UCC-ND

NUCLEAR
DIVISION

UNION
CARBIDE

RECEIVED BY TIC AUG 27 1980

COMPUTER SCIENCES DIVISION

METHODS FOR OBTAINING DISTRIBUTIONS OF URANIUM
OCCURRENCE FROM ESTIMATES OF GEOLOGIC FEATURES

C. E. Ford
R. A. McLaren

MASTER

OPERATED BY
UNION CARBIDE CORPORATION
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
Price: Printed Copy A06 ; Microfiche Copy: A01

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use or the results of such use of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

Reference to a company or product name does not imply approval or recommendation of the product by Union Carbide Corporation or the U. S. Department of Energy to the exclusion of others that may meet specifications.

K/CSD-13
Subject Category: UC-32,
Mathematics and Computers

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

METHODS FOR OBTAINING DISTRIBUTIONS OF
URANIUM OCCURRENCE FROM ESTIMATES OF GEOLOGIC FEATURES

C. E. Ford
R. A. McLaren

Sponsor: J. T. Bradbury

Date Published: April 1980

COMPUTER SCIENCES DIVISION
at
Oak Ridge Gaseous Diffusion Plant
Post Office Box P
Oak Ridge, Tennessee 37830

Union Carbide Corporation, Nuclear Division
operating the
Oak Ridge Gaseous Diffusion Plant • Oak Ridge National Laboratory
Oak Ridge Y-12 Plant • Paducah Gaseous Diffusion Plant
under Contract No. W-7405-eng-26
for the
Department of Energy

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

PAGES 2 to 4
WERE INTENTIONALLY
LEFT BLANK

CONTENTS

	PAGE
List of Tables	7
List of Figures	9
Acknowledgments	11
Abstract	13
Sections	
I INTRODUCTION	15
II THE PROBLEM	16
III SOLUTION PROCEDURE — OVERVIEW	18
IV FITTING DISTRIBUTIONS TO PERCENTILES AND CENTRAL VALUE	21
V FUNCTIONS OF RANDOM VARIABLES	26
VI FITTING DISTRIBUTIONS TO MOMENTS	29
VII THE PROBABILITY OF OCCURRENCE	32
VIII INDEPENDENCE CONSIDERATIONS	36
IX SAMPLE PROBLEM	43
X DISCUSSION OF ACCURACY	49
References and Related Publications	61
Appendices	
A The Lognormal Distribution	69
B Determining a Lognormal from Percentiles and Central Value	79
C Third and Fourth Moments of Sums and Products of Independent Random Variables	89

CONTENTS (cont'd)

	PAGE
Appendices (cont'd)	
D Approximating Pearson Distributions by Cubic Splines	93
E Effect of the Probability of Occurrence on Moments	101
F Moments of Sums of Perfectly Correlated Random Variables	105
G Computer Programs	107
Microfiche	Inside Back Cover

LIST OF TABLES

TABLE		PAGE
1.	Groups of Regions by Various Criteria in Data with 263 Regions	39
2.	Data for Sample Problem	44
3.	Analysis of Sample Data: Intermediate Results	45
4.	Percentiles of Total Uranium in Sample Regions	47
5.	Comparison of Percentiles from Simulation and Pearson Fits to Moments	53
6.	Results of Sensitivity Analysis, Single Regions	56
7.	Results of Sensitivity Analysis, Combination of Ten Regions	57
8.	Sensitivity Tests of Probability of Occurrence, P_0	59
9.	The Effect of Truncation on the Moments of Lognormal Distributions with Unit Mean and Variance	77

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

LIST OF FIGURES

FIGURE		PAGE
1.	Lognormal density with fifth percentile, mode and ninety-fifth percentile of two, three and ten respectively	25
2.	Ratio of unconditional to conditional central moments as a function of P_0	34
3.	Distribution of total uranium under different assumptions of independence	38
4.	Results of analysis of sample data: comparison of simulation and analytic methods	51
5.	Lognormal densities with mean five and unit variance . .	73
6.	Mode ratio for fifth and ninety-fifth percentiles ($v_1 = v_2 = 1.645$)	81
7.	The two lognormal densities with mode, thirtieth percentile and seventieth percentile of four, five and ten respectively	84
8.	Median ratio for fifth and ninety-fifth percentiles ($v_1 = v_2 = 1.645$)	86
9.	Mean ratio as a function of sigma	88
10.	Examples of Pearson density functions with zero means and unit variances	100

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

ACKNOWLEDGMENTS

The work discussed in this report was performed for the Grand Junction Office of the Department of Energy. Many individuals in that organization contributed to these developments; in particular, we wish to credit Messrs. David Grundy and David Blanchfield. Mr. Grundy, now with the U. S. Geological Survey, was a key participant in the formulative phase of the study and helped to develop the statistical representation of geological phenomena. Mr. Blanchfield monitored the effort as the techniques matured, continually steering our mathematical efforts into practical directions. His interface with management personnel at Grand Junction was invaluable. We wish also to thank the Department of Energy and Bendix Corporation geologists in the several field offices for their suggestions and cooperation.

We appreciate too our colleagues in the Nuclear Division who have supported and contributed to the effort. Our special thanks go to Mr. J. T. Bradbury, Operations Analysis and Planning Division, who provided our funds. We are indebted also to Vic Kane, John Beauchamp, and Marie Cuthbert for their careful reading and criticism of the draft. Finally, to our patient typist, Pat Parkey: Well done.

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

ABSTRACT

The problem addressed in this paper is the determination of a quantitative estimate of a resource from estimates of fundamental variables which describe the resource. Due to uncertainty about the estimates, these basic variables are stochastic. The evaluation of random equations involving these variables is the core of the analysis process.

The basic variables are originally described in terms of a low and a high percentile (the 5th and 95th, for example) and a central value (the mode, mean or median). The variable thus described is then generally assumed to be represented by a three-parameter lognormal distribution.

Expressions involving these variables are evaluated by computing the first four central moments of the random functions (which are usually products and sums of variables). Stochastic independence is discussed. From the final set of moments a Pearson distribution is obtained; the high values of skewness and kurtosis resulting from uranium data requires obtaining Pearson curves beyond those described in published tables. A cubic spline solution to the Pearson differential equation accomplishes this task.

A sample problem is used to illustrate the application of the process; sensitivity to the estimated values of the basic variables is discussed. Appendices contain details of the methods and descriptions of computer programs.

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

SECTION I

INTRODUCTION

The Grand Junction Office of the Department of Energy has the responsibility for estimating regional and national uranium resources. Many regions are only partially explored, resulting in uncertainty about the uranium content. This uncertainty is incorporated in the resource estimates by expressing quantities as random variables. Geologists familiar with the regions are asked to make several quantitative estimates which permit distribution functions of the random variables to be formed. Typically, these estimates are not of the uranium endowment itself, but of more basic variables such as the size of ore bodies and the grade of ore. The distribution of the resource is a function of these basic variables.

The representation of resource estimates as random variables is becoming increasingly attractive. Strategies about the use of the resource can be formulated which recognize the possibility of extreme outcomes. The resulting decisions can differ radically from those based only on average values.

In this report, the quantitative methods developed to support this resource estimating procedure are described.

SECTION II

THE PROBLEM

Uranium resource estimation usually begins with a geologist's description of variables which contribute to the resource endowment in a particular region. These individual variables are then combined to yield a variable which describes the estimate of uranium in the region; results for several regions are then summed to obtain the estimate throughout all regions. For example, a geologist may, for each region, describe the following:

f, the fraction of the area of the region thought to be underlain by deposits;

t, the tons of ore per square mile in parts of the region where deposits occur;

g, the average grade of ore (in percent) in the region; and

A, the area of the region.

The variable, u, describing the uranium in the region is then given by

$$u = \frac{A \cdot f \cdot t \cdot g}{100} \text{ tons} \quad (2.1)$$

Additionally, the geologist may wish to allow for the possibility that there is no resource at all in the region. This is accomplished by specifying a probability of uranium occurrence, P_o ; that is, P_o is the geologist's estimate of the probability that the region contains uranium ore as described by the variables above and $(1 - P_o)$ is the probability that the region contains no ore (in interesting quantities). The

variable u is thus a conditional estimation of the quantity of uranium in the region.

The problem is further complicated by the geologist's uncertainty about the estimates of fraction, tonnage, and grade (f , t , and g). This leads to the use of random variables to describe these quantities.

The geologist may express uncertainty by giving three estimates:

the most likely value of the variable,

the fifth percentile (that is, there is only one chance in twenty that the variable will be less than this quantity), and

the ninety-fifth percentile (there is a one-in-twenty chance of exceeding this quantity).

The first problem is to find distribution functions for the random variables described in this manner. We assume the densities to be unimodal, continuous functions; the most likely value is interpreted to be the mode. (Other quantities may be substituted for the above: the central measure may be the mean or median and any pair of percentiles may be substituted for the fifth and ninety-fifth.)

Next, functions of these random variables must be formed as shown in expression 2.1. The result is a conditional random variable representing the tons of mineral given the existence of any mineralization in the region. From this and the probability of occurrence, the unconditional distribution of mineral is found. Finally, the variables for all regions must be summed to yield the total mineral estimate.

Throughout all of these processes, questions of independence (or lack thereof) among random variables must be considered. For large numbers of regions computational efficiency is of concern. Finally, analysis methods must be capable of responding to changes in the form of the estimating equation 2.1; other variables and other functions may be used as the estimating methods continue to evolve.

SECTION III

SOLUTION PROCEDURE — OVERVIEW

The first problem, that of finding a distribution to represent the three values estimated by the geologist, is conceptually straightforward. By using a (three-parameter) lognormal distribution,* positively or negatively skewed, or a normal distribution, a suitable representation can be obtained for most estimates of percentiles and central value. If a lognormal distribution is to be chosen, three nonlinear equations must be solved for the three unknown parameters. Sometimes there is more than one such distribution, especially when the central quantity is the mean; in that case, the selection of the desired solution may be somewhat complicated.

Lognormal distributions have an infinite range in one direction. This is usually of no consequence, even when representing random variables of finite range. The probabilities in the tail of the distribution are sufficiently small to have little effect on the moments. In highly skewed distributions, though, this may not be true. In these cases, the distribution function is truncated at some appropriate point. This and other problems of lognormal fits to estimates are discussed in the next section.

After distributions of the basic random variables are obtained, the distribution of various functions of these variables must be found. Analytic solutions are available for only a relatively few special cases. For example, products of two-parameter lognormal distributions are also

*Fundamentals of the lognormal distribution are given in Appendix A.

lognormally distributed. There is not an equivalent result for the three-parameter case, however. Further, the distribution of the sum of the regional resources cannot be expressed usefully. For reasons such as these, functions of random variables have often been evaluated by simulation methods.

There are advantages to simulation: it is both flexible and easily understood. But it becomes costly when large numbers of random variables are involved. There is a further problem of determining and controlling the accuracy of the simulation results.

The resource estimation problem currently being addressed is of sufficient size to encourage the development of an alternative to simulation. The method chosen takes advantage of the moments of the random variables. (A brief discussion of moments is given in Appendix A.) In general, a distribution function is uniquely determined by its moments. That is, the function can be described to any degree of accuracy by knowing a sufficient number of its moments.* In practice, four moments (the mean, variance, and the third and fourth moments about the mean) are found to describe adequately most distributions.

The approach to evaluating a function of several random variables is first to obtain the moments of each of the basic variables. From these, four moments of the resulting random variable are computed (details

*There is a strict technical sense in which the lognormal distribution is not uniquely defined by its moments, but this causes no practical problem in the analysis, especially after the truncation (references: Heyde [2] and Wilks [8].)

are given in Section V). To obtain probability information, a distribution is found which has moments similar to those computed.* This is discussed in Section VI.

*Depending upon the moments, several distributions are candidates; these include the lognormal and normal, Student's t or other Pearson distributions.

SECTION IV

FITTING DISTRIBUTIONS TO PERCENTILES AND CENTRAL VALUE

The tasks discussed here have a common goal: the determination of a distribution function to represent a random variable x given only two percentiles, x_1 and x_2 , and a central value, m (either the mode, median or mean of x). Specifically, a lognormal distribution is chosen (with the normal distribution as a degenerate, i.e. symmetric, case).

It is important to note that after a lognormal is selected, it is truncated (restricted to a finite interval). The truncation is absolutely essential in order that the higher moments of the distribution not be disproportionately affected by the infinite tail (Table 9 in Appendix A shows the dramatic effects of truncation on the skewness and kurtosis, for instance).

Identifying the appropriate distribution is accomplished by analyzing the ratio,

$$r = (m - x_1)/(x_2 - x_1) \quad . \quad (4.1)$$

In the discussion that follows let v_2 be the percentile of the standard normal corresponding to x_2 ; and, v_1 , the absolute value of the percentile, corresponding to x_1 . It is assumed that the median of x lies strictly between x_1 and x_2 ; thus, $v_1 > 0$ and $v_2 > 0$. For example, if x_1 and x_2 are the fifth and ninety-fifth percentiles of x , $v_1 = v_2 = 1.645$.

Normal Distribution

If $r = v_1/(v_1 + v_2)$, then the data can be represented by a normal distribution with mean and standard deviation

$$\mu = m \quad , \text{ and} \quad (4.2)$$

$$\sigma = (x_2 - m)/v_2 \quad . \quad (4.3)$$

Note that for the normal distribution, the median, mode and mean are identical.

Lognormal from Percentiles and Mode

When m is the mode of x , the system of equations to be solved for the parameters μ , σ and d (see Appendix A) is

$$d + \exp(\mu - v_1\sigma) = x_1 \quad , \quad (4.4)$$

$$d + \exp(\mu + v_2\sigma) = x_2 \quad , \text{ and} \quad (4.5)$$

$$d + \exp(\mu - \sigma^2) = m \quad . \quad (4.6)$$

No solution exists unless $r < v_1/(v_1 + v_2)$; in that case there is a unique solution as long as $x_1 < m$. The situation $m < x_1$ requires careful analysis, especially for low confidence levels of x_1 : there is no solution unless m is sufficiently near x_1 ; even then there are generally two solutions. The solution with the lower skewness (smaller σ value) is chosen as the more reasonable distribution, although it is usually quite highly skewed, too. Appendix B contains further details.

