INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

University Microfilms International
300 N. Zeeb Road
Ann Arbor, MI 48106
Kurth, Robert Edward

THE DEVELOPMENT AND APPLICATION OF A NEW PROBABILISTIC ANALYSIS TECHNIQUE FOR NUCLEAR RISK CALCULATIONS

The Ohio State University

Ph.D. 1985

University Microfilms International 300 N. Zeeb Road, Ann Arbor, MI 48106

Copyright 1985 by Kurth, Robert Edward All Rights Reserved
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy _____
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages ✓
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) ___________ lacking when material received, and not available from school or author.
12. Page(s) ___________ seem to be missing in numbering only as text follows.
13. Two pages numbered __________. Text follows.
14. Curling and wrinkled pages _____
15. Other ____________________

University Microfilms International
THE DEVELOPMENT AND APPLICATION OF A NEW PROBABILISTIC ANALYSIS TECHNIQUE FOR NUCLEAR RISK CALCULATIONS

DISSERTATION
Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

by
ROBERT E. KURTH

The Ohio State University
1985

Reading Committee:  
Dr. T. Aldemir
Dr. R. Christensen
Dr. J. Waddell

Approved by:  
Adviser
Department of Nuclear Engineering
To my best friend, who also is my wife - Katie
ACKNOWLEDGMENTS

I gratefully acknowledge the following people without whom this work could not have been completed:
Dr. Richard Christensen, adviser
Ms. Barbara Foris, typist
Ms. Ellen Molnar, illustrator
VITA

October 11, 1953 .......... Born - Madison, Wisconsin
1975. .................. B.S., Mathematics, University of
Notre Dame, South Bend, Indiana
1975-1979 ............... Research Assistant, Nuclear Engineering,
The Ohio State University, Columbus,
Ohio
1977. .................. M.Sc., Nuclear Engineering, The
Ohio State University, Columbus,
Ohio
1979-1984 ............... Research Scientist, Battelle Memorial
Institute

PUBLICATIONS

publication, Volume 72, Random Life Prediction

Kurth, R.E. and Baybutt, P., "Uncertainties in LWR Meltdown Accident
Sequences", NUREG/CP-0017, 1982

Kurth, R.E., and Cox, D.C., "An Investigation of Discrete Probability
Distributions For Probabilistic Fracture Mechanics Analysis", accepted
for publication in Risk Analysis

FIELDS OF STUDY

Major Field: Nuclear Engineering

Studies in nuclear methods. Professor S. Nakamura

Studies in risk and safety analysis. Professor R. Christensen
TABLE OF CONTENTS

ACKNOWLEDGEMENT .............................................................. iii
VITA ............................................................................................... iv
LIST OF TABLES ............................................................................ vii
LIST OF FIGURES .......................................................................... ix

CHAPTER

1. INTRODUCTION .......................................................................... 1

2. PROBABILISTIC ANALYSIS METHODS ..................................... 5

   2.1 Introduction ........................................................................... 5

   2.2 Statistical and Probabilistic Analysis ................................. 6

   2.3 Statistical Formulas .............................................................. 7
       - Probability Density Function (PDF) ..................................... 7
       - Cumulative Distribution Function (CDF) ......................... 7
       - Moment Calculations ....................................................... 8
       - Coefficients of Skewness and Kurtosis ............................ 9
       - Distribution Types ......................................................... 9
         - Uniform Distribution .................................................... 11
         - Normal Distribution .................................................... 11
         - Lognormal Distribution .............................................. 13
         - Rayleigh Distribution ................................................ 13

   2.4 Methods for Probabilistic Analysis .................................... 15
       - Probabilistic Models ...................................................... 19
       - Approximate Analytic Models ......................................... 20
       - Discrete Probability Methods ................................. 22
       - Monte Carlo ............................................................... 22
       - Response Surface Method ........................................... 27
       - Latin Hypercube Sampling (LHS) Method ..................... 29
       - Markov Chain Models .................................................. 30
       - Discrete Probability Distributions (DPDs) .................... 32

3. THE DEVELOPMENT OF THE RANDOM SAMPLING CONDENSATION ALGORITHM ...................................................... 37

   Summary .................................................................................. 45
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. FAULT TREE QUANTIFICATION AND UNCERTAINTY ANALYSIS</td>
<td>46</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>46</td>
</tr>
<tr>
<td>4.2 Fault Tree Models</td>
<td>47</td>
</tr>
<tr>
<td>4.3 Fault Tree Quantification Results</td>
<td>52</td>
</tr>
<tr>
<td>4.4 Estimating Upper Tail Probabilities</td>
<td>58</td>
</tr>
<tr>
<td>4.5 Summary</td>
<td>65</td>
</tr>
<tr>
<td>5. UNCERTAINTY ANALYSIS OF NUCLEAR WASTE CANISTER CORROSION</td>
<td>67</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>67</td>
</tr>
<tr>
<td>5.2 Corrosion Model for Waste Canister Evaluation</td>
<td>68</td>
</tr>
<tr>
<td>5.3 Input Variable Definition</td>
<td>69</td>
</tr>
<tr>
<td>5.4 Time Dependent Uncertainty Analysis Results</td>
<td>72</td>
</tr>
<tr>
<td>5.5 Uncertainty Analysis Performance Evaluation</td>
<td>81</td>
</tr>
<tr>
<td>Percentile Measure of Uncertainty and Confidence Levels</td>
<td>82</td>
</tr>
<tr>
<td>5.6 Probabilistic and Sensitivity Analysis</td>
<td>85</td>
</tr>
<tr>
<td>5.7 Summary</td>
<td>87</td>
</tr>
<tr>
<td>6. PROBABILISTIC FRACTURE MECHANICS ANALYSIS FOR NUCLEAR PIPING MATERIAL</td>
<td>90</td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>90</td>
</tr>
<tr>
<td>6.2 Fracture Mechanics Model</td>
<td>91</td>
</tr>
<tr>
<td>6.3 Random Variable Definitions</td>
<td>95</td>
</tr>
<tr>
<td>6.4 PFM Analysis Results</td>
<td>98</td>
</tr>
<tr>
<td>7. THE OPTIMIZATION OF THE RASCAL ALGORITHM</td>
<td>104</td>
</tr>
<tr>
<td>7.1 Introduction</td>
<td>104</td>
</tr>
<tr>
<td>7.2 RASCAL Description</td>
<td>104</td>
</tr>
<tr>
<td>7.3 Optimization Algorithm</td>
<td>109</td>
</tr>
<tr>
<td>7.4 Summary</td>
<td>114</td>
</tr>
<tr>
<td>8. SUMMARY AND RECOMMENDATIONS FOR FURTHER RESEARCH</td>
<td>117</td>
</tr>
</tbody>
</table>

APPENDICES

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. DATA ANALYSIS AND STATISTICAL METHODS.</td>
<td>120</td>
</tr>
<tr>
<td>Statistical Analysis</td>
<td>120</td>
</tr>
<tr>
<td>Fitting Distributions to Data</td>
<td>126</td>
</tr>
<tr>
<td>B. THE RASCAL ALGORITHM: A USER'S ALGORITHM MANUAL.</td>
<td>131</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Distribution Definitions</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Random Sequencing for LHS Analysis</td>
<td>30</td>
</tr>
<tr>
<td>3.1 Numerical Simulation of Choosing Balls from a Jar</td>
<td>39</td>
</tr>
<tr>
<td>3.2 Sample Calculation for Random DPD Method</td>
<td>41</td>
</tr>
<tr>
<td>4.1 Fault Tree Definition for Study I</td>
<td>51</td>
</tr>
<tr>
<td>4.2 Fault Tree Definition for Study II</td>
<td>52</td>
</tr>
<tr>
<td>4.3 Mean and Standard Deviation for the Top Event Probability for Fault Tree Study I</td>
<td>54</td>
</tr>
<tr>
<td>4.4 Comparison of Percentile Calculation by RASCAL, LHS, and MOMOD Methods</td>
<td>57</td>
</tr>
<tr>
<td>4.5 Mean and Standard Deviation for the Top Event Probability for Fault Tree Study II</td>
<td>60</td>
</tr>
<tr>
<td>4.6 Probability for Each of 20 Intervals of Individual DPDs for Fault Tree Study I</td>
<td>62</td>
</tr>
<tr>
<td>4.7 Probability of Masses for Each of 50 Intervals of Individual DPDs for Fault Tree Study I</td>
<td>65</td>
</tr>
<tr>
<td>4.8 Comparison of Upper Tails Calculation by Monte Carlo and RASCAL for Fault Tree Study I</td>
<td>65</td>
</tr>
<tr>
<td>5.1 Corrosion Model Input Data Description</td>
<td>71</td>
</tr>
<tr>
<td>5.2 Discretization Used for the RASCAL and LHS Analysis</td>
<td>75</td>
</tr>
<tr>
<td>5.3 Mean Corrosion Depth as Predicted by Monte Carlo Simulation</td>
<td>77</td>
</tr>
<tr>
<td>5.4 Mean Corrosion Depth as Calculated by the RASCAL Method</td>
<td>79</td>
</tr>
<tr>
<td>5.5 Mean Corrosion Depth as Calculated by the LHS Method</td>
<td>79</td>
</tr>
<tr>
<td>5.6 Moments of the Distribution of Corrosion Depth at Year 1,000 by Monte Carlo Simulation</td>
<td>80</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Generic PDF Representation and Some Parameters</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Uniform Distribution</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Normal Distribution</td>
<td>12</td>
</tr>
<tr>
<td>2.4</td>
<td>Lognormal Distribution</td>
<td>14</td>
</tr>
<tr>
<td>2.5</td>
<td>Rayleigh Distribution</td>
<td>14</td>
</tr>
<tr>
<td>4.1</td>
<td>Sample Fault Tree</td>
<td>49</td>
</tr>
<tr>
<td>4.2</td>
<td>Small Fault Tree Analysis - MOMOD and RASCAL</td>
<td>55</td>
</tr>
<tr>
<td>4.3</td>
<td>Small Fault Tree Analysis - LHS and RASCAL</td>
<td>56</td>
</tr>
<tr>
<td>4.4</td>
<td>Emergency Feedwater System Fault Tree Analysis</td>
<td>59</td>
</tr>
<tr>
<td>4.5</td>
<td>Unequal Probability Interval Comparison</td>
<td>63</td>
</tr>
<tr>
<td>4.6</td>
<td>Upper Tail Calculation Using Unequal Probability Intervals</td>
<td>64</td>
</tr>
<tr>
<td>5.1</td>
<td>Mean Corrosion Depth As a Function of Time By Monte Carlo Analysis</td>
<td>73</td>
</tr>
<tr>
<td>5.2</td>
<td>Mean Corrosion Depth Calculated By Each Uncertainty Analysis Method</td>
<td>76</td>
</tr>
<tr>
<td>5.3</td>
<td>Standard Deviation Predicted By Each Uncertainty Analysis Method</td>
<td>89</td>
</tr>
<tr>
<td>6.1</td>
<td>LEFM Theory for Crack in an Infinite Plate</td>
<td>92</td>
</tr>
<tr>
<td>6.2</td>
<td>Coordinate System for Stress Calculation</td>
<td>92</td>
</tr>
<tr>
<td>6.3</td>
<td>Stability Region for Crack Growth</td>
<td>96</td>
</tr>
<tr>
<td>6.4</td>
<td>Mean Crack Size Versus Time as Predicted by Monte Carlo and RASCAL</td>
<td>101</td>
</tr>
<tr>
<td>6.5</td>
<td>Standard Deviation of Crack Size Versus Time as Predicted by Monte Carlo and RASCAL</td>
<td>102</td>
</tr>
<tr>
<td>6.6</td>
<td>Empirical CDF of Crack Size</td>
<td>103</td>
</tr>
</tbody>
</table>
7.1 Five Point Discretization of Variable X .................. 106
7.2 Twenty Point Discretization of Variable X ............... 106
7.3 Comparison of Interval Discretization Schemes .......... 111
7.4 Standard Deviation of Crack Size Using Optimized RASCAL Method ....................................................... 116
A.1 Regions of Parametric Families ......................... 130
B.1 RASCAL Main Program Flowchart ...................... 134
The design and construction of any engineered systems requires that sources of uncertain behavior be included in the design. Whether these uncertainties are due to true stochastic processes or the engineer's lack of knowledge about a specific process or subsystem, they must be accounted for to insure safe construction, operation, and retirement. Historically, these uncertainties have been handled by safety factors. This means that if the design called for a beam to be able to carry a load of $X$ then the material strength must actually be equal to or greater than $X$ multiplied by the safety factor so that some accounting of the uncertainty is made. A safety factor of 3.0 is still used commonly today in many nuclear applications presumably, whether it is actually true or is myth, to correspond to three standard deviations.

