES
      DO 300 ISET = 1,NOITER
C--   RETURN TO HERE IF THE RANDOM VALUE RESULTED IN AN UNACCEPTABLE ANSWER
 90   CONTINUE
      ANO = RANG(NAR)
      ATEST = 0.5 * (ALOG (ANO/(1.0-ANO) ) )
C--   FAIL SAFE MECHANISM, TO CHECK FOR INFEASIBLE SOLUTIONS
      IF(ATEST.LE.-1.987) GO TO 90                          ! -1.987 IS THE INTERCEPT
      IF(ATEST.LT.0.009.and.ATEST.GT.-0.009) GO TO 90       ! FUNCTION IS UNSTABLE AT ZERO

      IF(ATEST.GT.0.0) THEN
      AMIN = ATEST * 0.999
      AMAX = ATEST * 1.001
      END IF
      IF(ATEST.LT.0.0) THEN
      AMIN = ATEST * 1.001
      AMAX = ATEST * 0.999
      END IF

      ICAREA  =  0
      MAXITR = 1000
C-------------------------------------------------------------------------------
C--   ENTER THE FUNCTION TWICE SO THE PROGRAM CAN GET AN INITIAL
C--   PAIR OF SOLUTION VALUES
C--   START WITH THE MINIMUM VALUE FOR X
      X = -2.2
      Y1 = -1.987 + 6.732 * X -5.741 * (X**2) + 2.119 * (X**3)
      X1 = X
C--   START WITH THE MAXIMUM VALUE FOR X
      X = 2.2
      Y2 = -1.987 + 6.732 * X -5.741 * (X**2) + 2.119 * (X**3)
      X2 = X

      ITER = 0
100   CONTINUE
C--   STOP THE SEARCH IN A CIVILIZED MANNER RATHER THAN ITERATING FOREVER
      IF(ITER.GT.MAXITR) THEN
      WRITE(*,*) 'MAXIMUM ITERATIONS HIT'
      WRITE(*,5) ITER,X1,Y1,X2,Y2,ATEST
      WRITE(11,*) 'MAXIMUM ITERATIONS HIT'
      WRITE(11,5) ITER,X1,Y1,X2,Y2,ATEST
      STOP
      END IF
      ITER = ITER + 1
C--   SORT THE PAIRS OF X & Y'S SO Y2 > Y1
      IF(Y1.GT.Y2) THEN
      AX = X2
      AY = Y2
      Y2 = Y1
```

```
      X1 = AX
      Y2 = AY
      END IF

C--   CHECK FOR CONDITION A:  ATEST BETWEEN Y1 AND Y2
      IF(Y1.LT.ATEST.AND.ATEST.LT.Y2) THEN
      IF(IDEBUG.EQ.1) WRITE(*,*) 'AREA A'
C--   DETERMINE WHICH Y IS CLOSER TO ATEST
        IF((ATEST-Y1).LT.(Y2-ATEST)) THEN          ! REJECT Y2
        X1 = X1
        Y1 = Y1
        X2 = X1 + (X2-X1) * 0.50
        GO TO 200
        ELSE                                       ! REJECT Y1
        AX = X1
        X1 = X2    ! SAVE THE VALUE
        Y1 = Y2    ! SAVE THE VALUE
        X2 = X2 - (X2-AX) * 0.50
        GO TO 200
        END IF
      END IF

C--   DETERMINE IF Y IS AREA C: ATEST LESS THAN Y1
      IF(ATEST.LT.Y1) THEN
      IF(IDEBUG.EQ.1) WRITE(*,*) 'AREA C'
      ICAREA  = ICAREA + 1
      X2 = X1 - ((X1-X2)) * (0.5)
      IF((X1*0.999).LE.X2.AND.X2.LE.(X1*1.001)) THEN
      X2 = X1 * 0.85    ! CAST OUT AND TRY AGAIN
      END IF
      GO TO 200
      END IF

C--   DETERMINE IF Y IS AREA B: ATEST GREATER THAN Y2
      IF(ATEST.GT.Y2) THEN
      IF(IDEBUG.EQ.1) WRITE(*,*) 'AREA B'
      X2 = X2 + (X2-X1) * (0.5)
      IF((X1*0.999).LE.X2.AND.X2.LE.(X1*1.001)) THEN
      X2 = X1 * 1.15    ! CAST OUT AND TRY AGAIN
      END IF
        GO TO 200
      END IF
200   CONTINUE

C-------------------------------------------------------------------------
C--   FUNCTION TO EVALUATE
      Y1 = -1.987 + 6.732 * X1 -5.741 * (X1**2) + 2.119 * (X1**3)
      Y2 = -1.987 + 6.732 * X2 -5.741 * (X2**2) + 2.119 * (X2**3)
      IF(IDEBUG.EQ.1) WRITE(*,5) ITER,X1,Y1,X2,Y2,ATEST
C-------------------------------------------------------------------------
C--   TEST CRITERIA FOR THE DEGREE OF ACCURACY NEEDED IS ENTERED HERE
C--   STOP LOOP FOR A SUCCESSFUL SOLUTION IF ATEST > 0.0
      IF(Y2.GE.AMIN.AND.Y2.LE.AMAX) THEN
      XVALUE = X2
      YVALUE = Y2
      GO TO 400
      END IF

      GO TO 100
400   CONTINUE
      IF(IDEBUG.EQ.1) WRITE(*,2) ATEST,XVALUE,YVALUE,ITER
      OUT = 3.0 * 422.550 * XVALUE
      WRITE(11,6) ISET,ANO,ATEST,YVALUE,XVALUE,OUT
      WRITE(12,10) ISET,ITER,OUT
300   CONTINUE
      STOP
      END
C-------------------------------------------------------------------------
USER MUST PROVIDE THE RANG FUNCTION
C-------------------------------------------------------------------------
NO DATA ARE REQUIRED FOR SOLVE
```