Lognormal from Percentiles and Median

When m is the median of x , the equations are 4.4, 4.5 and

$$d + \exp(\mu) = m \quad . \quad (4.7)$$

There is a unique solution provided $x_1 < m < x_2$ and $r < v_1/(v_1 + v_2)$; if either condition fails to hold, there is no solution. See Appendix B for details.

Lognormal from Percentiles and Mean

When m is the mean of x , the system of equations consists of 4.4, 4.5 and

$$d + \exp(\mu + \sigma^2/2) = m \quad . \quad (4.8)$$

Two cases may be identified: in the first case, $v_1 v_2 > 1$. Then for $r < v_1/(v_1 + v_2)$, there are usually two solutions; the solution possessing the smaller σ value (and, hence, smaller skewness) is selected. If $r > v_1/(v_1 + v_2)$, there is a unique solution; however, the skewness is often alarmingly high, particularly if v_2 is large—hence, this condition should generally be avoided. In the second case, $v_1 v_2 \leq 1$; if $r > v_1/(v_1 + v_2)$, there is a unique solution. Otherwise, no solution exists. Refer to Appendix B for more detail.

Negative Lognormal from Percentiles and Central Value

When the corresponding lognormal algorithm above has failed to find a satisfactory solution, in most cases it is possible to find a negative lognormal distribution to represent the data. To accomplish the transformation, set

$$v'_1 = -v_2, \quad v'_2 = -v_1, \quad x'_1 = -x_2, \quad m' = -m, \quad \text{and} \quad x'_2 = -x_1 \quad .$$

If μ' , σ' and d' are the parameters of the lognormal representing the "prime" data, then the parameters of the negative lognormal for v_1 , v_2 , x_1 , m and x_2 are

$$\mu = \mu' , \sigma = \sigma' , \text{ and } d = -d' \quad (\text{see Appendix A}) .$$

Two-Parameter Lognormal by Least-Squares

A two-parameter lognormal distribution can be obtained which best represents the data in the sense of least squares. First, the parameter μ may be expressed in terms of the parameter σ by requiring that the fitted central measure (mode, mean or median) equal m . Then the value of σ is chosen to minimize the sum of squares of the difference between the fitted percentiles and the input, x_1 and x_2 .

Example: Find a Lognormal Distribution from Estimates

Given the estimates

$$\text{5th percentile: } x_1 = 2 ,$$

$$\text{mode: } m = 3 , \text{ and}$$

$$\text{95th percentile: } x_2 = 10 ; \text{ and}$$

solve the system of equations:

$$d + \exp (\mu - 1.645\sigma) = x_1 ,$$

$$d + \exp (\mu + 1.645\sigma) = x_2 , \text{ and}$$

$$d + \exp (\mu - \sigma^2) = m .$$

The ratio is $r = (m - x_1)/(x_2 - x_1) = .125$, from which the parameters are

$$\sigma = .6404 \quad ,$$

$$\mu = 1.1557 \quad , \quad \text{and}$$

$$d = .8924 \quad .$$

The density function, $f(x)$, is shown in Figure 1, below.

Figure 1. Lognormal density with fifth percentile, mode and ninety-fifth percentile of two, three and ten respectively.

SECTION V

FUNCTIONS OF RANDOM VARIABLES

In Section III the use of moments in the evaluation of functions of random variables was introduced. In this section the procedure is developed in detail.

Addition. Let x and y be stochastically independent random variables with means, variances, and third and fourth moments about the mean:

$$m_x = E(x), \text{ the mean of } x$$

$$v_x = E(x - m_x)^2, \text{ the variance of } x$$

$$t_x = E(x - m_x)^3, \text{ the third moment about the mean}$$

$$k_x = E(x - m_x)^4, \text{ the fourth moment about the mean.}^*$$

Similarly, m_y , v_y , t_y and k_y are defined. Now form the random variable z :

$$z = x + y.$$

It is well known that the mean and variance of z are given by

$$m_z = m_x + m_y \tag{5.1}$$

$$v_z = v_x + v_y. \tag{5.2}$$

*For conciseness, these last two quantities will be referred to simply as the third and fourth moments.

It is shown in Appendix C that the third and fourth moments of z are

$$t_z = t_x + t_y \quad (5.3)$$

$$k_z = k_x + 6v_x v_y + k_y \quad (5.4)$$

Clearly, associativity permits these processes to be extended to any number of variables.

Multiplication. Let x and y be independent random variables with moments m , v , t and k as above. Form the random variable z :

$$z = xy \quad .$$

The mean of z is well known:

$$m_z = m_x m_y \quad (5.5)$$

The variance of z is also available. See, for example, Kendall and Stuart [5], page 245:

$$v_z = v_x v_y + m_x^2 v_y + m_y^2 v_x \quad (5.6)$$

The third and fourth moments of z , developed in Appendix C, are

$$t_z = t_x t_y + t_x m_y (3v_y + m_y^2) + t_y m_x (3v_x + m_x^2) + 6m_x m_y v_x v_y \quad (5.7)$$

and

$$k_z = [k_x + 4m_x t_x + 6m_x^2 v_x + m_x^4] [k_y + 4m_y t_y + 6m_y^2 v_y + m_y^4] - 4m_z t_z - 6m_z^2 v_z - m_z^4 \quad (5.8)$$

Again, associativity allows products of any number of mutually stochastically independent random variables to be formed.

Division. Let x be a random variable having a two-parameter lognormal distribution, $\Lambda(\mu, \sigma^2)$. Then $z = 1/x$ is $\Lambda(-\mu, \sigma^2)$ (see Appendix A). Division can be accomplished by representing divisors as two-parameter lognormal random variables, inverting, and then using the multiplication techniques outlined above.

SECTION VI

FITTING DISTRIBUTIONS TO MOMENTS

Previous sections have shown how to evaluate the first four moments of certain random variables (for instance, the uranium endowment in a single study area or the total endowment in several areas). Usually the Pearson system of distributions is employed to convert these moments into a distribution function; occasionally the lognormal may be used to represent the resulting distribution function by matching the first three moments only. Procedures for obtaining the appropriate Pearson or lognormal distribution are described below.

Pearson Distributions

The Pearson system is a class of distribution functions governed by a differential equation. Let $f(x)$ be the density function for a member of this class, normalized to have zero mean and unit variance. Then there exist constants c_0 , c_1 and c_2 (which can be uniquely determined from the third and fourth moments) so that

$$df/dx = -(c_1 + x) f(x)/(c_0 + c_1x + c_2x^2) \quad (6.1)$$

This can easily be rewritten as

$$d[\ln f(x)]/dx = -(c_1 + x)/(c_0 + c_1x + c_2x^2) \quad ; \quad (6.2)$$

thus, by analyzing the roots of the quadratic in the denominator of the right-hand side, the solution may be obtained analytically except for a normalization constant so that the integral of the density is one.

Special cases include the normal, Student's t , gamma and beta distributions, among others not so familiar. In general, however, the distribution functions themselves cannot be expressed in closed-form. Also, when the specified moments are near the "boundary" between two types of Pearson distributions, numerical difficulties arise.

For the above reasons, another approach to the solution of equation 6.1 is employed. Let $F(x)$ represent the distribution function, i.e., $F'(x) = f(x)$. Then equation 6.1 may be written as a second-order linear differential equation in $F(x)$, namely

$$(c_0 + c_1x + c_2x^2) F''(x) + (c_1 + x) F'(x) = 0 \quad (6.3)$$

A sequence of nodes, $x_1, x_2, \dots, x_n$, is selected, based on the Pearson type (determined from c_0, c_1 , and c_2). Then a cubic-spline approximation to $F(x)$ is obtained which satisfies equation 6.3 at each of the nodes; $F(x_1) = 0$; and $F(x_n) = 1$. The equations for defining and manipulating the spline are given in Appendix D.

Cubic spline approximations to Pearson distributions have shown very close agreement to existing tables of Pearson percentiles. In addition, the calculations are usually insensitive to boundaries between the various types. References for Pearson distributions and tables of percentiles include [3], [4], [5], and [7].

The Lognormal Distribution

Given the mean, variance and third moment, m , v and t , of a lognormal distribution, the parameters μ , σ and d are easily obtained. First, the equation for the skewness (see equation A.13 in Appendix A),

$$G(\eta) = \eta^3 + 3\eta - \frac{t}{v^{3/2}} = 0 \quad (6.4)$$

is solved for η . This is straightforward, for instance, by Newton's method, since

$$G'(\eta) = 3(\eta^2 + 1) > 0 \quad . \quad (6.5)$$

The parameter σ is then computed from equation A.8, the definition of η ; μ and d follow rapidly from equations A.9 and A.10. See the description of subroutines LN4MOM and LN3MOM in Section XI for the actual computer algorithms.

SECTION VII

THE PROBABILITY OF OCCURRENCE

It was mentioned in Section II that geologists make estimates of resource occurrence for a region conditioned upon some significant quantity of uranium existing in the region. Accompanying this conditional estimate is a probability, P_o , representing the geologist's degree of belief that some uranium does indeed exist. The conditional moments are found for the region from the distributions fit to the various factors describing the quantity and grade of ore. In this section, the modification required to convert these to unconditional moments is discussed.

Finding the Unconditional Moments

Consider a region for which the conditional distribution of uranium has been described by a random variable, y , with density function $f_y(y)$ for $y > 0$. Let the mean, variance, and third and fourth moments be m_y , v_y , t_y and k_y , respectively. Further, let the probability of a deposit be P_o . Let x be the random variable describing the unconditional distribution of uranium. The density function of x , $f_x(x)$, is given by

$$f_x(x) = P_o f_y(x) \text{ when } x > 0 ; \text{ and}$$

$$\Pr(x=0) = 1 - P_o . \quad (7.1)$$

It is required to find the moments of the variable x in terms of P_o and the moments of y . Derivations are given in Appendix E.

The moments of x are

$$m_x = P_o m_y, \quad (7.2)$$

$$v_x = P_o [v_y + (1 - P_o) m_y^2], \quad (7.3)$$

$$t_x = P_o [t_y + 3(1 - P_o) m_y v_y + (1 - P_o) (1 - 2P_o) m_y^3], \quad (7.4)$$

$$k_x = P_o \left\{ k_y + 4(1 - P_o) m_y t_y + 6(1 - P_o)^2 m_y^2 v_y + (1 - P_o) [1 - 3P_o(1 - P_o)] m_y^4 \right\}. \quad (7.5)$$

The effect of P_o on the moments can be dramatic. Figure 2 illustrates an example of these changes. Note that the third moment changes sign for some values of P_o . The sensitivity of the unconditional moments to the value of P_o clearly indicates the importance of careful estimation of this probability. Effects of errors in the estimation of P_o are examined in Section X.

Probability of Occurrence as a Random Variable

As with all quantities estimated by the geologists, there is uncertainty about the appropriate value of P_o . This suggests expressing P_o as a random variable. However, subject to one key assumption, it is shown here that the results are not changed by doing so.

Again, consider a random variable, y , which represents the conditional uranium content in a region. Suppose now that the probability of occurrence is expressed as a discrete random variable with values

(a) Conditional moments: $m_y = 1$, $v_y = 0.1$, $t_y = .014$
(skew = .2), $k_y = .03$ (kurtosis = 3).

(b) Conditional moments: $m_y = 1$, $v_y = .1$, $t_y = .044$
(skew = 2), $k_y = .05y$ (kurtosis = 5).

Figure 2. Ratio of unconditional to conditional central moments as a function of P_0 .

$P_1, P_2, \dots, P_n$, each with respective probabilities $\pi_1, \pi_2, \dots, \pi_n$.

The mean probability of occurrence will be expressed as P_o . Thus

$$P_o = \sum_{i=1}^n \pi_i P_i$$

The key assumption is that the conditional distribution of y , $F_y(y)$, remains the same for all realizations of the probability of occurrence, P_i . The unconditional distribution of uranium, $F_x(x)$ is, then,

$$\begin{aligned} F_x(x) &= \Pr(X \leq x) = \Pr(X \leq x | P_1) \pi_1 + \dots + \Pr(X \leq x | P_n) \pi_n \\ &= \sum_{i=1}^n \pi_i \left\{ (1 - P_i) + P_i F_y(x) \right\} \\ &= \sum_{i=1}^n \left\{ \pi_i - \pi_i P_i + P_i \pi_i F_y(x) \right\} \\ &= (1 - P_o) + P_o F_y(x) \end{aligned} \quad (7.6)$$

This shows that the expected value of the probability of occurrence is the only quantity that affects the unconditional distribution. It should be noted that expressions 7.1 and 7.6 are equivalent. Expressions 7.2 through 7.5 can thus be used with the expected value of the probability of occurrence replacing the point estimate.

For simplicity, the above demonstration has used a discrete distribution for P_o . An equivalent result obtains for the continuous case.

SECTION VIII

INDEPENDENCE CONSIDERATIONS

Total uranium, u , is found from n regional occurrences, u_i , $i = 1, \dots, n$, by

$$u = \sum_{i=1}^n u_i .$$

In Section V the methods of computing the moments of u from those of the u_i were discussed assuming the u_i variables are mutually stochastically independent. This independence assumption is not always justified, however. Regions may have been divided by nongeologic criteria, such as ownership or political jurisdictions. Clearly in such cases, independence ceases to be a valid assumption.

Assessing the nature of dependence among regions is not straightforward. There are obviously degrees of geologic similarity and geographic compactness. There is, however, no information available to quantify these relationships. A typical assumption is that neighboring regions of identical geology are perfectly positively correlated; otherwise they are assumed to be independent.

Computation of moments for the sum of perfectly correlated random variables is thus required. Let x and y be two such variables and form

$$z = x + y . \quad (8.1)$$

The mean, variance, and third and fourth central moments of z (m_z , v_z , t_z and k_z , respectively) are given by

$$m_z = m_x + m_y \quad (8.2)$$

$$v_z = \left(v_x^{1/2} + v_y^{1/2} \right)^2 \quad (8.3)$$

$$t_z = \left(t_x^{1/3} + t_y^{1/3} \right)^3 \quad (8.4)$$

$$k_z = \left(k_x^{1/4} + k_y^{1/4} \right)^4 \quad (8.5)$$

Details are found in Appendix F.