Of course, this is the most simplistic approach and even after safety factors are applied, the entire design is rechecked. To illustrate how variables which are stochastic and therefore can credibly assume a wide range of values can affect the design life, consider the following example. Suppose a system which has three sources of loading applied to it is to be designed so that the maximum earthquake load expected over a 1,000 year period is to never be exceeded. If the sum of the 90th percentile values of each of the individual loads is greater than or equal to the 1,000 year earthquake, then there is a 32% chance that
the system will see the 1,000 year earthquake load before it is ever subjected to the earthquake.

Even if one becomes more sophisticated from a probabilistic viewpoint, errors can be made without a probabilistic analysis. For example, consider the fatigue of metal. In this case, the critical variable is the number of cycles of fatigue that the metal is subjected to at a given stress level. Assume that the metal being studied is subjected to two load sources which are both uniformly distributed between zero and one. Further, assume that superposition holds at the point of interest in the metal. The first instinct is that the composite load is uniformly distributed between zero and two (because of superposition). However, a simple probabilistic analysis will demonstrate that the composite load has a triangular distribution with the value of one being more likely to occur than any other value. This has serious implications for the fatigue analysis particularly if there is a threshold stress level below which no damage (due to fatigue) accumulates.

The purpose of these examples is to demonstrate that it is not always adequate to perform deterministic, engineering, "best-estimate" analysis. In fact, a probabilistic analysis may demonstrate that the engineering best-estimate is not the statistical best-estimate, i.e., the mean or modal value. The correct inclusion of uncertainties in an analysis can produce useful information on mean responses, most frequent responses, and credible bounds for responses of a system or process. This information can be used in decision making to maximize safety within the cost constraints of the project.

The most reasonable question to now ask is—why isn't probabilistic analysis always performed? The answer is two-fold: First, it is generally believed that the lack of data introduces more uncertainty into the problem than it illuminates and explains;
and, secondly, probabilistic analysis can be extremely costly. The first problem is generic to all analyses, and are problems in either deterministic, or probabilistic analysis. The second problem, while it is generic to probabilistic analyses alone, can be attacked and solved. The goal of this dissertation is the development and benchmarking of a new probabilitistic method which is as accurate as Monte Carlo yet much less costly. This method, hereafter denoted RASCAL, will be benchmarked and compared to other techniques currently in use in the nuclear field.

In the field of nuclear engineering, probabilistic analysis began with the Reactor Safety Study [1] (RSS). The methods used in the RSS were crude, basic techniques with almost no uncertainty analysis having been performed. Since the RSS was published, almost 10 years ago, significant work and progress has been made in the probabilistic analysis area, not only in the nuclear engineering area, but also the aerospace, petrochemical, oil, gas, and metals industries. The purpose of this study is to compare the three major methods used in the nuclear industry for the probabilistic analysis of three important problems in the nuclear field. The problems chosen for analysis are (1) determination of the corrosion depth of a nuclear waste canister; (2) determination of the size of the material defect in nuclear piping material; and (3) an uncertainty analysis of the quantification of the fault tree representation of a nuclear power plant safety system. These problems were chosen not only because they have important safety implications but also because there is available data or studies to judge these results against.

The three methods of analysis include Monte Carlo[2], Latin Hypercube Sampling[3] (LHS), and the random sampling condensation algorithm (RASCAL) developed during this study. The RASCAL method is an original probabilistic technique and, therefore, one of the goals of this study is to assess its performance against the standardly used Monte Carlo and LHS methods. Because
the motivation for the RASCAL method grew out of an examination of Discrete Probability Distribution (DPD) methods, these are also discussed to provide the motivation from which RASCAL grew, even though the RASCAL method is not a DPD algorithm.

The remainder of the dissertation is organized as follows. First, the basic literature review and the rudimentary theoretical concepts to be used are presented. The next chapter gives the development and description of the RASCAL. The following three chapters give the results of each case study. Next, an optimization procedure for the RASCAL method is derived which is based on the results of the case studies. Finally, a summary and recommendations for further research are given.
CHAPTER 2

PROBABILISTIC ANALYSIS METHODS

2.1 Introduction

There are many uses for probabilistic analysis methods in the design, development, and operation of engineered components or structures. During the design phase, it is important to determine the range of possible responses of individual components so that the design can be made in a safe manner to ensure that the system will not operate in a region where infrequent sets of conditions will cause its failure. During the development phase, it is necessary to select one of several design options. In this case, one of the criteria upon which the selection will be made is the safety, i.e., lack of risk, of the design. This criterion is normally based on some type of risk analysis which requires a probabilistic treatment to determine the failure probabilities. Finally, during the operational phase, it is necessary in many cases to calculate the remaining expected life of a component or system. Because many design variables are random, e.g., the magnitude and timing of loads applied to a structure, probabilistic methods are the only tools available to calculate the expected, as opposed to the engineered, life of the component.

In this chapter, some motivation for the present studies will be given. In order to identify the significance of these developments and applications, a survey of the literature is given. Because of the rapidly expanding nature of this field, the survey
does not attempt to be comprehensive but rather focuses on significant methods which are popular because they are currently accepted and/or used in the nuclear industry.

2.2 Statistical and Probabilistic Analysis

Throughout the literature, statistical and probabilistic analysis are used almost interchangeably. However, in order to provide clarity for the remaining discussion, it is important to define the type of analysis and techniques used in each of these overlapping, but distinct, areas.

Simply stated probabilistic analysis is concerned with the generation of a sample space which describes a random variable or model while statistical analysis is concerned with the analysis and characterization of this sample space. For example, if a variable is characterized by a probability density function (described below) then no probabilistic analysis can be performed since the sample space is defined by this function. However, parameters of this distribution, such as the mean, variance, and so forth, can be determined using statistical analysis. On the other hand, if two random variables are to be combined to find an output distribution of interest then the determination of the output distribution (i.e., the sample space of the output random variable) is performed using probabilistic analysis. The resultant distribution can then be characterized using statistical analysis. Given this basic definition, more mathematically oriented concepts are defined in the following paragraphs in order to provide clarity and common definitions when the survey of other probabilistic analysis work is given.
2.3 Statistical Formulas

Probability Density Function (PDF). The PDF of a random variable will be fundamental to all of the work performed during this study. The PDF, denoted \( f \), of a random variable, denoted \( X \), is given by [4]

\[
\sum_{i=1}^{N} f_i(X_i) = 1 
\]  

(2-1a)

for \( X \) a discrete variable whereas if \( X \) is continuous, then

\[
\int_{-\infty}^{\infty} f(x)dx = 1
\]

(2-1b)

This is the most general formulation for the PDF although initially it does not appear to give much information. Qualitatively, the PDF of a random variable \( X \) is any function such that the sum, if \( X \) is discrete, or integral overall possible values of \( X \), if \( X \) is continuous, is equal to one. For example, if

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp(- (x - \mu)^2 / 2\sigma^2)
\]

then this is a PDF for a random variable \( x \). (Later sections discuss this function in more detail.)

Cumulative Distribution Function (CDF). The CDF of a random variable \( X \) gives the probability that the value of the random variable is less than or equal to a specified value. If the CDF is represented by \( F(x) \), then mathematically,

\[
F(x) = \int_{-\infty}^{x} f(z)dz
\]

(2-2a)
where \( f(z) \) is the PDF of the random variable \( X \). If \( X \) is discrete, then

\[
F(x_m) = \sum_{i=1}^{m} f(x_i) \tag{2-2b}
\]

Many times in the literature, the complementary cumulative distribution function (CCDF) is used\[^{[6]}\]. The CCDF, denoted by \( C(x) \), is equal to

\[
C(x) = 1.0 - F(x) \tag{2-3}
\]

Qualitatively, then the CCDF represents the probability that \( X \) is greater than \( x \). The CDF will be used exclusively in this study since it is the more common probabilistic function.

**Moment Calculations.** For all of the probabilistic studies to be performed, it will be necessary to determine at least the first moment of distributions either obtained from experiments or computer simulations. If there are \( N \) discrete points representing a random variable \( X \), then the \( k \)th moment is given by

\[
M_k = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^k \tag{2-4a}
\]

where

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{2-4b}
\]

If \( X \) is a continuous variable then

\[
M_k = \int_{-\infty}^{\infty} (x - \mu)^k f(x) \, dx \tag{2-4c}
\]
where

$$u = \int_{-\infty}^{\infty} xf(x)dx$$  \hspace{1cm} (2-4d)

and $f(x)$ is the PDF of $X$.

**Coefficients of Skewness and Kurtosis.** In analyzing data representative of a random quantity, it may be necessary to examine the shape of the PDF of this data. A quick check of the shape of the distribution can be made by calculating the coefficient of skewness, denoted $\beta_1$, and the coefficient of kurtosis, denoted $\beta_2$, defined by [7]:

$$\beta_1 = \frac{M_3}{M_2^{3/2}}$$  \hspace{1cm} (2-5a)

$$\beta_2 = \frac{M_4}{M_2^{2}}$$  \hspace{1cm} (2.5b)

where $M_i$ is defined by Equation (2-4). The coefficient of skewness indicates the symmetry of the distribution. For $\beta_1$ equal to zero, the distribution is symmetrical about its mean value while for $\beta_1$ greater than zero, the distribution is skewed right, i.e., the mean value is greater than the modal value (see below). The coefficient of kurtosis (also identified as excess in the literature) measures how peaked the distribution is. For a normal distribution, $\beta_2$ equals three. If $\beta_2$ is less than three, then the distribution is less peaked, or flatter than the normal, and if $\beta_2$ is greater than three, it is more peaked.

**Distribution Types.** The mathematical representation of discrete data for a random parameter obtained from experiments requires that the features of this data be represented by the selected expression. Figure 2.1 presents a plot of a generic PDF in which three parameters are shown: (1) the modal value which
Figure 2.1 Generic PDF and Some Parameters
which is that value which divides the area under the PDF into
two equal parts, and (3) the mean which is the value at which
the center of gravity of the area under the PDF occurs. This
distribution, which is a Rayleigh type, is more meaningful for
physical calculations than, say, the Normal distribution since
it has a lower bound; the Normal distribution bounds are always
between negative infinity and positive infinity although the
probability of very large deviations from the mean occurring
is very small.

With these facts in mind, several distributions are discussed
below which represent those to be used most frequently in this
study.

Uniform Distribution. The uniform distribution, shown in Figure 2.2,
can be used when there are known upper and lower limits for
the random variable yet the probability of the variable assuming
a value between these bounds is equal for any value. In Figure 2.2,
the upper and lower limits, denoted b and a, respectively, together
with the mean value, denoted m, and the standard deviation,
σ, are shown. Also given are the first four moments of the
distribution. The mean is the arithmetic average between the
upper and lower limits. Since the third moment is equal to
zero, the distribution is symmetric.

Normal Distribution. The normal distribution is also identified
as the Gaussian distribution in the literature. The values
of the random variable have no limits, i.e., the random variable
can assume any value between −∞ and +∞. This is a two parameter
distribution defined by its mean value, m, and its standard
deviation, σ. The mean value is the most frequently occurring
value (or the mean equals the modal) because this distribution
is symmetric. The standard deviation is a measure of how "spread
out" the distribution is. For large σ, the distribution is
\[ m = \frac{(a+b)}{2} \]
\[ m_2 = \frac{(b-a)^2}{12} \]
\[ m_3 = 0 \]
\[ m_4 = \frac{(b-a)^4}{80} \]

Figure 2.2 Uniform Distribution

Figure 2.3 Normal Distribution
For the lognormal distribution shown in Figure 2.4, the standard deviation is equal to one and, thus, the mean value is approximately 4.5 times larger than the modal value. As \( \sigma \) becomes larger, the lognormal distribution approaches the uniform while \( \sigma \) approaching zero will give the distribution a more symmetrical shape. The important feature of the lognormal is the "long" upper tails. To illustrate this, Figure 2.4 has been plotted using the standard normal variables of \( \mu \) equal to zero and \( \sigma \) equal to one. In this case, the modal value occurs at the value of the random variable, \( x \), equal to 0.368, the mean equal to 1.649, and the value of \( x \) at which 95% of the area under the PDF is covered is 3.561, and 99% coverage occurs at \( x \) equal to 6.643. For the normal distribution, the value of \( x \) at which 99% of the values are less than or equal to that value divided by the 95% value is 1.2875 while it is 1.865 for the lognormal, i.e., the upper tails are "longer" for the lognormal distribution as opposed to the normal. It is noted that these values are obtained using the CDF of each distribution to obtain the area under the PDF curve via integration.

Rayleigh Distribution. The Rayleigh distribution, shown in Figure 2.5, is a special form of the Weibull distribution [8]. It is a nonsymmetric distribution skewed to the right. It is used in many probabilistic structures applications [9,10] because the upper tails do not introduce the degree of over conservatism inherent in the lognormal since the upper tails are not as long. Additionally, the mean value is much closer to the modal value, for the standard parameters of \( \mu \) equal to zero and \( \sigma \) equal to one.
\[ m = e^{\mu + 1/2\sigma^2} \]
\[ m_2 = m^2(e^{\sigma^2 - 1}) \]
\[ m_3 = m^3(e^{3\sigma^2 - 3\sigma^4 + 2}) \]
\[ m_4 = m^4(e^{6\sigma^2 - 4e^{3\sigma^2} + 6e^{\sigma^2} - 3}) \]

Figure 2.4 Lognormal Distribution

\[ m = \delta \sqrt{\pi/2} \]
\[ m_2 = \delta^2(4\pi) \]
\[ m_3 = \delta^3 \sqrt{\pi/2} (\pi - 3) \]
\[ m_4 = \delta^4(32\pi^2)/4 \]

Figure 2.5 Rayleigh Distribution
The above four distributions are the ones which will be used in the subsequent three chapters for the case studies. The parameters and formulas associated with each distribution are summarized in Table 2.1. A method for fitting distributions to experimental data is presented in Appendix A. The next topic to be covered is methods for combining distributions representing random variables, i.e., probabilistic analysis.

<table>
<thead>
<tr>
<th>Name</th>
<th>Range</th>
<th>Parameters</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>[a, b]</td>
<td>a, b</td>
<td>(1/(b-a))</td>
</tr>
<tr>
<td>Normal</td>
<td>(-(\infty), +(\infty))</td>
<td>(\mu, \sigma)</td>
<td>(\exp\left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right) / \sigma \sqrt{2\pi})</td>
</tr>
<tr>
<td>Lognormal</td>
<td>(0, +(\infty))</td>
<td>(\mu, \sigma)</td>
<td>(\exp\left(-\frac{1}{2} \frac{(\ln x - \mu)^2}{\sigma}\right) / \sigma x \sqrt{2\pi})</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>(0, +(\infty))</td>
<td>(\sigma)</td>
<td>(\exp\left(-\frac{1}{2} \left(\frac{x}{\sigma}\right)^2\right) x / \sigma^2)</td>
</tr>
</tbody>
</table>

2.4 Methods For Probabilistic Analysis

There are three primary reasons for performing a probabilistic analysis. First, one wishes to determine the probability that a critical value is exceeded. For example, if a crack in a structure reaches a certain length, say a\(^6\), then it is known the structure will fail. Because there are several different
defect sizes and the structure will be subjected to random loadings, there is some probability that the current crack size exceeds \( a^q \). This is a topic properly addressed by probabilistic analysis and will be investigated in Chapter 6. The second topic is the determination of the most probable value of a random variable which changes due to stochastic changes in its environment. An example of this calculation is given in Chapter 5 where the most probable corrosion depth in a nuclear waste canister is calculated. Of course, since one recognizes the variability in the process being analyzed, in this case metal corrosion, it is not adequate to simply present the most probable value but rather one must include the third topic appropriately addressed by probabilistic analysis: uncertainty analysis. This topic is also studied in Chapter 4. The fourth chapter addresses these topics for fault tree quantification which is a time independent problem as opposed to the crack growth and corrosion analysis which are time dependent.

The following discussion presents various methods for performing probabilistic analysis which have been used in previous studies. Afterwards, the three methods to be examined in this study are presented in detail.

The types of probabilistic methods currently used for risk analysis type studies can be categorized according to the detail of the output which can be obtained from the analysis. The greatest level of detail is when the entire PDF of the random variable of interest is determined\(^1\). The next level of detail is provided when either several moments or percentile values are calculated.

---

\(^1\)Obviously, for PDF's whose range is infinite, the entire function cannot be calculated. In this context, the word entire means between very small, e.g., the 0.1 percentile, and very large, e.g., the 99.9 percentile.
Finally, the least detail is provided when credible upper and/or lower bounds are determined.