An example of how these assumptions affect the analysis is now considered. Data from 263 regions was processed using three different criteria for ascribing perfect correlation among regions: 1) geography and age of rock, 2) geography, age of rock, and geologic formation, and 3) age and formation only. The number of groups and the number of regions in each group are shown in Table 1 for each criteria. The random variables within each group are then treated as being perfectly correlated and are added accordingly. The resultant variables for the groups are then treated as independent.

Figure 3 shows the distribution of total uranium under these assumptions. (The curve for grouping by age and formation is virtually identical to that for region, age, and formation.) For contrast, the results are displayed for the cases in which all regions are assumed to be independent and in which all regions are perfectly correlated.

Figure 3. Distribution of total uranium under different assumptions of independence.

Table 1

Groups of Regions by Various Criteria in
Data with 263 Regions

Number of Regions in the Group	Number of Groups of This Size When Regions are Grouped by		
	Region and Age of Rock	Region, Age, and Formation	Age and Formation
1	30	64	55
2	11	25	23
3	9	10	9
4	4		2
5	2		
7	3	5	4
8	1		
9	1	1	
10	1		1
11			1
12	1		
16	1	2	1
19			1
20	1		
21		1	1
22		1	1
24	1		
38	1		
Total Number of Independent Groups	67	109	99

Comments on the Linear Assumption of Correlation

The distributions separately developed for each of the correlated regions may be inconsistent with the assumption of perfect correlation. The assumption implies the relation between variables x and y

$$y = ax + b \quad .$$

For $n > 1$, the n^{th} central moment of y , $\lambda_n(y)$, is

$$\lambda_n(y) = E(y - m_y)^n = E[ax + b - am_x - b]^n = a^n \lambda_n(x) \quad .$$

The coefficient a is thus uniquely determined by any pair of higher moments of x and y . The three pairs of higher moments considered in this analysis imply (in general) three different relationships between x and y .

In order to understand the effect of expressions 8.3 through 8.5 relative to this inconsistency, consider the coefficient a to be known and define the following:

$\lambda_n(x)$, the n^{th} central moment of x .

$\lambda_n(y)$, the n^{th} central moment of y .

$\lambda_n(z)$, the n^{th} central moment of $z = x + y$ using
the given moments of x and y .

$\hat{\lambda}_n(y)$, the moment of y computed assuming the moment
of x is correct.

$\hat{\lambda}_n(z;x)$, the n^{th} central moment of z assuming $\lambda_n(x)$ is correct and computing the moment of y from the moment of x and the coefficient a . Throughout, the caret will indicate values computed assuming one or the other of the moments to be correct.

$\hat{\lambda}_n(x)$, $\hat{\lambda}_n(z;y)$ values similar to the two preceding but assuming the moment of y to be correct.

From the above definitions, the following expressions can be written

$$\lambda_n(z) = [\lambda_n(x)^{\frac{1}{n}} + \lambda_n(y)^{\frac{1}{n}}]^n \quad (8.6)$$

$$\hat{\lambda}_n(x) = a^{-n} \lambda_n(y) \quad (8.7)$$

$$\hat{\lambda}_n(y) = a^n \lambda_n(x) \quad (8.8)$$

$$\hat{\lambda}_n(z;x) = [\lambda_n(x)^{\frac{1}{n}} + \hat{\lambda}_n(y)^{\frac{1}{n}}]^n \quad (8.9)$$

$$\hat{\lambda}_n(z;y) = [\hat{\lambda}_n(x)^{\frac{1}{n}} + \lambda_n(y)^{\frac{1}{n}}]^n \quad (8.10)$$

Inconsistency between the moments of x and y can take two forms:

$$\text{Case 1: } \lambda_n(x) > \hat{\lambda}_n(x) \quad (8.11)$$

From expressions 8.11, 8.6, and 8.10 it is easy to see that

$$\lambda_n(z) > \hat{\lambda}_n(z;y) \quad (8.12)$$

But expressions 8.11 and 8.7 also imply that

$$\lambda_n(x) > \frac{1}{a} \lambda_n(y)$$

$$\text{or } a \lambda_n(x) > \lambda_n(y)$$

$$\text{and } \hat{\lambda}_n(y) > \lambda_n(y) \quad . \quad (8.13)$$

Now from 8.13, 8.6, and 8.9,

$$\hat{\lambda}_n(z;x) > \lambda_n(z) \quad . \quad (8.14)$$

From 8.12 and 8.14

$$\hat{\lambda}_n(z;y) < \lambda_n(z) < \hat{\lambda}_n(z;x) \quad . \quad (8.15)$$

$$\text{Case 2: } \lambda_n(x) < \hat{\lambda}_n(x)$$

A development similar to that in Case 1 leads to

$$\hat{\lambda}_n(z;x) < \lambda_n(z) < \hat{\lambda}_n(z;y) \quad . \quad (8.16)$$

It is seen from 8.15 and 8.16 that the answer obtained by using the given moments of x and y is always bracketed by those which would be obtained by assuming either the moment of x or the moment of y to be correct.

SECTION IX

SAMPLE PROBLEM

In this section the analysis methods previously discussed will be used to analyze data from 20 regions. The particular subset of data chosen avoids extreme values of skewness and the mean tons of uranium in each region are all of comparable magnitude. This was done since, with a small number of regions, one extreme distribution would dominate the results.

The data provided by the geologists for each region are tabulated in Table 2. The steps in the analysis, in summary, are listed below. Subroutines (described in Appendix G) applicable to each step are referenced.

1. For each region, a three-parameter lognormal* distribution is fit to the data for each of the three variables fraction (f), tonnage (t), and average grade (g) using methods discussed in Section IV. (Ref. subroutines L4MODE and L3MODE.)
2. The first four central moments of f_i , t_i and g_i are computed for each region, $i=1, \dots, 20$. The moments of the tons of uranium, u_i , with

$$u_i = \frac{A_i f_i t_i g_i}{100} \quad (\text{where } A_i \text{ is the area of the region})$$

*A normal or negatively skewed lognormal distribution may be used if applicable.

Table 2
Data for Sample Problem

Region	Area (sq. mi.)	Prob. of Occurrence	Fraction $\times 10^{-3}$			Tonnage/Sq. Mi. $\times 10^5$			Average Grade (%) $\times 10^{-3}$		
			5th	mode	95th	5th	mode	95th	5th	mode	95th
1	1	1.00	4.1	16.8	29.8	121	220	341	40	80	140
2	220	1.00	27.0	55.0	103.0	.489	.626	.799	20	40	75
3	42	.22	0.78	1.0	4.8	15	78.8	160	20	40	80
4	138	1.00	1.0	3.0	6.0	10	25	50	50	60	80
5	606	1.00	8.0	18.0	36.0	.3	.8	1.5	50	60	80
6	89	.90	.047	.152	.408	101	164	320	46	106	184
7	99	1.00	3.44	6.84	16.7	7	9.5	20	26	57	110
8	54	1.00	5.2	9.26	14.2	7.5	11.6	20	22	56	101
9	62	1.00	2.44	9.56	26.3	6.1	10.9	20	33	64	142
10	38	.75	1.84	9.64	32.1	4.6	9.0	23	33	52	98
11	92	.70	.64	3.28	34.8	6.7	8.9	25	26	51	81
12	32	1.00	3.2	9.12	29.1	6.7	9.3	22	34	66	137
13	52	1.00	3.75	9.66	34.1	5	8	15	22	50	98
14	35	1.00	3.34	12.4	44.8	5.2	9.4	20	31	54	110
15	19	1.00	6.24	12.8	34.8	9	13.6	20	32	60	148
16	50	1.00	3.08	9.68	27.6	6.9	11.8	24	23	64	112
17	57	.70	0.98	5.8	24.2	8.4	12.6	24	25	58	102
18	43	1.00	2.9	9.58	34.8	7.8	11.7	25	29	58	92
19	49	1.00	2.66	9.34	29.7	7.4	11.8	20	33	62	105
20	63	1.00	2.22	9.02	27.8	6.6	10.3	23	33	62	105

Table 3

Analysis of Sample Data: Intermediate Results

For each region, the moments of the unconditional uranium occurrence give the following:

Region	Mean Tons	Std. Dev.	Coeff. of Variation	Skewness	Kurtosis
1	3312	2188	0.66	1.29	5.50
2	3788	2137	0.56	1.22	5.29
3	7109	2061	0.29	5.67	56.60
4	8060	5327	0.66	1.30	5.55
5	6695	4179	0.62	1.25	5.39
6	3386	3060	0.90	1.81	8.57
7	6525	4827	0.74	2.08	10.63
8	4046	2302	0.57	1.25	5.57
9	7413	6273	0.85	2.12	10.59
10	2900	3539	1.22	2.46	13.45
11	4971	8009	1.61	3.97	31.29
12	4050	3574	0.88	2.54	14.58
13	4043	3528	0.87	2.32	12.08
14	4925	4521	0.92	2.46	13.44
15	3623	2752	0.76	2.10	10.42
16	6067	5126	0.84	2.06	10.03
17	3618	4631	1.28	2.34	11.84
18	5594	4810	0.86	2.23	11.43
19	5660	4465	0.79	1.87	8.70
20	6998	5857	0.84	2.86	10.39

are found using methods in Section V. The random variable represented by these moments is a conditional distribution of uranium, the condition being that a deposit indeed exists in the region. (Ref. subroutines QNTLN4, QNTLN3, TRNCL4, TRNCLN, PRD4M.)

3. If the geologist has allowed for the possibility that no deposits occur (that is, the probability of occurrence is less than one), the moments are modified as described in Section VII. These, then, are the moments of the unconditional quantity of uranium estimated to be in each region. These values are reported in Table 3. (Ref. subroutine UNCMOM.)
4. The total uranium in the sample regions is then found, starting with the proper summation of moments as discussed in Section V. All regions in this sample are assumed to be independent. The moments of the random variable representing total uranium are:

Mean:	96384 tons
Variance:	3.9×10^8
Third moment:	5.17×10^{12}
Fourth moment:	6.31×10^{17}
Skew:	.667
Kurtosis:	4.11

(Ref. subroutine SUM4MI.)

5. A Pearson distribution which fits these moments is found as discussed in Section VI. For comparison, a three-parameter lognormal distribution is also fit. Both results are shown in Table 4. (Ref. subroutines LN4MOM, LN3MOM, INPRSN, GNODES, GCPRSN and QLPRSN.)

Table 4

Percentiles of Total Uranium in Sample Regions

P	<u>Pth Percentile of Uranium (10³ tons)</u>	
	Pearson Fit to Moments	Lognormal Fit to Moments
5	67.6	67.8
10	73.0	72.9
20	79.9	79.6
30	85.2	84.8
40	89.9	89.6
50	94.5	94.3
60	99.3	99.3
70	104.8	104.9
80	111.6	111.9
90	122.1	122.5
95	131.6	132.1

In this problem, the need for Pearson distributions instead of the lognormal for representing the final result is not illustrated; the moments are such that the lognormal gives a good fit. In other cases, particularly when skewness or kurtosis is high, this is not the case.

This same problem will be used in the next section to compare the method of moments to simulation.

SECTION X

DISCUSSION OF ACCURACY

There are two major topics to be included in this section. First, the use of moments to analyze functions of random variables is unfamiliar to many users. This results in a lingering skepticism about the results obtained. There have been requests for comparisons of distributions obtained by moments with those which result from the better-known method, simulation. An example of such a comparison is presented.

Secondly, the numbers used throughout these analyses are estimates, founded in geologic theory, but still somewhat subjective. It is desirable that the final results be reasonably insensitive to small variations in these numbers. A feature of the analysis that makes this an interesting issue is the fact that products of random variables with high skewness are involved; the potential for serious consequences cannot be dismissed lightly. Several case studies of the sensitivity of results to input value variations are presented. The conclusion is that, over a wide range of conditions, input perturbations are not amplified (and are generally attenuated quickly) as the analysis proceeds.

Simulation Comparison

The sample data analyzed by moments in Section IX was, for comparison, also analyzed by simulation. The simulation proceeded as follows:

1. Three parameter lognormal distributions were fit to the fraction (f_i), tonnage (t_i), and grade (g_i) for each region, $i=1, \dots, 20$, of the sample data.
2. A random observation was taken from each factor of each region.

That is, a sample of f_i , t_i , and g_i was obtained for each of

the 20 regions. An observation of total uranium was obtained by

$$u = \frac{1}{100} \sum_{i=1}^{20} A_i f_i t_i g_i .$$

Five hundred such observations then formed a sample of the distribution of u .

3. Twenty samples were obtained in this way. From these the mean value of each percentile and the variance of that value were found.

Figure 4 shows a plot of the results. The solid line is the Pearson distribution computed in the last section. The points are the simulation results; the spread around each point is the 95% confidence interval. It is apparent that the method of moments and simulation give, in this case, highly comparable results. The only variations of any significance occur in the mid-range of percentiles, a region typically of little interest. However, it is important to note that each of the Pearson percentiles lies within the 95% confidence interval determined by simulation.

Another observation can be made by comparing the true moments to those obtained by simulation:

	<u>Analytical Moments</u>	<u>Simulation Moments</u>
Mean	9.638×10^4	9.651×10^4
Variance	3.918×10^8	3.838×10^8
Third Moment	5.171×10^{12}	4.433×10^{12}
Fourth Moment	6.313×10^{17}	5.478×10^{17}

Figure 4. Results of analysis of sample data: comparison of simulation and analytic methods.

The imprecision of the simulation is clearly illustrated by the above differences. The moments obtained by the simulation were used in one additional manner. A Pearson distribution was fit to the simulation moments; the result was a curve which even more closely reproduced the simulation distribution. Table 5 shows a comparison of the percentiles from simulation, the Pearson fit to simulation moments, and the Pearson fit to the real moments. These results tend to confirm the aptitude of a Pearson distribution to represent the distribution of u . A further implication is that, had the simulation process generated the correct moments, the agreement between simulation percentiles and those from the method of moments would have been improved.