Examining methods which determine credible bounds to the value of the random variable limits one almost exclusively to sensitivity or uncertainty analysis methods. A good comparative survey of such methods are presented in Cox and Baybutt [11]. Two popular approaches exist for obtaining the estimates of the bounds.

First are moment propagation techniques which propagate selected moments, usually the second, of the random input variables to obtain an estimate of the corresponding moment of the resultant random variable [12,13]. The major drawback to such methods is that they only work when the functional relationship is well-behaved and smooth and when the input variables are defined by a limited set of possible distributional families, usually limited to normal and lognormal distribution types. Modifications to these schemes have been devised so that non-normal distributions, e.g., Weibull or extreme value, are transformed to a standard normal distribution and then moments or percentile values are determined from this transformed distribution. Examples include the Rackwitz and Fiessler scheme [14] and the Chin-Lind [15] algorithm. These methods have been applied to a structural integrity problem by Wirsching [16] but still suffer from the drawback that the functional relationship must be analytic.

The second method includes response surface techniques. These methods are described in Myers [17] but the basic premise is the replacement of the true functional relationship between the inputs and the output, denoted the response, by a simple analytic polynomial approximation. The variances of the inputs are propagated through this approximation, called the response
surface, to obtain an estimate of the variance of the output. Again, this technique represents a good method for limited distributional forms. It has been applied in several uncertainty analyses by Kurth, Baybutt, and Cox [18]. The response surface generation is performed using one of several sampling methods including fractional factorial designs [19] and Latin hypercube sampling [20] which is described below.

If moment propagation methods are not used, there are methods for calculating the first time a random variable reaches (either falls below a selected level or exceeds the level) a specified value. Included in these methods are barrier problems [21], first crossing methods [22], and up crossing techniques. For example, one may be interested in the probability that the stress at a critical location in a structure exceeds a specified level, many times the elastic limit or the ultimate strength. This stress level is treated as a "barrier" and calculations are made to determine either the expected time at which the barrier is reached (barrier and first crossing) or the probability that the barrier is exceeded (up-crossing). A good review of the mathematics of such techniques is given in reference [23], however, they can only be used to perform bounding calculations and are usually of little use in a design calculation.

Techniques for estimating several moments or percentile values without calculating the entire response PDF are limited. If normal or lognormal distributions are used to represent the input variable PDF's then additional formulas can be derived for propagating higher order moments although it is usually quite tedious [24]. In most cases, a response surface method is used and only approximations to these values are made. One notable exception is the adjoint method for sensitivity analysis [25] in which the adjoint problem is formed and from its solution a great deal of probabilistic information can be obtained.
Its major drawback is the need to solve a problem different from the physical conditions the engineer is interested in solving. If a computer code is being used, this technique requires significant recoding to be applied.

The best method for performing probabilistic analysis is one which determines, either analytically or numerically, the PDF of the response variable. It should be noted that this is mathematically equivalent to determining the CDF of the variable. If one is interested in determining the CDF for a variable, call it $Z$, which is a function of two random variables $X$ and $Y$ then

$$ F(Z) = \int_{-\infty}^{Z} g(f_X(x), f_Y(y)) dx dy \quad (2-6) $$

where $f_X$ and $f_Y$ are the PDF's of the variables $X$ and $Y$. Of course, if more than two variables are involved then more variables must be added to the integration process.

**Probabilistic Models**

A probabilistic model accepts distributions describing random variables and combines these distributions in the context of some rule or law describing a physical process. It provides a measure of some important statistical quantity or quantities as output. Probabilistic models can be divided into two categories according to the type of output each one produces. The first provides a point estimate or bounds for the probability of some event occurring (e.g., probability of failure, probability of a load being exceeded, etc.). The second type of model gives a distribution as the output, from which various statistical parameters can be calculated, such as the probability of failure.
The former type of probabilistic models are denoted approximate analytic models since they usually use some fast probability integration technique or moment matching or propagation method to obtain the probability estimate while the latter is denoted discrete probability methods since the input and/or output of the models are contained in a discrete form. The most popular forms of each type are discussed in the next subsections.

**Approximate Analytic Models.** The basic goal of these types of probabilistic models is to calculate, or at least bound, the probability that a level or barrier is exceeded or that the limit state (a surface in the multidimensional sense) is entered. To examine how various rules or models are used, an example with the individual component load effects are represented by $L_1(t), L_2(t), \ldots, L_N(t)$ is used. Assume that the probability of failure, denoted $P_F$, is desired. The limit state condition can be written as

$$q(R-Q) = R - Q$$

where $R$ is the measure of the structural strength, and $Q$ is the measure of the combined load. The limit state probability $P_F$ can be determined as

$$P_F = P(R < Q) = \int F_R(x)f_Q(x)dx = \int [1 - F_Q(x)]f_R(x)dx$$

provided the cumulative distribution functions (CDF) $F_R(x)$ and $F_Q(x)$, respectively, of $R$ and $Q$ and the corresponding probability density functions (PDF) $f_R(x)$ and $f_Q(x)$ are obtained. It should be noted that if more than one of the component loads $L_i(t)$ must be regarded as a random (stochastic) process rather than a random variable, then the calculation of $F_Q(x)$ becomes difficult. Therefore, for this discussion, assume that only one of the $L_i(t)$, say $L_N(t)$, is truly a random process, while all others.
are random variables. Then finding the distribution of $F_{QT}(x)$ is equivalent to either the problem of evaluating the probability that the random process $Y_N(t)$ will exceed a specified level $x > 0$ in a given time interval $[0, T]$ or to the problem of deriving the CDF $F_T(t)$ of the time $T$ at which the process crosses the level for the first time and from below. The first problem is referred to as a level crossing problem while the latter is a first passage time problem. Clearly,

$$F_{QT}(t) = 1 - F_T(T).$$

Therefore, once the CDF of $Q_T$ is found, the limit state probability, i.e., the probability of failure, can be determined.

Unfortunately, no exact solutions are available for first excursion probabilities such as those outlined above, except in simple cases. Shinozuka [12] and Shinozuka and Yao [13] were among the first to study these types of problems in the context of structural safety and reliability. The results of their early work were techniques for bounding the probability of failure. However, all of this work deals with linear loads where $Q$ is the sum of the individual loads. For nonlinear load problems, fast integration techniques may be used for estimating limit state probabilities. This concept is based on a generalized safety index introduced by Hansofer and Lind [25], which only deals with normally distributed random variables, was modified by Rackwitz and Fiessler [14] by transforming each variable to equivalent normal variables. Wu and Wirsching [26] further extended the algorithm by approximating the limit state as a polynomial in the neighborhood of the design point. Additional accuracy can be obtained through the use of response surface methods [17] to obtain the polynomial approximation. These methods are discussed below.
In addition to these reliability-based techniques, simple rules, i.e., models, have been developed to either estimate limit state probabilities or define the limit state surface. Some of these include linear combination rule [27]; load reduction factor method in which the design combined load is equal to the maximum among the individual loads and their sum multiplied by a reduction factor; and companion action factor load [28]. Such rules are adequate for providing bounds for limit state probabilities when all the variables can be assumed to be independent. However, when dependencies exist among the loads, these rules must be modified, otherwise nonconservative results will be obtained. For example, earthquake loads may be highly correlated with aerodynamic (wind) and thermal (fire) loads when an analysis of a marine structure is performed. Therefore, these rules or models are not applicable unless modifications can be derived to account for such dependencies.\(^2\)

**Discrete Probability Methods.** The discrete probability methods are characterized by their ability to estimate a distribution as output from a model which uses distributions as input. While several models exist, the most important are (1) Monte Carlo, (2) Monte Carlo with importance sampling, (3) Markov chain, and (4) Discrete Probability Distribution (DPD). The Monte Carlo method has been described previously; however, the basic method is reviewed in order to adequately describe importance sampling. To provide a clear example, the discussion is made in the context of nonlinear loads.

**Monte Carlo.** The Monte Carlo technique is a simple method for adding a probabilistic structure to a deterministic model. Suppose a composite load, \(L_{c}(t)\) is related to individual loads \(L_{i}(t)\), by a function

\(^2\)This is a problem given to all probabilistic methods.
where the function \( f \) may not even be analytic, e.g., a computer program. If each of the individual loads has been characterized by a PDF during data analysis, then the following procedure is used during a Monte Carlo simulation. The CDF of each individual load is generated by integrating the PDF. A random number between 0 and 1 is generated, call it \( r_1 \). The CDF is inverted and the value for the load, \( L_j(1) \), is determined. This method of choosing the value is repeated for each individual load. A value of \( L_c \) is then calculated as

\[
L_c(1) = f(L_1(1), L_2(1), \ldots, L_n(1))
\]  

The entire process is repeated a large number of times, say \( M \). What results is an \( M \)-dimensional vector of composite loads: \((L_c(1), L_c(2), \ldots, L_c(M))\). This vector is used to construct a histogram of the composite load. This histogram can now be analyzed statistically to obtain estimates of the mean, kurtosis, probability of a load being exceeded and so on. Obviously, in the limit as \( M \) tends to infinity, the continuous distribution will be asymptotically approached. Equally obvious, the computer time will also increase. As an example, let \( N(m,s^2) \) represent the normal distribution with mean \( m \) and variance \( s^2 \). Consider estimating the probability that a random variable described by the standard normal distribution, \( N(0,1) \), is greater than 3 using Monte Carlo. It is known analytically that this probability is \( 1.34 \times 10^{-3} \). The variance of the Monte Carlo estimator based on \( M \) trials is

\[
S^2/M = P(1 - P)/M
\]

where \( P \) is the probability that the random variable exceeds 3. Thus, a relative standard error of 10\% implies

\[
S/P \sqrt{M} = 0.1
\]

where \( P \) is the probability that the random variable exceed 3.
Thus, a relative standard error of 10% implies

\[ \frac{S}{P} \sqrt{M} = 0.1 \]

or

\[ \frac{(1 - P)}{PM} = 0.1 \]

Thus, \( M \) is approximately equal to 75,000. The large number of Monte Carlo samples needed to estimate \( P \) is due to its small value: when one samples from a normal density, values greater than 3 are so rare that many trials are needed to get enough data to estimate \( P \) accurately. These problems and inefficiencies of the Monte Carlo method have led to modified sampling methods described next.

Suppose the mean \( m \) of a function \( f(x_1, \ldots, x_n) \) of the random variables \( x_1, \ldots, x_n \) is to be estimated. Let \( p(x_1, \ldots, x_n) \) be the joint PDF. Then, \( M - E_p f(x_1, \ldots, x_n) = f(x_1, \ldots, x_n)p(x_1, \ldots, x_n)dx_1 \ldots dx_n = fp \), where \( E_p( - ) \) denotes expected value taken with respect to the density \( p \). The basic Monte Carlo method involves sampling repeatedly from the joint PDF and estimating \( m \) by the usual sample average of the values of \( f \). The idea of importance sampling \([\ldots]\) is to sample not from \( p \) but from a different PDF, \( q \), chosen to decrease the variance of the Monte Carlo estimator. The values of \( f \) must be appropriately weighted to produce an unbiased estimator of \( m \). Specifically, when sampling from \( q \), we have

\[ E_q f(x_1, \ldots, x_n) = fq \neq m \]

Instead,

\[ E_q [f(x_1, \ldots, x_n)/q(x_1, \ldots, x_n)] = q(fp/q) = fp - m \]
Thus, the appropriate weight for a sample point \((x_1, \ldots, x_n)\) is the ratio of the densities \(P(x_1, \ldots, x_n)/q(x_1, \ldots, x_n)\). Hence, \(m\) is now estimated by the sample average of the values of \(fp/q\), rather than \(f\).

The variance of the estimator \(fp/q\) is

\[
S^2 = \left(\frac{fp}{q}\right)^2 - m^2
\]

\[
= \frac{f^2 p^2}{q^2} = m^2
\]

If the importance sampling density is chosen as \(q = fp/m\) (assuming \(f \geq 0\)), then \(S^2 = m \frac{fp}{q} - m^2 = m^2 - m^2 = 0\). This is also obvious from \(fp/q \equiv m\) (i.e., the estimator is identically equal to \(m\) for all values of \((x_1, \ldots, x_n)\)) in this case. Of course, this is a nonsensical choice since it requires knowing the value of \(m\). However, it points the way toward choosing \(q\): the shape of \(q\) should be similar to that of \(fp\), the product of the function \(f\) and the original density \(p\). When \(n \geq 2\), of course, achieving this is not easy and pragmatic choices usually must be made.

Consider the use of normal importance densities in the previous example. It makes sense to sample from a normal density which increases the frequency of observed values greater than 3. Suppose we use \(N(a,1)\). In the notation used above, we have only one variable and therefore \(n = 1\) and

\[
f(x) = 1 \quad , \text{if } x \geq 3
\]

\[
0 \quad , \text{if } x \leq 3
\]

\[
p(x) = (\sigma^2 \pi)^{-1} \exp(-x^2/2)
\]

\[
q(x) = (\sigma^2 \pi)^{-1} \exp(-(x-a)^2/2)
\]
Hence, the weighted estimator is \( \frac{f(x)p(x)}{q(x)} = \exp(-ax + a^2/2) \), if \( x > 3 \)

\[
0, \quad \text{if} \quad x \leq 3.
\]

The variance of this estimator can be calculated as

\[
S_1^2 = \exp(a^2)[1 - F(a + 3)] - p^2
\]

\( S_1 \) = importance sampling variance,

where \( F \) is the standard normal CDF. For \( a \geq 0 \), we have

\[
1 - F(a + 3) = ((a + 3) 2\pi)^{-1} \exp(-(a + 3)^2/3) .
\]

The approximate minimum value of \( S_2^2 \) can now be found by differentiation to be \( a = \sqrt{10} = 3.16 \). Values of \( S_2 \) for various values of \( a \) are

<table>
<thead>
<tr>
<th>( a )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_2 )</td>
<td>3.65 x 10^{-2}</td>
<td>9.18 x 10^{-3}</td>
<td>3.72 x 10^{-3}</td>
<td>2.49 x 10^{-3}</td>
</tr>
</tbody>
</table>

The sample size needed to achieve 10% relative standard deviation for various values of \( a \) (various importance sampling densities)
can now be found. The results are

<table>
<thead>
<tr>
<th>a</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>10</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size:</td>
<td>74,700</td>
<td>4,710</td>
<td>775</td>
<td>347</td>
<td>341</td>
<td>536</td>
<td>2,400</td>
<td>27,100</td>
</tr>
</tbody>
</table>

needed to achieve 10%
C.O.V.

The optimal choice reduces the sample size by a factor of \(\frac{74,700}{341} = 219\). However, very substantial reductions are obtained by intuitive choices; e.g., \(a = 3\) is an obvious choice since it produces 50% values above 3 and 50% below, a reasonable balance.

For this simple example, we find that importance sampling can have a dramatic reduction on the number of Monte Carlo trials needed to accurately predict quantities of interest. However, it must be clearly stated that when the number of random variables is greater than 2, the choice of the weighting density is non-trivial. Depending on the number of variables, the effort required to derive an important sampling scheme may not be justified for the predicted reduction in Monte Carlo trials.

Even for those cases in which importance sampling can be used to achieve a reduction in the number of Monte Carlo runs required if the evaluation of the function \(f\) in Equation (2-7) is either very costly or time consuming then the Monte Carlo simulation of \(f\) may not be possible. To deal with these cases, response surface methods can be used to derive a statistically based approximation to \(f\) which is analytic, easy to evaluate, and, for special distributional forms, can be used in moment propagation methods. Response surface techniques are described next.

**Response Surface Method.** The response surface method was not identified as one of the discrete probability methods since, strictly speaking, it is not a probabilistic method but rather
is a statistically based tool for obtaining a polynomial approximation to a complex function. Returning to the example in the previous subsection, if the applied (composite) load is given by

\[ L_c = f(L_1, L_2, \ldots, L_n) \]

where the time dependency has been removed as discussed previously then a polynomial approximation to \( f \), denoted a response surface, is sought in the form

\[ L_c = a_0 + \sum_{i=1}^{n} a_i L_i + \sum_{i<j} a_{ij} L_i L_j \quad (2-8) \]

For the discussion, quadratic and higher order terms have been assumed to be negligible although they can be included if desired.\(^3\) For the determination of the applied load, this is a reasonable approximation since a quadratic term, such as \( L_k^2 \), would imply that load \( L_k \) interacting with itself would have a significant effect on \( L_c \), the response. This is not expected. In Equation (9), the first term is referred to as the mean effect, the second term as the primary effects, and the last term as the interaction effects. In order to determine the coefficients in Equation (9), statistical design methods are used. Examples of statistical designs include Latin square [29] and factorial [30] designs. Statistical designs provide the basis for selecting the optimum number of evaluations of \( f \) in order to accurately calculate the response surface coefficients. If the Taylor series expansion of \( f \) in Equation (2-7) is made then it is immediately seen that the coefficients \( a_i \) correspond to \( \partial L_c / \partial L_i \) and \( a_{ij} \) to \( \partial^2 L_c / \partial L_i \partial L_j \) and are thus direct measures of the sensitivity of \( L_c \) to

---

\(^3\)A complete treatment of response surfaces is given in Reference (17).
the individual loads. For normalized loads,

\[ P_i = \frac{(L_i - m_i)}{\sigma_i} \]

where \( m_i \) is the mean value and \( \sigma_i \) is the standard deviation. The variance of \( L_c \) can be obtained by

\[ \text{Var}(L_c) = \sum_{i=1}^{n} a_i^2 \text{Var}(P_i) + \sum_{i=1}^{n} a_i^2 \text{Var}(P_i) \text{Var}(P_j) \]

where

\[ \text{Var}(X) = \text{variance of variable X} \]

Finally, if Equation (2-8) provides an accurate approximation to the function \( f \), it can be used in a Monte Carlo simulation in place of \( f \).

While statistical designs can be used to define response surfaces, they can also be used directly to estimate statistical quantities. Since one of these methods, Latin hypercube sampling, is used in these studies, it is described below.

**Latin Hypercube Sampling (LHS) Method.** The basic premise of LHS is to insure that all portions of the sample space of the input variables \( X_j \) are sampled. Each \( X_j \) is divided into \( K \) strata of equal marginal probability \( 1/K \), denoted by \( X_{jk} \) where \( k = 1, 2, \ldots, K \). For each stratum, one value of \( X_{jk} \) is obtained by sampling in this interval. LHS then matches the \( X_{jk} \) randomly to use as input to the function being evaluated to obtain values for the response. For example, assume that there are three variables, each divided into equal probability intervals, or strata, as shown in Table 2.2. For this example, the first response,
denoted \( R(1) \), is obtained by sampling stratum 1 of variable 1, stratum 3 of variable 2, and stratum 2 of variable 3, i.e., \( X_{11}, X_{23}, X_{32} \), which are then input to the function \( F \) to obtain \( R(1) \). Strata 3, 2, and 1 are used to obtain the inputs to \( F \) for variables 1, 2, and 3, respectively, to calculate \( R(2) \). Finally, strata 2, 1, and 3 are used to calculate \( R(3) \). This method of sampling is an extension of quota sampling [31]. Additionally, it can be viewed as a \( K \)-dimensional extension of Latin square sampling [32], from which it obtains its name.