A final comparison of the two methods involves the computation time. The following processing times were obtained.*

Input and edit data for 20 regions; find lognormal parameters for 60 random variables; perform calculations to get moments of the sum of 20 regions:	1.32 seconds.
Fit a Pearson distribution to the final moments:	.53 seconds.
Simulation (20 samples of 500 observations each) to obtain the distribution of the sum of 20 regions:	315.47 seconds.

Several points need to be made about these times. First, the simulation performed contained more observations than necessary for an adequate result. Also, variance reduction techniques were not explored. A good simulation could undoubtedly be performed in considerably less time. On the other hand, only 20 regions were involved; a typical real problem

*The machine used was a Digital Equipment Corporation KL-10 processor.

Table 5

Comparison of Percentiles from Simulation
and Pearson Fits to Moments

p	p th Percentile of Uranium (10 ³ Tons)		
	Pearson Fit to Correct Moments	Pearson Fit to Simulation Moments	Simulation Results
5	67.6	67.7	67.4
10	73.0	73.1	73.0
20	79.9	80.0	80.0
30	85.2	85.3	85.3
40	89.9	90.0	90.0
50	94.5	94.7	94.8
60	99.3	99.6	99.5
70	104.8	105.1	104.9
80	111.6	112.0	111.9
90	122.1	122.2	122.3
95	131.6	131.4	131.2

involves several hundred regions. While processing for either calculation of moments or simulation would increase in approximate proportion to the number of regions, the processing for the Pearson fit would remain unchanged.

Input Error Analysis

The objective now is to examine the impact on the results of changes in input quantities. Clearly there are too many combinations of situations and types of errors to permit an exhaustive study. The approach which follows involves a simplified model of the total data analysis procedure.

Let x_1 , x_2 and x_3 be stochastically independent random variables represented by identical two-parameter lognormal distributions truncated at the 99.5th percentile. Form the variable u with

$$u = x_1 \cdot x_2 \cdot x_3$$

and find the fifth and ninety-fifth percentiles, u_{05} and u_{95} , and the mean, u_m , of u . Then consider a perturbation on one of the original variables, x_1 , which would represent an error in estimating a basic piece of input data by the geologist. For example, suppose the fifth percentile of x_1 had been estimated 10% too high. Let this perturbed variable be x_1^* . Now form the result u^* with

$$u^* = x_1^* \cdot x_2 \cdot x_3$$

The percentiles and mean of u^* are compared to those of u to expose the impact of the perturbation of x_1 .

Table 6 gives the key results from a number of such tests. There are three groups of cases, representing x_i variables with low, moderate and relatively high skewness (σ equal to .1, .5 and 1.0 respectively).

From these cases several observations are made, all of which are substantiated by other evidences in our investigation. Most importantly, errors do tend to be damped out as the analysis proceeds. Changes of 10% in input variables rarely result in a 10% difference in the means or percentiles of the product (case 21 is an exception); usually the differences are much reduced. The effect of a change is roughly proportional to the perturbation; compare for example cases 13, 14 and 15 with cases 7, 9 and 11. The sign of the effect is generally predictable, cases 18 and 19 showing some modest exception.

It is further seen that changes in the fifth percentile are more significant when the distributions are less skewed and the 95th percentile is more influential for highly skewed data. (Contrast cases 1, 2 and 3 with cases 18, 20 and 22.)

The resulting random variable, u^* , represents the perturbed estimate of uranium distribution for a single area. We now wish to examine the effect of the change when this area is combined with other areas to give a total uranium endowment over some larger geographical division. Table 7 shows an example when ten areas are combined. The objective is to compare the result when all ten areas are similar to the original unperturbed random variable, u , with the case when one of the ten is u^* . Thus

$$u_t = u_1 + \dots + u_{10} \quad , \quad \text{and}$$

$$u_t^* = u^* + u_1 + \dots + u_9$$

are contrasted.

Table 6

Results of Sensitivity Analysis, Single Regions

Case	With σ of	Percentage Changes in These Input Parameters			Result in These Percentage Changes in the Final Distribution		
		x_{05}	Mode	x_{95}	u_{05}	u_m	u_{95}
1	.1	10	0	0	5.69	2.56	.74
2		0	10	0	1.54	3.74	2.82
3		0	0	10	.89	3.57	6.87
4		-10	0	0	-7.36	-2.69	-.64
5		0	-10	0	-3.09	-4.23	-2.87
6		0	0	-10	-1.36	-3.62	-6.05
7	.5	10	0	0	4.13	.92	.10
8		-10	0	0	-4.52	-.98	-.13
9		0	10	0	-1.28	3.19	2.14
10		0	-10	0	2.11	-3.49	-2.17
11		0	0	10	7.40	5.68	7.79
12		0	0	-10	-7.17	-5.77	-7.67
13		20	0	0	7.93	1.76	.16
14		0	20	0	-2.05	6.20	4.26
15		0	0	20	14.93	11.30	15.71
16		10	0	-10	-3.53	-4.77	-7.50
17	1.0	-10	0	10	2.43	4.77	7.71
18		10	0	0	.77	-.34	.91
19		-10	0	0	-.72	.31	-.77
20		0	10	0	.49	.87	-1.74
21		0	-10	0	-.54	-2.53	2.34
22		0	0	10	8.70	7.90	11.24
23		0	0	-10	-8.77	-7.96	-11.08

It is not surprising that the perturbation is diluted by this process. It is reassuring, though, that the dilution proceeds as orderly as it does. Comparing cases 7, 9 and 11 from Table 6 with the results in Table 7 shows an average dilution of the change by almost exactly a factor of 10.

Table 7

Results of Sensitivity Analysis, Combination of Ten Regions

Case	With σ of	Percentage Changes in These Input Parameters			Results in These Percentage Changes in the Final Distribution		
		x_{05}	Mode	x_{95}	u_{05}	u_m	u_{95}
24	.5	10	0	0	.17	.09	.05
25		0	10	0	.37	.32	.25
26		0	0	10	.40	.57	.70

One would intuitively expect compensating perturbations to be much attenuated in the final results. To test this conjecture the following comparison was made. Two similar regions with moderate skewness ($\sigma=.5$) were combined. Another such combination was performed, but with a change in one factor (x_1) of each area: in one area, the fifth percentile was set 10% higher and in the other it was 10% lower. The results, which confirm our expectations, follow:

	<u>Fifth Percentile</u>	<u>Mean</u>	<u>Ninety-fifth Percentile</u>
Base case	.61426	1.91710	4.4597
Perturbed case	.61346	1.91653	4.4701
% Difference	.13	.03	- .23

The final comment concerns P_o , the estimate of the probability that a deposit occurs. It was seen in Section VII that small changes in P_o can have large impacts on the moments of the unconditional random variable. Figure 2 (page 34) illustrates this effect well. It should therefore be expected that errors in estimating the probability of occurrence will have important consequences.

To quantify this, comparisons were made as follows: let the distribution functions of f , t and g be identical two-parameter lognormal functions and form the conditional random variable,

$$u = f \cdot t \cdot g \quad . \quad (\text{Unit area is assumed.})$$

Apply a probability of occurrence to u and find the moments of the unconditional random variable. Sum ten of these and find the percentiles and mean of the total. These results are then compared to a case in which one of ten regions has had a change of one percentage point in P_o . The results are shown in Table 8. Clearly the impact can be extreme, particularly for data with high skewness ($\sigma=1$).

Table 8

Sensitivity Tests of Probability of Occurrence, P_o

Data with low skewness ($\sigma=.1$ for f, t and g).				
	P_o	u_{05}	u_{mean}	u_{95}
	.90	8.820	9.851	10.823
	.91	8.864	9.864	10.798
% Difference		.51	.13	-.24
	.50	8.292	9.453	10.671
	.51	8.299	9.465	10.661
% Difference		.08	.12	-.10
	.10	8.165	9.055	10.097
	.11	8.166	9.065	10.118
% Difference		.01	.11	.21
Data with high skewness ($\sigma=1.0$ for f, t and g).				
	P_o	u_{05}	u_{mean}	u_{95}
	.90	4.047	8.414	24.230
	.91	4.387	8.511	23.730
% Difference		8.40	1.15	-2.06
	.50	3.878	8.074	23.657
	.51	4.056	8.132	23.364
% Difference		4.57	.72	-1.24
	.10	3.734	7.734	23.073
	.11	3.772	7.754	23.018
% Difference		1.01	.25	-.24

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

REFERENCES
AND
RELATED PUBLICATIONS

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

REFERENCES

1. Aitchinson, J., and Brown, J. A. C., The Lognormal Distribution with Special Reference to Its Uses in Economics, Cambridge University Press, London, 1969.
2. Heyde, C. C., "On a Property of the Lognormal Distribution," *Journal of the Royal Statistical Society*, Series B, Vol. 25, 1963, p. 392.
3. Johnson, N. L., "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika* (1949), Vol. 36, p. 149.
4. Johnson, N. L., et al., "Table of percentage points of Pearson curves, for given $\sqrt{\beta_1}$ and β_2 , expressed in standard measure," *Biometrika* (1963), Vol. 50, p. 459.
5. Kendall, M. G., and Stuart, A., The Advanced Theory of Statistics, Vol. I, Hafner Publishing Co., New York, 1969.
6. Miller, C. F., et al., "User's Guide to PDFPLOT; A Code for Plotting Certain Statistical Distribution Functions," TREE-1290, UC-32, Idaho National Engineering Laboratory, September 1978.
7. Solomon, H., and Stephens, M. A., "Approximations to Density Functions Using Pearson Curves," *Journal of the American Statistical Association*, Vol. 73, No. 361, March 1978.
8. Wilks, Samuel S., Mathematical Statistics, John Wiley and Sons, Inc., New York, 1962. (Section 5.5 "Determination of Distribution Functions from Moments")

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

RELATED PUBLICATIONS

1. Abramowitz, M., and Stegun, I. A., Eds., Handbook of Mathematical Functions, AMS 55, National Bureau of Standards, Washington, D.C., 1968. (26.2 "Normal or Gaussian Probability Function")
2. Cox, N. D., and Miller, C. F., "User's Description of Second-Order Error Propagation (SOERP) Computer Code for Statistically Independent Variables," TREE-1216, UC-32, Idaho National Engineering Laboratory, March 1978.
3. Matthews, S. D., and Poloski, J. P., "MOCARS: A Monte Carlo Code for Determining Distribution and Simulation Limits and Ranking System Components by Importance," TREE 1138, UC-32, Idaho National Engineering Laboratory, August 1978.
4. Nelson, Lloyd S., "Using the Tenth and Ninetieth Percentiles and Mode to Characterize Lognormal Distributions," *The American Statistician*, Vol. 31, No. 1, February 1977.
5. Pearson, E. S., and Hartley, H. O., Eds., Biometrika Tables for Statisticians, Vol. I, Cambridge University Press, London, 1970. (Table 42, p. 230)

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

APPENDICES

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

APPENDIX A

THE LOGNORMAL DISTRIBUTION

The following is a brief summary of fundamental properties of the lognormal distribution. The notation employed in this report is, for the most part, identical to that used in Aitchinson and Brown [1]; in fact, most of the results given below can be found in Chapter 2 of that book. Another frequently encountered system of notation is discussed in Chapter 6 of Kendall and Stuart [5].

Definition of the Lognormal

A random variable x is said to have a (three-parameter) lognormal distribution with parameters μ , σ and d , if the random variable $y = \ln(x - d)$ has a normal distribution with mean μ and variance σ^2 . The above statements may be abbreviated " x is $\Lambda(\mu, \sigma^2, d)$ " and " y is $N(\mu, \sigma^2)$ ". Let x be $\Lambda(\mu, \sigma^2, d)$ and $z = [\ln(x - d) - \mu]/\sigma$ ($\sigma > 0$). Then z is $N(0, 1)$, a standard normal random variable, and

$$x = d + \exp(\mu + \sigma z) \quad . \quad (A.1)$$

If Φ denotes the distribution function of z , then the distribution function of x is

$$\begin{aligned} F(x_0) &= \Pr\{x < x_0\} = \Pr\{z < [\ln(x_0 - d) - \mu]/\sigma\} \\ &= \Phi([\ln(x_0 - d) - \mu]/\sigma) \quad . \end{aligned} \quad (A.2)$$

Similarly, if

$$\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$$

is the density function of z , then the density function of x is

$$f(x) = \begin{cases} \phi([\ln(x - d) - \mu]/\sigma)/[\sigma(x - d)] & , \text{ for } x > d \\ 0 & , \text{ for } x \leq d \end{cases} .$$

The mode (or most likely value) of x is

$$\text{mode}(x) = d + \exp(\mu - \sigma^2) \quad . \quad (\text{A.3})$$

Percentiles of the Lognormal

If $F(x_p) = \Pr\{x \leq x_p\} = p/100$ ($0 < p < 100$), then x_p is called the "p-th percentile of x ." Let v_p be the p-th percentile of z ; then from equations A.1 and A.2, one obtains

$$x_p = d + \exp(\mu + v_p \sigma) \quad , \quad 0 < p < 100 \quad . \quad (\text{A.4})$$

In particular, $v_{05} \doteq -1.645$ and $v_{95} \doteq 1.645$, so that the fifth and ninety-fifth percentiles of x are

$$x_{05} = d + \exp(\mu - 1.645\sigma) \quad \text{and} \quad (\text{A.5})$$

$$x_{95} = d + \exp(\mu + 1.645\sigma) \quad . \quad (\text{A.6})$$

The median of x is the fiftieth percentile, i.e.,

$$\text{median}(x) = x_{50} = d + \exp(\mu) \quad , \quad (\text{A.7})$$

since $v_{50} = 0$.