Markov Chain Models. In discussing the Markov model, some of the conventions put forth by Bogdanoff [33] are used. In the Markov model, the loads are defined by discrete states with time being measured by duty cycles. In addition, the Markov assumption is made, i.e., the probability that the load which is currently in state \( i \) will be in state \( j \) during the next duty cycle is only dependent on its present state and not on the previous load history. Consider

\[
t = 0, 1, 2, 3, ...
\]

which need not be of equal duration. The load is defined by
variable states

\[ L = 0, 1, 2, \ldots, n \]

where \( L = 0 \) implies no load and \( L = n \) may be defined as the limit state. The initial distribution of loads is defined by

\[ I = (i_1, i_2, \ldots, i_n) \]

by design codes or other data. The transition probabilities, \( P_{ij} \), are defined as the probability that given the load is current in state \( i \), it will next be in state \( j \). The evolution of the load process is given by

\[ P_t = I M^T \]

where

\[ M = \text{matrix of transition probabilities } P_{ij} \]

\[ P_t = (P_t(1), P_t(2), \ldots, P_t(n)), \text{PDF of loads at time } t. \]

Several difficulties arise in using a Markov model. The first of these is the definition of the load states since, ultimately, the number of parameters required to define the load state at the critical components may be significant. Thus, the use of the Markov model may be limited. Second, the calculation of the new PDF for the loads is only simple if the process is stationary. Thus, the previous equation defining the evolution of the loads is constant only if material and thermodynamic properties are constant, and the spectral density of various quantities of interest are independent of time. If this is
not the case, then

\[ P_T = \prod_{t=0}^{T} M(t) \]

where the transition probabilities are now functions of time.

Another area requiring detailed consideration and potential modification is the Markov assumption itself. Under this assumption, the future states of a system are governed entirely by its present state. This difficulty may be addressed in various ways. For example, as a first approximation, the effects on the calculated failure probabilities of the potential inadequacy of the Markov assumption could be addressed in sensitivity and uncertainty analysis. Alternatively, additional parameters can be added to the definition of the load state so that history dependencies are included in the present state of the system. This would restore the Markov nature of the load process. A third possibility is the use of a multiple Markov Chain model. Here, the future states of the system are allowed to depend not only on the present state but on a fixed number of past states. Thus, the process is allowed to have a finite memory.

Discrete Probability Distributions (DPDs). The description of DPDs follows the conventions set forth by Kaplan [31*] and Kurth and Cox [35]. In this method, the initial values of the loads are discretized into \( m \) values. Each value of each variable is then assigned a probability of occurrence. Additionally, the various forms of any probabilistic function are assigned a probability of being correct. If these discrete values are paired with their probabilities, the following vectors of ordered
pairs result for two loads $X$ and $Y$:

$$X = [X_1, p_1), (X_2, p_2), \ldots, (X_m, p_m)]$$

$$Y = [Y_1, q_1), (Y_2, q_2), \ldots, (Y_m, q_m)]$$

The number of discrete points in each of these vectors has been chosen to be the same although it is not necessary to do so. The addition of two discrete vectors is defined by

$$Z = Y + X.$$

$$Z = (Y_i, p_i) + (X_j, q_j),$$

and

$$Z = (X_j + Y_i, P_i^q_j)$$

for all $i$ and $j$.

Therefore, the addition of two vectors containing $m$ ordered pairs each results in a vector which has $m^2$ ordered pairs. The multiplication of DPs is similarly defined:

$$Z = X \cdot Y$$

$$= [X_j \cdot Y_i, p_i^q_j)]$$

for all $i$ and $j$.

For the combination of a large number of loads, the amount of computer storage increases very quickly. If there are $k$ loads, each described by $M$ discrete points, then the vector will contain $M^k$ ordered pairs. Since, even for relatively small values of $M$ and $K$ (on the order of 20), the computer storage capability will quickly be exceeded, it is necessary to examine some procedure for reducing this vector's size. This leads to an examination of the condensation procedure discussed below.

In order to illustrate the condensation procedure, assume that
the initial DPD for two loads contains 20 ordered pairs, respectively. The composite load will then be a vector of 400 ordered pairs after each of the individual load DPDs have been combined. However, it has been assumed that 20 ordered pairs adequately describe the load distribution. Suppose the range of possible loads is divided up into equal intervals. Further, for the sake of example, assume that the new values between 144 and 188 in this composite load vector, denoted $L'$, fall in the 6th interval. Then

$$\hat{p}_6 = \sum_{i=144}^{188} p_i$$

and

$$z_6 = \frac{1}{\hat{p}_6} \sum_{i=144}^{188} p_i z_i$$

where $L' = [(z_i, p_i)]$, $i = 1, 2, \ldots, 400$. This procedure can be written in general as

$$z_i = \frac{1}{\hat{p}_i} \sum_{j} \frac{B_i}{B_j} p_j z_j$$

where

$$B_i = \{ j | d_i < z_j \leq d_i + 1 \} = \frac{(a_{\text{max}} - a_{\text{min}})}{20}$$

$$d_i = a_{\text{min}}, \quad d_i + 1 = d_i + \Delta$$

For many DPD applications, it is not possible to use equally spaced intervals. In fact, logarithmically spaced or other time independent unequal interval spacing schemes may lead to the same problem. A method for calculating time dependent bin sizes
has been devised and is used to condense the DPD at each time step. In this method, after each time step, the largest and smallest values of the vector $L'$ are determined. The intervals are then determined from

$$b_1 = a_{\text{min}}$$

$$b_i = b_{i-1} + 2(i - 1)(a_{\text{max}} - a_{\text{min}})/N(N - 1), \quad i = 2, \ldots, N$$

where

$$a_{\text{min}} = \text{minimum value}$$

$$a_{\text{max}} = \text{maximum value}$$

$$N = \text{number of discrete intervals}$$

$$b_i = \text{interval endpoints for condensation.}$$

For fault tree applications, another, more serious problem has been presented in the literature (see reference [26] for example). Because several terms in the Boolean expression representing the fault tree may be dependent, it is not possible to condense until all of the dependent terms have been combined. As an example of this problem, consider a small fault tree represented by the expression

$$T = X \cdot Y + Y \cdot Z + X \cdot Z$$

where $T$ is the top event probability and $X$, $Y$, and $Z$ are the basic event probabilities. In this case, the first term cannot be condensed until the second term has been calculated since both depend upon $Y$. Similarly, the third term cannot be condensed until the second term is calculated since they both depend on $Z$. Thus, all terms in the equation must be calculated before
any condensation can be performed. In this example, no condensation procedure is able to be used until after the entire set of calculations has been completed! As we shall see in one of the actual examples studied in this dissertation, the situation can be much worse than this simple example presented for discussion. This problem has led to the development of the RASCAL method which has fixed storage requirements and is described in the subsequent chapter.
CHAPTER 3

THE DEVELOPMENT OF THE RANDOM SAMPLING CONDENSATION ALGORITHM

The previous chapter discussed and presented a variety of methods for performing a variety of probabilistic calculations. Limit state methods, barrier crossing methods, and all other similar first order second moment methods provide a quick inexpensive approximation to the probabilistic quantity of interest. If the problem is relatively straightforward and only has normally distributed or lognormally distributed random variables then solutions can be found and used with some confidence. As the problem becomes more and more complex, additional assumptions must be introduced into the analysis in order to make it tractable, and therefore there is an associated loss of accuracy. If additional accuracy is needed then more complex simulation techniques can be employed. These include Monte Carlo and Markov Chain. However, the increased computational costs cannot always be afforded. The DPD method provides a faster algorithm than Monte Carlo but instead of being time limited it is storage limited in that all but the simplest problems quickly require more storage than is available.

While these disadvantages do exist there are also advantages of each technique which we would like to retain. For example, the use of the DPD method for determining the statistical characteristics of a response of interest is appealing because it can easily be put in a data base format, and the logic is independent of the distributional form. It would also be desirable to have a technique that becomes more accurate as more calculations
are performed as in the Monte Carlo method. As an example, consider a response, \( R \), which is simply equal to the sum of all the \( x_i \)'s, each of which has been divided into twenty intervals. Thus,

\[
R = \sum_{i=1}^{10} x_i
\]

and

\[
\text{DPD}[X_1] = [(\hat{x}_{1,1}, p_{1,1}), (\hat{x}_{1,2}, p_{1,2}), \ldots, (\hat{x}_{1,20}, p_{1,20})]
\]

In this case, the number of discrete values which will be generated by the standard DPD algorithm to define \( R \) is equal to \( 20^{10} \). From a statistical standpoint, most of these \( 20^{10} \) points are not meaningful if the PDF for \( R \) is being approximated. The situation is analogous to the problem of determining how many white balls are in a jar of black and white balls. If a jar could be found which is large enough to hold \( 20^{10} \) balls which contains an equal number of black and white balls, one solution is to tip over the jar and count all of the white balls. This is what the DPD method does in an analogous fashion. A much better method is to randomly select a ball from the jar to determine if it is white or black and repeat the process. Table 3.1 presents the results of a numerical simulation in which the true average of white balls is equal to 0.50. As this table shows, even after 500 samples, the predicted average of number of white balls is within 0.8% of the true average. After 100,000 samples, it is within 0.05% of the true value. Obviously, not all of the \( 20^{10} \) balls need to be counted to correctly estimate the
TABLE 3.1 NUMERICAL SIMULATION OF CHOOSING BALLS FROM A JAR

<table>
<thead>
<tr>
<th>Number of Samples</th>
<th>Estimate Number of White Balls</th>
<th>Number of Black Balls</th>
<th>White Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6</td>
<td>4</td>
<td>60.00%</td>
</tr>
<tr>
<td>25</td>
<td>12</td>
<td>13</td>
<td>48.00</td>
</tr>
<tr>
<td>50</td>
<td>24</td>
<td>26</td>
<td>48.00</td>
</tr>
<tr>
<td>100</td>
<td>45</td>
<td>55</td>
<td>45.00</td>
</tr>
<tr>
<td>250</td>
<td>122</td>
<td>128</td>
<td>48.80</td>
</tr>
<tr>
<td>500</td>
<td>252</td>
<td>248</td>
<td>50.40</td>
</tr>
<tr>
<td>1,000</td>
<td>490</td>
<td>510</td>
<td>49.00</td>
</tr>
<tr>
<td>5,000</td>
<td>2,445</td>
<td>2,555</td>
<td>48.90</td>
</tr>
<tr>
<td>10,000</td>
<td>4,977</td>
<td>5,023</td>
<td>49.77</td>
</tr>
<tr>
<td>40,000</td>
<td>20,036</td>
<td>19,964</td>
<td>50.09</td>
</tr>
<tr>
<td>100,000</td>
<td>50,024</td>
<td>49,976</td>
<td>50.024</td>
</tr>
</tbody>
</table>

number of white balls from the sample.¹ This analogy leads to the question: If all of the 20¹⁰ values for R are available, is there not some way to sample from the space of R to estimate the PDF? The answer is, of course, yes; and the sample procedure is explained below.

For the sake of example, assume that equal probability intervals are used for each DPD, i.e., \( P = 1/N_D = p_{ij} \) for all \( i \) and \( j \). Because each value in each variable's DPD is equally likely to occur, a value is selected from an interval or bin for each variable. Based on a random number generator, the following values would be used to generate a particular response, denoted \( R_i \):

¹In this example, it is assumed that there is no bias in the sample, i.e., the jar is well-mixed so that removing a ball from the top of the jar does not affect the result. If all of the black balls are placed in the jar first and then the white balls are placed on top, then this procedure would predict no black balls, however, this sample is then biased and must be sampled differently.
\[ R_1 = x_{1,10} + x_{2,3} + x_{3,4} + x_{4,10} + x_{5,2} + x_{6,17} + x_{7,6} + x_{8,8} + x_{9,18} + x_{10,10} \]

This process is repeated hundreds, or perhaps thousands, of times. The result is an N-dimensional array, denoted \( V \), of the responses:

\[ V = (R_1, R_2, \ldots, R_N) \]

where \( N \) is the number of calculations of \( R \) which have been performed. To further illustrate the procedure, assume that \( N \) is equal to 1,000. Then, this procedure has generated a subset of the DPD space which is approximately \( 10^{10} \) times smaller than the full DPD space! Define as the space of responses generated by taking all possible combinations of the discrete values of the input variables. Then

\[ \lim_{N \to \infty} V_R = \Omega \]

where \( V_R \) means that multiple values of \( R \) are removed from the space \( V \). That is, since the space \( V \) is generated by randomly sampling the input DPDs as \( N \) becomes large, there will be duplicate combinations of the input variables made. Thus, in order to obtain \( \Omega \) by this procedure, many more than \( N_D N_V \) calculations must be made. To illustrate this point, assume that there are two variables, each with two discrete points. The full DPD analysis implies that \( 2^2 = 4 \) calculations are needed. Table 3.2 gives the results of the two calculations. In this example, two extra calculations are made before the DPD space has been reproduced.

Table 3.2 also illustrates another problem with the random sampling method in that the output will almost never be a PDF. As Table 3.2 illustrates, the sum of the probabilities for the random sampling
### Table 3.2 Sample Calculation for Random DPD Method

<table>
<thead>
<tr>
<th>Variable</th>
<th>DPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>(1, 0.5)</td>
</tr>
<tr>
<td>$X_2$</td>
<td>(5, 0.5)</td>
</tr>
</tbody>
</table>

R = $X_1 + X_2$

#### DPD Result

<table>
<thead>
<tr>
<th>Value of R</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.25</td>
</tr>
<tr>
<td>9</td>
<td>0.25</td>
</tr>
<tr>
<td>7</td>
<td>0.25</td>
</tr>
<tr>
<td>10</td>
<td>0.25</td>
</tr>
</tbody>
</table>

#### Random Sampling

<table>
<thead>
<tr>
<th>Value of R</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.25</td>
</tr>
<tr>
<td>7</td>
<td>0.25</td>
</tr>
<tr>
<td>10</td>
<td>0.25</td>
</tr>
<tr>
<td>9</td>
<td>0.25</td>
</tr>
<tr>
<td>10</td>
<td>0.25</td>
</tr>
<tr>
<td>6</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Scheme is 1.5 instead of the value of 1.0 needed for a PDF. To generate the PDF for the response, the probability sum must be normalized to one. To obtain a PDF, it is necessary to examine two cases: one, when the probability sum from the random sampling method is greater than one; two, when the sum is less than one.

When the sum of the probabilities in the array V is greater than one, it is necessary to remove all duplicate ordered pairs from V to obtain $V_R$. If the probability sum is equal to one, then the calculation is finished, otherwise, the procedure described below is used. Because the N samples from $\Omega$ are chosen in a
random manner and are unbiased, the generation of a PDF when the sum of the probabilities in $V$ is less than one is straightforward. If this sum is denoted as $P_s$, then the partial DPD for $V$ is given by

$$V = [(R_1, P_1/P_s), (R_2, P_2/P_s), \ldots, (R_N, P_N/P_s)]$$

where

- $R_i$: value of the response
- $P_i$: probability of $R_i$.

A procedure has been described which will always produce a discrete PDF for a response of interest. However, the DPD of the response will rarely be determined when $N_D^{N_V}$, where $N_D$ is the number of discrete points and $N_V$ is the number of variables, is greater than 10,000. Thus, this method is not, strictly speaking, a DPD method.

The next question that arises is what if the process being modeled is a time dependent process? In such cases, the response may be a function of its previous value and a condensation procedure is needed. If so, the array $V$ is treated just as a DPD, and the condensation procedure previously described is used. Thus, if $N$ is equal to 1,000 and it is desired that the response be represented by 20 discrete points, then

$$P_j = \sum_{i=k}^{l} \frac{P_i}{P_s} \quad k = (j - 1) 50 + 1$$

$$V_j = \frac{1}{P_j} \sum_{i=k}^{l} \frac{P_i}{P_s} R_i$$

For a completely general solution,

$$k = \frac{N}{N_D} (j - 1) + 1$$
Up to this point, it has been assumed that the probability of each variable's discrete value are all equal. If each of the intervals are not of equal probability then what the random sampling scheme produces is an importance sampling method since the random sampling of the discrete intervals will sample from an interval with a lower probability as frequently, on the average, as it does from an interval with a higher weight.

This method, denoted RASCAL (Random Sampling Condensation Algorithm), is summarized as follows. The input variables are discretized into DPD's denoted

$$k = \frac{N}{N_D}$$

where

$$I = 1, 2, \ldots, NV$$

$$J = 1, 2, \ldots, ND$$

$$NV = \text{number of variables}$$

$$ND = \text{number of discrete points}$$

$$X_{I,j} = \text{conditional mean of variable } I \text{ in discrete interval } J$$

$$P_{I,j} = \text{probability of } X_{I,j}$$

$$X_I = \text{input variable } I.$$
denoted $f$. Thus,

$$ R = f(X_1, X_2, \ldots, X_N) $$

At this point, it is decided how many points one wishes to use to characterize the conditional mean of each interval in $R$ after condensation, denote this number by $M$. If

$$ N_D \cdot M > N_D^N $$

then the standard DPD algorithm is used. This is done to minimize the calculational time since, for this case, duplicate values will be produced. If $N_D \cdot M$ is less than $N_D^N$ then $N_v$ random integers between 1 and $N_d$ are generated:

$$ k_1, k_2, \ldots, k_{N_v} $$

A response, $R_1$, is calculated by

$$ R_1 = f(x_1, k_1, x_2, k_2, \ldots, x_N, k_{N_v}) $$

and the associated probability is given by$^2$

$$ q_1 = \prod_{i=1}^{N_v} P_{R_i, k_i} $$

The process is repeated $N_D \cdot M$ times to give an array of ordered pairs of the responses:

$$ V = [(R_1, q_1), (R_2, q_2), \ldots, (R_L, q_L)] $$

$$ L = N_D \cdot M $$

$^2$This calculation assumes that the input variables are independent.
Using the condensation equations from above to condense \( V \) to \( V_R \) thus:

\[
V_R = [(r_1, q_1), (r_2, q_2), \ldots, (r_{ND}, q_{ND})]
\]

The use of RASCAL will be compared to standard methods in the subsequent chapters and its effectiveness for use as an importance sampling scheme will be examined.

**Summary**

This chapter has provided the rationale and development of a new probabilistic technique RASCAL. The purpose of the study is the benchmarking of RASCAL to standard or accepted methods for performing uncertainty and probabilistic analyses. Since it is not feasible to examine all methods, the RASCAL technique will be compared to Monte Carlo and LHS analysis which are currently the most widely accepted and used techniques in the nuclear industry.

Appendix B contains the description of the computer model, its input, and a sample output, of the RASCAL model. The comparisons will be made for three important topics in the nuclear industry: waste canister corrosion modeling; fatigue crack growth; and fault tree quantification. The first two topics are time dependent analyses while the third is a time independent calculation. The results of these analyses are presented in the following three chapters.
4.1 Introduction

The quantification and subsequent uncertainty analysis of fault trees has been the topic of many investigations (see Chapter 2 and references [36] for example) primarily in the nuclear energy and aerospace industries. The goal of a fault tree analysis is to determine all credible ways in which an undesired event may occur. The quantification of the fault tree is performed by assigning probabilities of basic events occurring and then, using Boolean algebra, determine the probability of the undesired event occurring.

This study examines the quantification and uncertainty analysis of two fault trees using Monte Carlo, LHS, and RASCAL. The first fault tree is a small tree which was studied by Chang, Park, and Kim [38] using the MOMOD method. The second fault tree is for the emergency feedwater system at the Arkansas Nuclear One (ANO) Unit 1 power plant. This fault tree is relatively large and contains 33 basic events. The analysis of this fault tree is given in Martz, et al. [36], where they demonstrate that the standard DPD algorithm cannot be applied to this fault tree. However, the RASCAL modification is applicable and is used in this study.