Moments of the Lognormal

If G is any function of x , the "expectation of $G(x)$ " is defined as

$$E[G(x)] = \int_d^{\infty} G(s)f(s)ds \quad .$$

Several important statistics of x are defined by expectations. Set

$$\alpha = \exp(\mu + \sigma^2/2) \text{ and } \eta^2 = \exp(\sigma^2) - 1 \quad (\eta > 0) \quad . \quad (A.8)$$

Then the mean (or average value or expected value or first moment) of x is

$$m = E[x] = d + \alpha = d + \exp(\mu + \sigma^2/2) \quad . \quad (A.9)$$

The variance of x (or second moment of x about the mean) is

$$v = \text{Var}(x) = E[(x - m)^2] = \alpha^2 \eta^2 \quad ; \quad (A.10)$$

the third moment of x (about the mean),

$$t = \lambda_3(x) = E[(x - m)^3] = \alpha^3 (\eta^6 + 3\eta^4) \quad ; \quad (A.11)$$

and the fourth moment of x (about the mean),

$$k = \lambda_4(x) = E[(x - m)^4] = \alpha^4 (\eta^{12} + 6\eta^{10} + 15\eta^8 + 16\eta^6 + 3\eta^4) \quad . \quad (A.12)$$

The (coefficient of) skewness of x is

$$\gamma_1(x) = \lambda_3(x)/[\text{Var}(x)]^{3/2} = \eta^3 + 3\eta \quad ; \quad (A.13)$$

and the (coefficient of) kurtosis of x,

$$\gamma_2(x) = \lambda_4(x)/[\text{Var}(x)]^2 = \eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^2 + 3 \quad . \quad (\text{A.14})$$

From the above equations, one can see that

$$\text{mode}(x) < \text{median}(x) < \text{mean}(x) \quad ;$$

and $\gamma_1(x) > 0$, i.e. the distribution of x always has positive skewness — though for σ^2 small, the skewness can be arbitrarily small and, hence, the distribution nearly symmetric. Examples of lognormal density functions with differing values of skewness are shown in Figure 5.

The Two-Parameter Lognormal Distribution

If x is $\Lambda(\mu, \sigma^2, 0)$, one may write " x is $\Lambda(\mu, \sigma^2)$ " and say that x has a two-parameter lognormal distribution. The reader should note that, unless otherwise specified, use of the term "lognormal" in this paper refers to the three-parameter lognormal; elsewhere in statistical literature the two-parameter lognormal is often implied. Familiar properties of the logarithm function and the normal distribution yield important results for the family of two-parameter lognormal distributions:

1. Reproductive Properties — Let x be $\Lambda(\mu_1, \sigma_1^2)$ and y be $\Lambda(\mu_2, \sigma_2^2)$ with x and y stochastically independent.
 - (a) If $p = xy$, then p is $\Lambda(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
 - (b) If $p = cx$ where $c > 0$ is a constant, then p is $\Lambda(\mu_1 + \ln(c), \sigma_1^2)$.
 - (c) If $p = x^c$ where c is any nonzero constant, then p is $\Lambda(c\mu_1, c^2\sigma_1^2)$. A case of special significance in this report is $c = -1$; the reciprocal of a $\Lambda(\mu, \sigma^2)$ random variable is a $\Lambda(-\mu, \sigma^2)$ random variable. (A.15)

Figure 5. Lognormal densities with mean five and unit variance.

2. Multiplicative Central Limit Theorem (see Aitchinson and Brown [1], p. 14) — Under fairly general conditions, the distribution of the product of many independent positive random variables is approximately a two-parameter lognormal.

The Negative Lognormal Distribution

A random variable y has a "negative lognormal distribution" if

$$y = d - \exp(\mu + \sigma z) \quad (\text{A.16})$$

where z is $N(0,1)$ and d, μ and $\sigma > 0$ are constant. Calculations of mode, percentiles and moments of a negative lognormal random variable are made easier by the following fact: if y has a negative lognormal distribution with parameters μ, σ and d , and $x = -y$, then x is $\Lambda(\mu, \sigma, -d)$. Thus

$$E(y) = -E(x), \text{Var}(y) = \text{Var}(x), \lambda_3(y) = -\lambda_3(x) \text{ and } \lambda_4(y) = \lambda_4(x) .$$

Then $\gamma_1(y) = -\gamma_1(x)$ and so the negative lognormal distribution has negative skewness. Also, the mode of y is $d - \exp(\mu - \sigma^2)$. Finally, if y_p, x_p and v_p are the p^{th} percentiles of y, x and z (as above), then

$$y_p = d - \exp(\mu - v_p \sigma) = -x_{1-p} \quad (\text{A.17})$$

since $v_{1-p} = -v_p$,

The Truncated Lognormal Distribution

Consider a lognormally distributed random variable which is truncated (or restricted) to the interval (a,b) . It is required to determine the impact of this truncation on the moments of the random variable.

First, it should be noted that only two-parameter distributions need to be considered since the third parameter, d , simply causes a linear displacement of the variable.

Thus, let y be a two-parameter lognormally distributed random variable truncated to the interval (a, b) . The resulting random variable, x , has the distribution function

$$\begin{aligned} F(x_0) &= \Pr[x < x_0] = \Pr[y < x_0 \mid a < y < b] \\ &= \frac{\Phi[(\ln x_0 - \mu)/\sigma] - \Phi[(\ln a - \mu)/\sigma]}{\Pr[a < y < b]} \end{aligned} \quad (A.18)$$

where

$$\Pr[a < y < b] = \Phi[(\ln b - \mu)/\sigma] - \Phi[(\ln a - \mu)/\sigma] \quad (A.19)$$

To evaluate the moments of x , define the standard normal random variable

$$z = (\ln y - \mu)/\sigma ;$$

and set

$$z_a = (\ln a - \mu)/\sigma , \text{ and}$$

$$z_b = (\ln b - \mu)/\sigma .$$

The n^{th} moment of x about the origin, λ'_n , is

$$\begin{aligned} \lambda'_n &= E[x^n] \\ &= E[y^n \mid a < y < b] \\ &= E[\exp(n\ln y) \mid z_a < z < z_b] , \end{aligned}$$

Then

$$\Pr[a < y < b] \lambda'_n = \frac{1}{\sqrt{2\pi}} \int_{z_a}^{z_b} \exp(n\mu + n\sigma z) \exp(-z^2/2) dz \quad . \quad (\text{A.20})$$

Completing the square in the exponent gives

$$\begin{aligned} \Pr[a < y < b] \lambda'_n &= \exp(n\mu + n^2\sigma^2/2) \frac{1}{\sqrt{2\pi}} \int_{z_a}^{z_b} \exp[-(z - n\sigma)^2/2] dz \\ &= \exp(n\mu + n^2\sigma^2/2) \frac{1}{\sqrt{2\pi}} \int_{z_a - n\sigma}^{z_b - n\sigma} \exp(-t^2/2) dt \\ &= \exp(n\mu + n^2\sigma^2/2) [\Phi(z_b - n\sigma) - \Phi(z_a - n\sigma)] \quad . \quad (\text{A.21}) \end{aligned}$$

Substituting from (A.19) and rearranging yields

$$\lambda'_n = \frac{\exp(n\mu + n^2\sigma^2/2) [\Phi(z_b - n\sigma) - \Phi(z_a - n\sigma)]}{\Phi(z_b) - \Phi(z_a)} \quad . \quad (\text{A.22})$$

The moments of x about the mean can be evaluated from expression A.22.

Table 9 illustrates the effect of truncation. Lognormal distributions with unit means and variances but with differing values of skewness were truncated at several higher percentiles. The impact on the mean and variance should be contrasted to the larger reductions in skewness and kurtosis.

Table 9
The Effect of Truncation on the Moments of Lognormal
Distributions with Unit Mean and Variance

Sigma	p	Statistics of lognormal distribution truncated at p th percentile			
		Mean	Variance	Skewness	Kurtosis
.25	97.50	0.9248	0.7890	0.324	2.63
	99.00	0.9645	0.8801	0.479	2.94
	99.50	0.9802	0.9244	0.566	3.17
	99.95	0.9973	0.9859	0.721	3.76
	*	1.0000	1.0000	0.778	4.10
.50	97.50	0.9093	0.6669	0.825	3.26
	99.00	0.9547	0.7920	1.067	4.08
	99.50	0.9737	0.8595	1.220	4.76
	99.95	0.9960	0.9675	1.556	6.82
	*	1.0000	1.0000	1.750	8.90
.75	97.50	0.8960	0.5189	1.299	4.46
	99.00	0.9449	0.6679	1.658	6.16
	99.50	0.9666	0.7583	1.911	7.69
	99.95	0.9942	0.9292	2.593	13.72
	*	1.0000	1.0000	3.263	26.54
1.00	97.50	0.8878	0.3623	1.754	6.13
	99.00	0.9366	0.5144	2.260	9.16
	99.50	0.9598	0.6190	2.651	12.18
	99.95	0.9920	0.8572	3.919	27.27
	*	1.0000	1.0000	6.185	113.94
1.25	97.50	0.8871	0.2208	2.195	8.22
	99.00	0.9320	0.3514	2.871	13.09
	99.50	0.9548	0.4537	3.432	18.36
	99.95	0.9896	0.7396	5.572	51.01
	*	1.0000	1.0000	13.148	800.46
1.50	97.50	0.8952	0.1143	2.622	10.67
	99.00	0.9327	0.2066	3.484	17.90
	99.50	0.9531	0.2895	4.240	26.25
	99.95	0.9876	0.5777	7.528	88.08
	*	1.0000	1.0000	33.468	10078.25
1.75	97.50	0.9110	0.0490	3.034	13.44
	99.00	0.9391	0.1017	4.090	23.49
	99.50	0.9556	0.1561	5.058	35.73
	99.95	0.9864	0.3946	9.709	140.14
	*	1.0000	1.0000	105.554	229898.09
2.00	97.50	0.9312	0.0170	3.431	16.48
	99.00	0.9501	0.0410	4.684	29.79
	99.50	0.9619	0.0694	5.870	46.65
	99.95	0.9866	0.2285	12.021	207.14
	*	1.0000	1.0000	414.359	9220559.98

*Untruncated

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

APPENDIX B

DETERMINING A LOGNORMAL FROM PERCENTILES AND CENTRAL VALUE

In Section IV the procedure for determining a density function to represent two percentiles and a central value was briefly described. Here we give further important details for the special case of determining a positively skewed lognormal distribution to fit the data. The notation used will conform to Section IV and Appendix A. For convenience many of the important equations will be repeated here.

Given the central value m and two percentiles, x_1 and x_2 , the equations to be solved for the unknown parameters μ , σ and d are

$$d + \exp(\mu - v_1\sigma) = x_1, \quad (B.1)$$

$$d + \exp(\mu + v_2\sigma) = x_2, \quad (B.2)$$

and either

$$d + \exp(\mu - \sigma^2) = m, \quad \text{or} \quad (B.3)$$

$$d + \exp(\mu) = m, \quad \text{or} \quad (B.4)$$

$$d + \exp(\mu + \sigma^2/2) = m, \quad (B.5)$$

depending on whether m is the mode, median or mean, respectively. Define the ratio,

$$r = (m - x_1)/(x_2 - x_1). \quad (B.6)$$

Note that r is the corresponding central value (mode, median or mean) if the linear transformation is made so that the percentiles are at 0 and 1.

By subtracting equation B.1 from the other two equations and dividing the two resulting differences, the following system of equations can be obtained:

$$R(\sigma) = r, \quad (B.7)$$

$$\mu = \ln\{[x_2 - x_1]/[\exp(v_2\sigma) - \exp(-v_1\sigma)]\}, \quad \text{and} \quad (B.8)$$

$$d = x_1 - \exp(\mu - v_1\sigma); \quad (B.9)$$

where the left-hand side of equation B.7, $R(\sigma)$, is a function of v_1 , v_2 and σ alone. Clearly, the exact form of $R(\sigma)$ depends on whether m is the mode, median or mean; these three cases will be discussed separately below. Whatever the interpretation of m , the function $R(\sigma)$ is nonlinear; hence, the Newton-Raphson method is used to find σ , the solution of B.7. Then the values of μ and d can be obtained by substitution into equations B.8 and B.9. Thus, questions about the existence and uniqueness of solutions to the system B.7-B.9 can be addressed by analyzing only the function $R(\sigma)$.

Mode

When m is the mode,

$$R(\sigma) = [\exp(-\sigma^2) - \exp(-v_1\sigma)]/[\exp(v_2\sigma) - \exp(-v_1\sigma)]. \quad (B.10)$$

Figure 6 shows a typical graph: $R(\sigma)$ is positive for σ between 0 and v_1 ; $R(\sigma)$ is negative for $\sigma > v_1$; from its maximum value at $\sigma = 0$, the function

(a) $R(\sigma) > 0$. (b) Detail when $R(\sigma) < 0$.

Figure 6. Mode ratio for fifth and ninety-fifth percentiles
 $(v_1 = v_2 = 1.645)$.

falls steadily through zero; eventually the curve turns upward and is asymptotic to 0 for large values of σ . Thus, there are unique solutions when r is positive; and two solutions when r is negative. Further, if r is too large (positive) or too small (negative), there are no solutions.

To make these notions more precise requires some care since $R(0)$ is an indeterminate form. The analysis of $R(\sigma)$ for small positive values of σ is easily accomplished through the use of Taylor's series:

$$\exp(-v_1\sigma) = 1 - v_1\sigma + (v_1^2/2)\sigma^2 - (v_1^3/6)\sigma^3 + \dots, \quad (\text{B.11})$$

$$\exp(v_2\sigma) = 1 + v_2\sigma + (v_2^2/2)\sigma^2 + (v_2^3/6)\sigma^3 + \dots, \text{ and} \quad (\text{B.12})$$

$$\exp(-\sigma^2) = 1 - \sigma^2 + (1/2)\sigma^4 - (1/6)\sigma^6 + \dots \quad (\text{B.13})$$

From equations B.10-B.13,

$$R(\sigma) = [v_1 - (1 + v_1^2/2)\sigma + \dots] / [(v_1 + v_2) + (v_2^2 - v_1^2)/2\sigma + \dots] \quad (\text{B.14})$$

Thus as σ decreases to zero, the maximum value of $R(\sigma)$ for $\sigma > 0$ is

$$R_{\max} = R(0+) = \lim_{\sigma \rightarrow 0+} R(\sigma) = v_1 / (v_1 + v_2) \quad ; \quad (\text{B.15})$$

and the slope at $\sigma = 0$ is

$$R'(0+) = \lim_{\sigma \rightarrow 0+} R'(\sigma) = - (1 + v_1 v_2 / 2) / (v_1 + v_2) < 0 \quad (\text{B.16})$$

Since the slope, $R'(v_1)$, is negative and $R'(\sigma)$ is positive for σ sufficiently large, there is a minimum value, $R_{\min}$, for $R(\sigma)$. If r is not between $R_{\min}$ and $R_{\max}$, there is no solution.

If $R_{\min} < r < 0$, there will usually be two solutions. The assumption throughout this study has been that the desired solution corresponds to the smaller value of σ , and, hence, the smaller coefficient of skewness. Figure 7 shows two lognormal density functions, each with mode equal to 4 and thirtieth and seventieth percentiles equal to 5 and 10, respectively. Algorithms based on the above analysis will always choose the solid curve ($\sigma = 0.98$) instead of the dashed curve ($\sigma = 1.5$).

The first step in the solution algorithm is the determination of $R_{\min}$ (and the corresponding maximum allowable value for σ); if r does not lie between $R_{\min}$ and $R_{\max}$, there is no solution. For permissible values of r , an initial approximation to σ is generated from equation B.14,

$$r = [v_1 - (1 + v_1^2/2) \sigma_o] / [(v_1 + v_2) + (v_2^2 - v_1^2) \sigma_o/2], \text{ or } (B.17)$$

$$\sigma_o = 2[(1 - r) v_1 - r v_2] / [(1 - r) v_1^2] \quad (B.18)$$

Successive iterates of σ are generated by

$$\sigma_{n+1} = \sigma_n - [R(\sigma_n) - r] / R'(\sigma_n) \quad , \quad (B.19)$$

until $|R(\sigma_n) - r|$ is sufficiently small. Then let $\sigma = \sigma_n$, and solve for μ and d from B.8 and B.9.

Figure 7. The two lognormal densities with mode, thirtieth percentile and seventieth percentile of four, five and ten respectively.

Median

When m is the median, the analysis proceeds as above with

$$R(\sigma) = [1 - \exp(-v_1\sigma)] / [\exp(v_2\sigma) - \exp(-v_1\sigma)] \quad . \quad (B.20)$$

From equations B.11 and B.12,

$$\begin{aligned} R(\sigma) = [v_1 - v_1^2\sigma/2 + \dots] / [(v_1 + v_2) \\ + (v_2^2 - v_1^2)\sigma/2 + \dots] \quad . \end{aligned} \quad (B.21)$$

Clearly $R(\sigma) > 0$ for all $\sigma > 0$; and it is not difficult to show that $R'(\sigma) < 0$ for all $\sigma > 0$. Thus, for $\sigma > 0$, the maximum value of $R(\sigma)$ is, as in B.15,

$$R_{\max} = R(0+) = \lim_{\sigma \rightarrow 0+} R(\sigma) = v_1 / (v_1 + v_2) \quad ; \quad (B.22)$$

also, as σ increases without bound, $R(\sigma)$ and $R'(\sigma)$ decrease to zero.

For $0 < r < R_{\max}$ the unique value of σ is found using the initial approximation from equation B.21,

$$\sigma_0 = 2[(1 - r)v_1 - rv_2] / [(1 - r)v_1^2 + rv_2^2] \quad , \quad (B.23)$$

and the solution procedure described at B.19 above. The graph of $R(\sigma)$ for fifth and ninety-fifth percentiles is shown in Figure 8.

Mean

When m is the mean, the function is

$$R(\sigma) = [\exp(\sigma^2/2) - \exp(-v_1\sigma)] / [\exp(v_2\sigma) - \exp(-v_1\sigma)] \quad . \quad (B.24)$$

Figure 8. Median ratio for fifth and ninety-fifth percentiles ($v_1 = v_2 = 1.645$).

Using the Taylor's series,

$$\exp(\sigma^2/2) = 1 + \sigma^2/2 + \sigma^4/8 + \dots, \quad (\text{B.25})$$

gives the expansion

$$\begin{aligned} R(\sigma) = & [v_1 + (1 - v_1^2) \sigma/2 + \dots] / [(v_1 + v_2) \\ & + (v_2^2 - v_1^2) \sigma/2 + \dots] \quad . \end{aligned} \quad (\text{B.26})$$

As before, $R(0+) = v_1/(v_1 + v_2)$. But in this case,

$$R'(0+) = (1 - v_1 v_2) / [2(v_1 + v_2)] \quad . \quad (\text{B.27})$$

If $v_1 v_2 \leq 1$, then $R'(0+) \geq 0$ and one expects a unique solution for all $R > v_1/(v_1 + v_2)$. An example of this case is shown in Figure 9(a) using thirtieth and seventieth percentiles ($v_1 = v_2 = 0.524$). The initial approximation to σ is determined from equation B.26,

$$\sigma_o = 2[v_1 - r(v_1 + v_2)] / [r(v_2^2 - v_1^2) - (1 - v_1^2)] \quad , \quad (\text{B.28})$$

and the algorithm proceeds as in B.19.

When $v_1 v_2 > 1$, the slope $R'(0+) < 0$. Since $R'(\sigma)$ is positive and arbitrarily large for sufficiently large values of σ , there are usually two solutions for σ if $r \leq R(0+) = v_1/(v_1 + v_2)$. To be consistent in selecting the density function possessing the smaller skewness, the algorithm must insure that $R'(\sigma) \leq 0$ if $r \leq R(0+)$. If $r > R(0+)$, one expects a unique solution to be found from B.28 and B.19.

(a) Mean ratio for thirtieth and seventieth percentiles ($v_1 = v_2 = .524$).

(b) Mean ratio for fifth and ninety-fifth percentiles ($v_1 = v_2 = 1.645$).

Figure 9. Mean ratio as a function of sigma.

APPENDIX C

THIRD AND FOURTH MOMENTS OF SUMS AND PRODUCTS
OF INDEPENDENT RANDOM VARIABLES

Sums of Random Variables

Let x and y be independent random variables with means m_x and m_y , variances v_x and v_y , and third and fourth moments t_x , t_y , k_x , and k_y .

Consider the random variable z where

$$z = x + y .$$

Then

$$\begin{aligned} t_z &= E(z - m_z)^3 = E[(x+y) - (m_x + m_y)]^3 \\ &= E[(x - m_x) + (y - m_y)]^3 \\ &= E(x - m_x)^3 + 3E(x - m_x)^2 E(y - m_y) + 3E(x - m_x) E(y - m_y)^2 + E(y - m_y)^3 . \end{aligned}$$

Since $E(y - m_y) = 0$ and $E(x - m_x) = 0$,

$$t_z = E(x - m_x)^3 + E(y - m_y)^3 .$$

Thus

$$t_z = t_x + t_y . \tag{C.1}$$

Similarly, for fourth moments

$$\begin{aligned}
 k_z &= E(z - m_z)^4 = E[(x+y) - (m_x + m_y)]^4 \\
 &= E[(x - m_x) + (y - m_y)]^4 \\
 &= E[(x - m_x)^4 + 4(x - m_x)^3 (y - m_y) + 6(x - m_x)^2 (y - m_y)^2 \\
 &\quad + 4(x - m_x) (y - m_y)^3 + (y - m_y)^4] \quad ; \quad \text{or} \\
 k_z &= k_x + 6v_x v_y + k_y \quad . \quad (C.2)
 \end{aligned}$$

Products of Random Variables

Let x and y be independent random variables with moments as above.

Let the random variable z be

$$z = x \cdot y \quad .$$

Then

$$\begin{aligned}
 t_z &= E(z - m_z)^3 = E(xy - m_x m_y)^3 \\
 &= E(x^3) E(y^3) - 3m_x m_y E(x^2) E(y^2) + 3m_x^2 m_y^2 E(x) E(y) \\
 &\quad - m_x^3 m_y^3 \quad . \quad (C.3)
 \end{aligned}$$

But

$$\begin{aligned}
 t_x &= E(x^3) - 3m_x E(x^2) + 2m_x^3 \\
 &= E(x^3) - 3m_x [E(x^2) - m_x^2] - m_x^3 \\
 &= E(x^3) - 3m_x v_x - m_x^3 .
 \end{aligned}$$

Thus, $E(x^3) = t_x + 3m_x v_x + m_x^3$; (C.4)

also, $E(x^2) = v_x + m_x^2$. (C.5)

Similar relations exist for y; substituting equations C.4 and C.5 into C.3 gives

$$\begin{aligned}
 t_z &= (t_x + 3m_x v_x + m_x^3) (t_y + 3m_y v_y + m_y^3) \\
 &\quad - 3m_x m_y (v_x + m_x^2) (v_y + m_y^2) + 2m_x^3 m_y^3 .
 \end{aligned}$$

Expanding and cancelling terms gives

$$t_z = t_x t_y + t_x m_y (3v_y + m_y^2) + t_y m_x (3v_x + m_x^2) + 6m_x m_y v_x v_y . \quad (C.6)$$

For fourth moments,

$$\begin{aligned}
 k_z &= E(xy - m_x m_y)^4 \\
 &= E(x^4 y^4) - 4m_x m_y E(x^3 y^3) + 6m_x^2 m_y^2 E(x^2 y^2) - 3m_x^4 m_y^4 .
 \end{aligned}$$

But

$$E(x^3 y^3) = t_z + 3m_z v_z + m_z^3$$

$$E(x^2 y^2) = v_z + m_z^2$$

$$E(x^4 y^4) = E(x^4) E(y^4), \text{ and}$$

$$E(x^4) = k_x + 4m_x t_x + 6m_x^2 v_x + m_x^4,$$

and similarly for $E(y^4)$. Thus

$$\begin{aligned} k_z &= E(x^4 y^4) - 4m_z (t_z + 3m_z v_z + m_z^3) + 6m_z^2 (v_z + m_z^2) - 3m_z^4 \\ &= E(x^4) E(y^4) - 4m_z t_z - 12m_z^2 v_z - 4m_z^4 + 6m_z^2 v_z + 6m_z^4 - 3m_z^4, \end{aligned}$$

which gives

$$\begin{aligned} k_z &= [k_x + 4m_x t_x + 6m_x^2 v_x + m_x^4] [k_y + 4m_y t_y + 6m_y^2 v_y + m_y^4] \\ &\quad - 4m_z t_z - 6m_z^2 v_z - m_z^4. \end{aligned} \quad (C.7)$$

APPENDIX D

APPROXIMATING PEARSON DISTRIBUTIONS BY CUBIC SPLINES

The density function, $f(x)$, of any Pearson distribution, normalized to have zero mean and unit variance, satisfies the differential equation,

$$df/dx = -(c_1 + x) f(x) / (c_0 + c_1 x + c_2 x^2) \quad (D.1)$$

The coefficients c_0 , c_1 and c_2 are determined from the skewness (γ_1) and the kurtosis (γ_2) as follows: set

$$\beta_1 = \gamma_1^2, \quad (D.2)$$

$$\beta_2 = \gamma_2, \quad (D.3)$$

$$A' = 10\beta_2 - 12\beta_1 - 18, \quad \text{and} \quad (D.4)$$

$$\kappa = \beta_1(\beta_2 + 3)^2 / [4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)] \quad (D.5)$$

Then

$$c_0 = (4\beta_2 - 3\beta_1) / A', \quad (D.6)$$

$$c_1 = \gamma_1(\beta_2 + 3) / A', \quad \text{and} \quad (D.7)$$

$$c_2 = (2\beta_2 - 3\beta_1 - 6) / A' \quad (D.8)$$

Also, it is convenient to let α_1 and α_2 denote the roots of the denominator, $c_0 + c_1 x + c_2 x^2$ (α_1 and α_2 may be identical, or they may be complex).

Determination of Pearson Type

The constants above may be used to determine the appropriate "Pearson type" distribution, and, thus, determine the analytical form of the density function. References for the Pearson system include [3], [4], [5] and [7]. It should be pointed out that within the Pearson system are distributions not considered appropriate for the work encompassed by this paper. These include certain "J-shaped" and "U-shaped" distributions with singularities at one end of the domain of the density function. For the most part, the cubic spline approach is quite insensitive to the particular Pearson type. However, several types have restricted domains—for this reason, a brief description of the various Pearson types is appropriate (see Reference [6]).

1. Type I (Beta distributions — $A' = 0$ or $\kappa < 0$): Here $\alpha_1 < 0 < \alpha_2$ and the domain of $f(x)$ is (α_1, α_2) .
2. Type II ($1 \leq \beta_2 < 3$ and $\beta_1 = 0$): This is a special case of Type I, $\alpha_1 = -\alpha_2$; the domain is $(-\alpha_2, \alpha_2)$.
3. Type III (Gamma distributions — $c_2 = 0$): The domain is $(-c_0/c_1, \infty)$ if $c_1 > 0$ or $(-\infty, -c_0/c_1)$ if $c_1 < 0$.
4. Type IV ($0 < \kappa < 1$; α_1 and α_2 are imaginary): The domain is $(-\infty, \infty)$.
5. Type V ($\kappa = 1$; α_1 and α_2 are real with the same sign): Let $L = c_1/(2c_2)$. Then the domain of $f(x)$ is $(=L, \infty)$ if $(c_1 - L)/c_2 \geq 0$ or $(-\infty, -L)$ if $(c_1 - L)/c_2 < 0$.
6. Type VI ($\kappa > 1$; α_1 and α_2 real with the same sign): The domain is (α_2, ∞) if $\alpha_1 < \alpha_2 < 0$ or $(-\infty, \alpha_1)$ if $0 < \alpha_1 < \alpha_2$.
7. Type VII (Student's t distribution — $\beta_1 = 0$, $\beta_2 > 3$): The domain is $(-\infty, \infty)$.

8. Normal ($\beta_1 = 0, \beta_2 = 3$): The domain is $(-\infty, \infty)$.
9. Exponential ($\beta_1 = 4, \beta_2 = 9$): The domain is $(-1, \infty)$ for $c_1 > 0$
or $(-\infty, 1)$ for $c_1 < 0$.

Specifying the Nodes of the Spline

The location of the nodes significantly affects the accuracy of the cubic spline approximation. It is especially important to have more nodes where the distribution function is steep and fewer where it is relatively flat. In particular, if the distribution is unimodal, then it is easily seen from equation D.1 that the mode occurs at $x = -c_1$; in this case it seems reasonable to place more nodes near the mode and fewer in the tails to improve the accuracy.

The first step is the determination of the theoretical domain of the distribution function (from the type designations above). If either tail is theoretically infinite the corresponding exterior node is set to -10 or +10. Although many of the Pearson distributions can be adequately approximated without resorting to such extreme values, using ten standard deviations away from the mean was required to represent uniformly well all the distributions encountered in this study.

In theory, any number of nodes could be placed between the limits determined above. In practice, the number of nodes necessary to achieve the desired accuracy is over 100. To keep the computer memory requirements on the DECsystem-10 small, 151 nodes are located. For unimodal distributions, 75 nodes are spaced "logarithmically" away from the mode toward each exterior node (this means that interval lengths are smallest near the mode and largest near the outer limits; and the ratio of the lengths of neighboring intervals is a constant). Note that the

specifications of initial interval length and interval length ratio are not unique; based on empirical evidence, these values are selected as a function of skew. (Details are described in Appendix G under SUBROUTINE GNODES.) If $x = -c_1$ falls outside the limits, the distribution is "J-shaped"; the entire 151 nodes are logarithmically spaced from the end-point nearest $x = -c_1$ toward the other.

Definining the Cubic Spline

Let $F(x)$ denote the distribution function whose density, $F'(x) = f(x)$, satisfies equation D.1. Let

$$P(x) = c_1 + x, \quad \text{and}$$

$$Q(x) = c_0 + c_1 x + c_2 x^2.$$

Then equation D.1 may be written

$$Q(x) F''(x) + P(x) F'(x) = 0. \quad (D.9)$$

Let $S(x)$ be the cubic spline approximation to $F(x)$ with nodes

$x_1 < x_2 < \dots < x_n$. This means, first of all, that $S(x)$ is a piecewise cubic, i.e. there exist constants a_{1k} , a_{2k} , a_{3k} and a_{4k} , so that for $x_k \leq x < x_{k+1}$,

$$S(x) = a_{1k} + a_{2k}(x - x_k) + a_{3k}(x - x_k)^2/2 + a_{4k}(x - x_k)^3/6, \quad (D.10)$$

for $1 \leq k \leq n-1$. Note that for $x_k \leq x < x_{k+1}$, the derivatives of $S(x)$ are

$$S'(x) = a_{2k} + a_{3k}(x - x_k) + a_{4k}(x - x_k)^2/2, \quad \text{and} \quad (\text{D.11})$$

$$S''(x) = a_{3k} + a_{4k}(x - x_k). \quad (\text{D.12})$$

Thus

$$S(x_k) = a_{1k}, \quad S'(x_k) = a_{2k}, \quad S''(x_k) = a_{3k}, \quad \text{and} \quad S'''(x_k) = a_{4k}. \quad (\text{D.13})$$

Since there are $(n-1)$ intervals between nodes, there are $4(n-1)$ unknown coefficients (a_{ik}) to be determined. First, because $S(x)$ is a distribution function, it must satisfy the two boundary conditions, $S(x_1) = 0$ and $S(x_n) = 1$, or

$$a_{11} = 0, \quad \text{and} \quad (\text{D.14})$$

$$\begin{aligned} a_{1,n-1} + a_{2,n-1}(x_n - x_{n-1}) + a_{3,n-1}(x_n - x_{n-1})^2/2 \\ + a_{4,n-1}(x_n - x_{n-1})^3/6 = 1. \end{aligned} \quad (\text{D.15})$$

Second, $S(x)$ is required to satisfy the differential equation D.9 at each of the nodes,

$$P(x_k) a_{2k} + Q(x_k) a_{3k} = 0, \quad 1 \leq k \leq n. \quad (\text{D.16})$$

The remaining $3(n-2)$ equations are the continuity constraints — we require that $S(x)$, $S'(x)$ and $S''(x)$ be continuous at each of the interior nodes (i.e. that the various cubic segments join together smoothly at common nodes),

$$S: \quad a_{1k} + d_k a_{2k} + (d_k^2/2) a_{3k} + (d_k^3/6) a_{4k} = a_{1,k+1} \quad , \quad (D.17)$$

$$S': \quad a_{2k} + d_k a_{3k} + (d_k^2/2) a_{4k} = a_{2,k+1} \quad , \quad \text{and} \quad (D.18)$$

$$S'': \quad a_{3k} + d_k a_{4k} = a_{3,k+1} \quad , \quad (D.19)$$

for $2 \leq k \leq n-1$ (and we define $d_k = x_{k+1} - x_k$). The system of $4(n-1)$ equations given in D.14 through D.19 is linear in the unknown coefficients (a_{ik}); in fact, if the equations are sorted on k , the resulting coefficient matrix is a band matrix of width 5 (2 diagonals above and below the main diagonal). By utilizing the narrow band structure of the coefficient matrix, the solution may be found efficiently and economically, in terms of both computer memory and time.

After the coefficients of the spline have been determined, percentiles, probabilities, etc., relating to the true distribution may be approximated. Let $F(x)$ denote the true distribution function with mean μ and standard deviation σ , and $f(x)$, the true density function. Then the approximation is

$$F(x) = S[(x - \mu)/\sigma] \quad , \quad \text{and} \quad (D.20)$$

$$f(x) = S'[(x - \mu)/\sigma]/\sigma \quad . \quad (D.21)$$

Examples of Pearson distributions with zero means and unit variances are shown in Figure 10. The types shown are typical of those generated in this study; no attempt has been made to illustrate the full range of flexibility of these distributions.

(a) Densities with positive skewness.

(b) Symmetric densities ($\beta_1 = 0$).

Figure 10. Examples of Pearson density functions with zero means and unit variances.

APPENDIX E

EFFECT OF THE PROBABILITY OF OCCURRENCE ON MOMENTS

Let m_y , v_y , t_y , k_y be the mean, variance and third and fourth moments about the mean of the conditional distribution, $f_y(y)$, describing the magnitude of an event and let P_o be the probability that the event occurs. Let x be the unconditional quantity described by

$$f_x(x) = P_o f_y(x) \text{ when } x > 0 \quad ; \quad \text{and}$$

$$\Pr(x=0) = P_o \quad . \quad (7.1)$$

The mean of x is

$$m_x = E(x) = (1 - P_o) E(x|x = 0) + P_o E(x|x > 0)$$

$$\text{with} \quad E(x|x > 0) = E(y) = m_y \quad .$$

Thus

$$m_x = P_o m_y \quad . \quad (E.1)$$

The variance of x is

$$\begin{aligned} v_x = E(x - m_x)^2 &= (1 - P_o) E[(x - m_x)^2|x = 0] \\ &+ P_o E[(x - m_x)^2|x > 0] \end{aligned}$$

with $E[(x - m_x)^2 | x = 0] = (-m_x)^2 = P_o^2 m_y^2$

and $E[(x - m_x)^2 | x > 0] = E(y - P_o m_y)^2$.

Thus
$$v_x = (1 - P_o) P_o^2 m_y^2 + P_o E(y^2 - 2P_o m_y y + P_o^2 m_y^2)$$

$$= (1 - P_o) P_o^2 m_y^2 + P_o (v_y + m_y^2 - 2P_o m_y^2 + P_o^2 m_y^2)$$

which simplifies to

$$v_x = P_o [v_y + (1 - P_o) m_y^2] \quad (E.2)$$

The unconditional third moment about the mean is

$$\begin{aligned} t_x &= (1 - P_o) E[(x - m_x)^3 | x = 0] + P_o E[(x - m_x)^3 | x > 0] \\ &= (1 - P_o) (-P_o m_y)^3 + P_o E[(y - P_o m_y)^3] \\ &= (1 - P_o) (-P_o m_y)^3 + P_o [E(y^3) - 3P_o m_y E(y^2) \\ &\quad + 3P_o^2 m_y^2 E(y) - P_o^3 m_y^3] \\ &= - (1 - P_o) P_o^3 m_y^3 + P_o [t_y + 3m_y v_y + m_y^3 - 3P_o m_y v_y \\ &\quad - 3P_o m_y^3 + 3P_o^2 m_y^3 - P_o^3 m_y^3] \end{aligned}$$

which simplifies to

$$t_x = P_o [t_y + 3(1-P_o) m_y v_y + (1-P_o) (1-2P_o) m_y^3] \quad (E.3)$$

A similar development for fourth moments is outlined:

$$\begin{aligned} k_x &= (1-P_o) E[(x-m_x)^4 | x = 0] + P_o E[(x-m_x)^4 | x > 0] \\ &= (1-P_o) (-m_y P_o)^4 + P_o E(y - P_o m_y)^4 \\ &= (1-P_o) m_y^4 P_o^4 + P_o [E(y^4) - 4P_o m_y E(y^3) + 6P_o^2 m_y^2 E(y^2) \\ &\quad - 4P_o^3 m_y^3 E(y) + P_o^4 m_y^4] \end{aligned}$$

which simplifies to

$$\begin{aligned} k_x &= P_o [k_y + 4(1-P_o) m_y t_y + 6(1-P_o)^2 m_y^2 v_y \\ &\quad + (1-P_o) (1-3P_o + 3P_o^2) m_y^4] \quad (E.4) \end{aligned}$$

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

APPENDIX F

MOMENTS OF SUMS OF PERFECTLY CORRELATED

RANDOM VARIABLES

Let x and y be random variables which are perfectly positively correlated. Form

$$z = x + y \quad . \quad (F.1)$$

The means of x , y , and z are denoted m_x , m_y , and m_z with

$$m_z = m_x + m_y \quad . \quad (F.2)$$

The n^{th} central moment (with $n > 1$) of x is written as $\lambda_n(x)$; similar notation holds for y and z .

The assumption of perfect positive correlation implies that

$$y = ax + b \quad , \quad (F.3)$$

where a and b are coefficients and $a > 0$. Then

$$m_y = E(ax + b) = am_x + b \quad ; \quad (F.4)$$

and

$$\begin{aligned} \lambda_n(y) &= E(ax + b - m_x - b)^n \\ &= a^n \lambda_n(x) \quad . \end{aligned} \quad (F.5)$$

Form

$$\begin{aligned}
 \lambda_n(z) &= E(z - m_z)^n \\
 &= E(x + y - m_x - m_y)^n \\
 &= E(x + ax + b - m_x - am_x - b)^n \\
 &= E[(1 + a)x - (1 + a)m_x]^n \\
 &= (1 + a)^n \lambda_n(x) \quad . \quad (F.6)
 \end{aligned}$$

Taking the n^{th} root of both sides gives

$$\begin{aligned}
 \lambda_n(z)^{1/n} &= (1 + a) \lambda_n(x)^{1/n} \\
 &= \lambda_n(x)^{1/n} + a \lambda_n(x)^{1/n} \quad . \quad (F.7)
 \end{aligned}$$

Using (F.5), this gives

$$\lambda_n(z)^{1/n} = \lambda_n(x)^{1/n} + \lambda_n(y)^{1/n}$$

and hence

$$\lambda_n(z) = [\lambda_n(x)^{1/n} + \lambda_n(y)^{1/n}]^n \quad . \quad (F.8)$$

Special cases of equation F.8 are given in Section VIII.

APPENDIX G

COMPUTER PROGRAMS

The subroutines for performing the computations developed in this report are described in this Appendix. Main control routines which accept data, call subroutines to perform the desired analysis, and write reports have not been listed; these routines are tailored to particular problems, data formats, and report requirements and are thus not considered to be of general interest.

The routines are all written in FORTRAN IV and were prepared on a DECsystem-10 computer. The codes should be readily convertible to other machines; known exceptions are the use of IMPLICIT DOUBLE PRECISION statements and the storage of ten alphanumeric characters in a double-precision word. Unless otherwise indicated, all real variables are double precision. A microfiche listing of the routines is included in the envelope at the end of the report.

The following summaries provide a statement of the purpose of each routine, a description of arguments in the order of their appearance in the argument list, and other routines or functions called by the routine. In the argument descriptions, use is made of the symbols introduced in Section IV and in Appendix A. Additionally, frequent reference is made to a distribution "flag." This integer variable is used to communicate to a subroutine the type of distribution being analyzed. The codes are:

- 1 for a negatively skewed lognormal,
- 0 for a normal, and
- +1 for a positively skewed lognormal distribution.

Routines for Finding Lognormal Distributions Given Percentiles and Central Measure

SUBROUTINE L4MODE: From a lower percentile, mode, and upper percentile, the parameters of a normal or three-parameter lognormal distribution are found.

Arguments: lower percentile, mode, and upper percentile
 associated percentiles of the standard normal distribution
 three computed parameters, μ , σ , and d
 the distribution flag (-1, 0, 1) or, if input data is improper, a value of 2 is returned.

Subroutine called: L3MODE, finds the parameters for lognormal distributions.

SUBROUTINE L3MODE: The parameters of a three-parameter lognormal distribution are found given two quantiles and the mode.

Arguments: lower quantile, mode, and upper quantile
 associated lower and upper quantiles of the standard normal distribution
 returns μ , σ , and d
 an error flag for normal return (flag = 0), convergence problem (flag = 1), or improper data (flag = 2).

SUBROUTINE L4MEDN: This is similar to L4MODE except the median replaces the mode as the central value. With this exception, all arguments are the same.

Subroutine called: L3MEDN, finds the parameters for a lognormal distribution starting with a lower percentile, median, and upper percentile.

SUBROUTINE L3MEDN: Given two percentiles and the median, a three-parameter lognormal distribution is found (the two percentiles must bracket the median).

Arguments: the lower percentile, median, and upper percentile

 the associated lower and upper percentiles of the standard normal distribution

 the parameters of the lognormal distribution, μ , σ , and d

 the distribution and error flag, as in L4MODE.

SUBROUTINE L3MEAN: This routine finds the parameters of a three-parameter distribution from the fifth and ninety-fifth percentiles and the mean.

Arguments: the fifth percentile, mean, and ninety-fifth percentile

 the parameters μ , σ , and d

 an error flag indicating success (0), convergence failure (1), or that a fit is impossible (2).

SUBROUTINE L2MODE: This program obtains the best least-squares fit of a two-parameter lognormal distribution to given quantiles and mode.

Arguments: lower quantile, mode, and upper quantile

 associated lower and upper quantiles of the standard normal distribution

 returns μ and σ

 an error flag equal to zero (normal return) or one (convergence failure).

SUBROUTINE L2MEAN: Given a fifth percentile, mean, and ninety-fifth percentile, this routine finds the best two-parameter lognormal distribution using a least-squares criterion. A distribution is found that has the given mean and which minimizes the sum of squares of the differences between the input and computed percentiles.

Arguments: the input values of fifth percentile, mean, and ninety-fifth percentile

the parameters μ and σ returned.

SUBROUTINE L2MEDN: This routine is similar to L2MEAN using the median instead of the mean as the central value for the distribution.

Routines for Finding Moments of Lognormal Distribution

SUBROUTINE L4CMOM: This routine finds the first four central moments of normal or lognormal distributions (untruncated).

Arguments: μ , σ , and d , the parameters of the distribution

the distribution flag, (-1, 0, 1)

the four moments.

SUBROUTINE TRNCL4: This is a control routine for finding the first four moments of lognormal distributions truncated to an interval (a, b). Normal distributions are not truncated.

Arguments: μ , σ , and d of the distribution to be truncated

the distribution flag, (-1, 0, 1)

the first four central moments of the truncated distribution are returned

the probability enclosed by the interval (a, b) in the original (untruncated) distribution.

Subroutines called: TRNCLN, to truncate lognormal distributions

L4CMOM, to set return arguments for normal distributions.

SUBROUTINE TRNCLN: This routine computes four moments of a three-parameter lognormal distribution truncated to an interval (a, b).

Arguments: μ , σ , and d of the distribution
the interval limits, (a, b)
the four moments to be returned
the probability enclosed by the interval (a, b) in the untruncated distribution.

Subroutine called: Function CDFN01, calculates the cumulative distribution function for the standard normal.

Routines for Obtaining Moments of Functions of Random Variables

SUBROUTINE PRD4M: The moments of the product of two independent random variables are computed as described in Section V.

Arguments: the first four central moments of the first factor
the four moments of the second factor
the four moments of the product.

SUBROUTINE SUM4MI: This routine computes the first four moments of the sum of two independent random variables (see Section V).

Arguments: the first four central moments of the first variable in the sum
the four moments of the second variable
the four moments of the sum.

SUBROUTINE PCSUM: This routine computes the first four moments of the sum of two perfectly correlated random variables (see Section VII). Arguments are the same as in SUM4MI.

SUBROUTINE UNCMOM: Given the conditional moments and the probability of occurrence, P_0 , the unconditional moments are computed (see Section VII).

Arguments: the first four central moments of the conditional distribution

 the probability of occurrence

 the four unconditional moments, returned.

Routines for Fitting Lognormal Distributions to Moments

SUBROUTINE LN4MOM: This is a control routine for supplying the parameters of a normal or a lognormal distribution which will have the specified moments.

Arguments: the first three central moments of a distribution

 the three parameters, μ , σ , and d , of the distribution which has those moments

 the distribution flag, (-1, 0, 1).

Subroutine LN3MOM obtains the parameters for lognormal
called: distributions.

SUBROUTINE LN3MOM: This routine computes the three parameters of the lognormal distribution which has the three specified moments.

Arguments: the first three central moments

 the parameters, μ , σ , and d , returned.

Routines for Obtaining Probabilities for Lognormal Distributions

FUNCTION CDFLN4: For a normal or lognormal distribution, F , this control function returns the cumulative distribution, $F(x)$.

Arguments: the distribution parameters, μ , σ , and d .
the distribution type flag (-1, 0, 1)
 x for which $F(x)$ is to be computed.

Subroutines Called: Function CDFLN3 obtains $F(x)$ for lognormal distributions
Function CDFN01 obtains $F(x)$ for standard normal distributions.

FUNCTION CDFLN3: For a lognormal distribution, F , with parameters μ , σ , and d , the probability, $F(x)$, is found for any value of x .

Arguments: μ , σ , and d
 x .

Function called: CDFN01 obtains $F(x)$ for standard normal distributions.

A Routine for Finding Quantiles

FUNCTION QNTLN4: For a lognormal or normal distribution, this function returns the p^{th} quantile.

Arguments: the three parameters of the distribution, μ , σ , and d
the flag for distribution type (-1, 0, 1)
the probability, p , for which the quantile is to be found.

Subroutine called: Function QNTN01 finds the p^{th} quantile of the standard normal distribution.

Routines Concerning Standard Normal Distributions

FUNCTION CDFN01: This function returns the probability that a standard normal random variable is less than the argument x .

FUNCTION QNTN01: For the standard normal distribution function, F , this function returns the value x such that $F(x) = p$, where p is the given argument.

Routines Concerning Pearson Distributions

SUBROUTINE INPRSN: From the first four moments of a distribution, this routine determines the type of Pearson distribution to be used and computes the constants in the Pearson differential equation to be solved by subroutine GCPRSN (see Appendix E). These constants are found such that the distribution has a mean of zero and variance of one. The true mean and variance are stored for later use.

Arguments: the four moments of the distribution to be represented

on completion, the lower and upper limits of the Pearson distribution which is fit to the moments

on completion, the values of β_1 , β_2 , A' , and κ (Appendix D, expressions D.2 through D.5)

on completion, the constants in the differential equation, c_0 , c_1 , and c_2 (Appendix D, expressions D.6 through D.8)

the type of distribution (1 through 7) is returned in an integer variable

an alphanumeric description of the type of distribution (e.g., "TYPE IV") is returned in a real variable.

Subroutine called: PTYPE finds the distribution type from values of A' and κ .

- FUNCTION CDFN01: This function returns the probability that a standard normal random variable is less than the argument x .
- FUNCTION QNTN01: For the standard normal distribution function, F , this function returns the value x such that $F(x) = p$, where p is the given argument.

Routines Concerning Pearson Distributions

SUBROUTINE INPRSN: From the first four moments of a distribution, this routine determines the type of Pearson distribution to be used and computes the constants in the Pearson differential equation to be solved by subroutine GCPRSN (see Appendix E). These constants are found such that the distribution has a mean of zero and variance of one. The true mean and variance are stored for later use.

Arguments: the four moments of the distribution to be represented

on completion, the lower and upper limits of the Pearson distribution which is fit to the moments

on completion, the values of β_1 , β_2 , A' , and κ (Appendix D, expressions D.2 through D.5)

on completion, the constants in the differential equation, c_0 , c_1 and c_2 (Appendix D, expressions D.6 through D.8)

the type of distribution (1 through 7) is returned in an integer variable

an alphanumeric description of the type of distribution (e.g., "TYPE IV") is returned in a real variable.

Subroutine called: PTYPE finds the distribution type from values of A' and κ .

SUBROUTINE PTYPE: This routine determines the type of Pearson distribution (if any) which will satisfy given values of β_1 and β_2 .

Arguments: β_1 , β_2 , A' , and κ (see Appendix D)

an integer value indicating the type found (1-7); set to zero if in the impossible region

an alphanumeric description of the type is returned

a tolerance used at the boundaries between the types; set in subroutine INPRSN.

SUBROUTINE GCPRSN: This routine finds the cubic spline representation of the solution of the Pearson differential equation

$$P(x) F''(x) + Q(x) F'(x) = 0 \quad .$$

Details of the solution procedure are contained in Appendix D. Involved in this process is the solution of a set of linear equations, $Ax = c$. The matrix, A , is a band matrix of bandwidth 5. The matrix A is placed, in band storage mode, into an array W . This storage mode is described under Subroutine BNDSTO.

Arguments: a vector of length n of predetermined values of the random variable at which nodes of the spline are to be located. Subroutine GNODES can be used to obtain these points

n , the vector length (an integer)

a vector, c , of length $4(n-1)$ which contains, on completion, the coefficients of the spline segments as defined in Appendix E. Thus, $c(1)$ through $c(4)$ will contain the coefficients for the first interval, a_{11} through a_{14} , $c(5)$ through $c(8)$ will contain the second interval's coefficients, a_{21} through a_{24} , and so forth. (Also, during computation, c is used as the right-hand side vector.)

a double-precision work array, W , required by the linear algebra routines, dimensioned $(7, 4(n-1))$

an integer work vector dimensioned $4(n-1)$

SUBROUTINE GCPRSN:
(cont'd)

Arguments: an integer code is returned containing a condition
(cont'd) code from the linear algebra routine (see "Sub-
routines called," below)

an integer code set in the calling routine which triggers output within this subroutine; if set to zero, no output will occur; if negative, the equations and their solutions will be printed; if positive, the solutions only will print.

Subroutines called: Entry PRSNQP (subroutine INPRSN) is called to find values of P(x) and Q(x)

repeated calls to BNDSTO load the work array, W, with the coefficients matrix in band storage mode and the right-hand side vector, c; these to be solved by the following subroutine

subroutine DGBFA with entry DGBSL are routines resident on all major computers in the Oak Ridge complex for solving $Ax = c$ with A a band matrix stored in array W. Other routines could be substituted for these; the manner in which the matrix A is loaded into W is described under subroutine BNDSTO.

SUBROUTINE BNDSTO: This routine stores a row of coefficients into the work area matrix, W, in a form compatible with the band matrix routines DGBFA and DGBSL. These are routines resident on all major computers in the Oak Ridge complex. Elements of the band matrix, W, are derived from the elements, m_{ij} , of the square matrix, M, by the operation

$$m_{ij} \rightarrow w_{kj}$$

where

$$k = i - j + u + 1$$

and u is the number of diagonals above the main diagonal.

Arguments: the row vector of coefficients within the band
the number of elements in the row vector (an integer)

SUBROUTINE BNDSTO:
(cont'd)

Arguments: the row number of matrix M being stored (an integer)

(cont'd)

the column of M in which the band of this row begins (an integer)

the number of diagonals below the main diagonal (an integer)

the number of diagonals above the main diagonal (an integer)

the band matrix, W, dimensioned (n_w , $4(n - 1)$) in the calling routine

n_w , the first dimension of W (in the present problem, n_w is always equal to 7).

SUBROUTINE GNODES: This control routine may be used to determine the logarithmically spaced nodes for the cubic spline solution of the Pearson differential equation.

Arguments: the minimum value at which a point is to be placed

the mode of the distribution

the maximum point

the vector of nodes returned

the number of nodes (an integer)

the value of β_1 , used to compute the spacing ratio, r, which is sent to subroutine LGMPTS. The relationship between r and β_1 was chosen by observing the effect of r on the accuracy of the 5th and 95th percentiles for various values of β_1 and β_2 for which results are published. The relationship is

$$r = 1 + .032 \exp(-1.73\sqrt{\beta_1})$$

Subroutine called: LGMPTS obtains the logarithmically spaced nodes.

SUBROUTINE LGMPTS: This routine returns n nodes spaced logarithmically with spacing ratio, r .

Arguments: x , the vector of nodes

n , the number of nodes

a and b , the bounds on the nodes. Note that the spacing is from a toward b , regardless of their values.

r , the spacing ratio.

FUNCTION QLPRSN: This function returns the α^{th} quantile of the spline approximation to a distribution.

Arguments: x , a vector defining the location of the nodes of the spline

n , the length of the vector x (an integer)

c , a vector of coefficients of the cubic spline segments (see subroutine GCPRSN)

the probability, α .

Common variables: the mean and standard deviation of the distribution.

Subroutines called: Function INTVL2 finds the spline interval in which a value of the random variable falls

Function FCPRSN gives the distribution function for a value of the random variable.

FUNCTION FCPRSN: Let F be the spline approximation to the distribution function of a random variable. This function returns values of $F(x)$ for any x in the range of the distribution. Entry point DCPRSN returns values of the density function, $f(x)$, and entry point F2PRSN returns values of $f'(x)$.

Arguments: x , the value at which the function is to be evaluated

an integer giving the spline interval in which x falls (see function INTVL2)

FUNCTION FCPRSN:
(cont'd)

Arguments:	the vector of values at which nodes of the spline are located
(cont'd)	the number of nodes (an integer)
	the vector of spline coefficients (see subroutine GCPRSN).
Common variables:	the mean and standard deviation of the distribution.

FUNCTION INTVL2: This function returns the spline interval in which a value, x , of the random variable falls. Binary search is used.

Arguments:	x
	an integer estimate of the interval (optional)
	the vector of node values
	the number of nodes (an integer).
Common variables:	the mean and standard deviation of the distribution.

Internal Distribution

1. L. L. Anthony	21. J. F. Jamison
2. J. W. Arendt	22. V. E. Kane
3. J. J. Beauchamp	23. J. D. McGaugh
4. K. O. Bowman	24-35. R. A. McLaren
5. H. P. Carter/A. A. Brooks/
ORNL CSD Library	36-38. D. S. Napolitan
6. R. C. Durfee	39. N. H. Van Wie
7-18. C. E. Ford	40. G. W. Westley
19. P. W. Gaffney	41. ORGDP CSD Library
20. J. E. Hawthorne	42-44. ORGDP Library
	45. ORGDP Records - RC

External Distribution

- 46-145. David M. Blanchfield, U. S. Department of Energy, Grand Junction Office, P. O. Box 2567, Grand Junction, Colorado 81501
146. David Grundy, U. S. Geological Survey, MS-937, P. O. Box 25046, Federal Center, Denver, Colorado 80225
147. Dr. D. P. Harris, Department of Mining and Geological Engineering, University of Arizona, Tucson, Arizona 85724
148. Richard B. McCammon, U. S. Geological Survey, National Center 920, Reston, Virginia 22092
- 149-150. Greg Piepel, Battelle Pacific Northwest Laboratories, P. O. Box 999, Richland, Washington 99352
- 151-334. DOE/TIC-4500, UC-32, Mathematics and Computers