The remainder of this chapter is divided into four sections. First, fault tree models are described. Next, the results of
the quantification of both fault trees is presented followed by the uncertainty analysis results of these fault trees. Finally, a summary section is given.

4.2 Fault Tree Models

A fault tree analysis is simply an analytic technique to determine all credible ways in which an undesired event may occur. The fault tree itself is only a graphical tool, its quantification being the topic of interest. The fault tree represents the various parallel and sequential combinations of faults which will lead to the undesired event.

A fault tree is a large complex of entities, denoted gates, which permit or inhibit the passage of fault logic up the tree. Events are defined as being lower if they serve as input to the gates, while higher events are outputs of the gate. For this study, only AND and OR gates are included although other types of logic gates exist [37].

Once the fault tree has been constructed, its quantification is performed using Boolean algebra [37]. Since it is not the intent of these studies to present all of the available fault tree analysis techniques, the remaining discussion is restricted to the necessary methods for the analysis to be performed.

An AND gate implies that two or more of the events which are inputs to the gate must occur (usually a failure event) in order for the event which is the output of the gate to occur. For example, an overspeed of an electric motor may require two relays to remain closed, however, if one remains open then no overspeed occurs. Conversely, an OR gate implies that only one of the events occur from two or more inputs to the gate for the output
event to occur. For the sake of example, assume that the fault tree shown in Figure 4-1 is being studied. In this figure the events A, B, D, and E are basic events, while C, F, and G are higher events than the basic events since they are outputs of the gates. Finally, T is the top event, i.e., the undesired event. From basic probability theory \[ P_{A \cap B} = P_A \cdot P_B \] (4-1)

where \( P_A \) is the probability of event A occurring, \( P_B \) is the probability of event B occurring, and \( P_{A \cap B} \) is the probability that both events A and B occur simultaneously. Similarly, the probability of either A or B occurring is given by

\[ P_{A + B} = P_A + P_B - P_A \cdot P_B \] (4-2)

where \( P_{A + B} \) is the probability of either event A or event B occurring. For the fault tree in Figure 4-1, the equivalent Boolean expression is

\[ T = ((A \cap B) \cap (A \cup D) \cap (B \cup E)). \] (4-3)

Using the idempotent and distributive laws of Boolean algebra, Equation (4-3) can be reduced to

\[ T = A \cap B \] . (4-4)

Thus, for this example, both events A and B must occur for the top event to occur. Equation (4-4) represents the minimal cut sets of the fault tree. That is, the occurrence of events A and B is the smallest combination of occurrences of basic events which, if they all occur, will cause the top event to occur. By definition, a minimal cut set is a combination of basic events.
Figure 4.1 Sample Fault Tree
sufficient for the top event to occur. The combination is minimal since, if any one of the basic events in the minimal cut set does not occur, then the top event cannot take place. Any fault tree can be reduced to a finite number of minimal cut sets. The one component minimal cut sets, if they exist, represent those single events whose occurrence will cause the top event. The two-component minimal cut sets represent the double event occurrences which together will cause the top event to occur. In general, the n-component minimal cut set requires the n components which together will cause the top event. The top event is then given as the union of the minimal cut sets:

\[ T = M_1 + M_2 + \ldots + M_N \]  

(4-5)

where \( M_k \) is the k-component minimal cut set.

Using Equations (4-1) and (4-2), the fault tree can now be quantified. For the example, the probability of event T, \( P_T \) occurring is

\[ P_T = P_A P_B \]

Two fault tree models have been selected for analysis in this study. First, a sample fault tree containing only six basic events and analyzed in reference [38] was examined. Table 4.1 gives the equivalent Boolean expression for the fault tree together with the parameters defining the basic event variables. In Table 4.1, \( P_T \) is the probability of the top event occurring if \( X_i \) is the probability of the event i occurring. For the lognormal distribution, the parameters for the sampling scheme in the Monte Carlo simulation are given by

\[ \mu = \log_e (\text{median}) \]
\[ \sigma = \log_e (\text{error factor})/1.645 \]
where \( \mu \) is the mean and \( \sigma \) is the standard deviation. The fault tree definition for the second study is contained in Table 4.2. The equation for \( P_T \) in Table 4.2 is a good example of how the DPD method can quickly become unmanageable, and why the RASCAL method is needed for these types of calculations. As pointed out in Chapter 2, if the output DPD is to be condensed back into 31 intervals, this implies that, on the average, over one billion data points are available to characterize each interval which is statistically accurate but not sufficiently more precise than using one thousand points per interval to justify the cost. Therefore, for the second fault tree study, only the RASCAL method is appropriate.

**TABLE 4.1. FAULT TREE DEFINITION FOR STUDY I**

<table>
<thead>
<tr>
<th>Event</th>
<th>Distribution</th>
<th>Median</th>
<th>Error Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>Lognormal</td>
<td>( 1.0 \times 10^{-3} )</td>
<td>3.0</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>Lognormal</td>
<td>( 3.0 \times 10^{-2} )</td>
<td>3.0</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>Lognormal</td>
<td>( 1.0 \times 10^{-2} )</td>
<td>3.0</td>
</tr>
<tr>
<td>( X_4 )</td>
<td>Lognormal</td>
<td>( 3.0 \times 10^{-2} )</td>
<td>3.0</td>
</tr>
<tr>
<td>( X_5 )</td>
<td>Lognormal</td>
<td>( 1.0 \times 10^{-2} )</td>
<td>3.0</td>
</tr>
<tr>
<td>( X_6 )</td>
<td>Lognormal</td>
<td>( 3.0 \times 10^{-3} )</td>
<td>3.0</td>
</tr>
</tbody>
</table>

At this point, the fault tree models have been defined and the input data identified. The evaluation of the fault trees is presented in the following section.
### Table 4.2. Fault Tree Definition for Study II

<table>
<thead>
<tr>
<th>Event</th>
<th>Distribution</th>
<th>Median</th>
<th>Error Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>Lognormal</td>
<td>1.50 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂</td>
<td>Lognormal</td>
<td>2.30 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₃</td>
<td>Lognormal</td>
<td>2.40 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₄</td>
<td>Lognormal</td>
<td>1.90 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₅</td>
<td>Lognormal</td>
<td>2.70 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₆</td>
<td>Lognormal</td>
<td>4.10 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₇</td>
<td>Lognormal</td>
<td>1.00 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₈</td>
<td>Lognormal</td>
<td>2.00 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₉</td>
<td>Lognormal</td>
<td>1.00 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₀</td>
<td>Lognormal</td>
<td>4.10 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₁</td>
<td>Lognormal</td>
<td>4.10 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₂</td>
<td>Lognormal</td>
<td>1.00 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₃</td>
<td>Lognormal</td>
<td>2.00 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₄</td>
<td>Lognormal</td>
<td>1.00 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₅</td>
<td>Lognormal</td>
<td>1.10 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₆</td>
<td>Lognormal</td>
<td>2.30 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₇</td>
<td>Lognormal</td>
<td>1.90 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₈</td>
<td>Lognormal</td>
<td>2.70 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₁₉</td>
<td>Lognormal</td>
<td>2.70 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₀</td>
<td>Lognormal</td>
<td>5.50 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₁</td>
<td>Lognormal</td>
<td>5.50 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₂</td>
<td>Lognormal</td>
<td>5.10 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₃</td>
<td>Lognormal</td>
<td>5.50 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₄</td>
<td>Lognormal</td>
<td>5.50 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₅</td>
<td>Lognormal</td>
<td>2.70 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₆</td>
<td>Lognormal</td>
<td>2.70 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₇</td>
<td>Lognormal</td>
<td>5.10 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₈</td>
<td>Lognormal</td>
<td>3.60 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₂₉</td>
<td>Lognormal</td>
<td>3.70 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₃₀</td>
<td>Lognormal</td>
<td>3.60 E-3</td>
<td>3</td>
</tr>
<tr>
<td>X₃₁</td>
<td>Lognormal</td>
<td>1.00 E-4</td>
<td>3</td>
</tr>
<tr>
<td>X₃₂</td>
<td>Lognormal</td>
<td>1.00 E-4</td>
<td>3</td>
</tr>
<tr>
<td>X₃₃</td>
<td>Lognormal</td>
<td>1.00 E-4</td>
<td>3</td>
</tr>
</tbody>
</table>

#### 4.3 Fault Tree Quantification Results

The results of the quantification of each of the two fault trees will proceed in series, that is, the results of study I (the small fault tree) will be presented first after which the results of study II will be given.
The results of the calculation of the mean and standard deviation for the top event probability for the smaller fault tree are given in Table 4.3. In order to make the comparisons as meaningful as possible, three Monte Carlo runs were made using 1000, 2000, and 5000 simulations, respectively. At the Monte Carlo 5000 run, the confidence bands on the results were very tight, and therefore, the calculation was discontinued. No further Monte Carlo runs were made.

Also shown in Table 4.3 are the results of two RASCAL analyses in which 10 and 31, denoted RASCAL-10 and RASCAL-31, respectively, discrete points were used. Similarly, two LHS analyses, denoted LHS-10 and LHS-31, were also performed using again 10 and 31 discrete strata. If the 5000 simulation Monte Carlo analysis is taken to represent the exact value, then it is immediately obvious that the LHS results are almost meaningless while the RASCAL results are predicting the mean very accurately and, when 31 data points are used, predicts the standard deviation within 3%.

Of course, when one is performing a risk analysis, it is rarely sufficient to use mean values in the calculation of the top event probability. Rather, a conservative analysis is usually required. For this reason, it is necessary to examine each methods ability to estimate the CDF of the top event probability. The 10th, 50th, and 90th percentile, as calculated by each method, are given in Table 4.4. The results of the new Monte Carlo
TABLE 4.3. MEAN AND STANDARD DEVIATION FOR THE TOP EVENT PROBABILITY FOR FAULT TREE STUDY I

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Relative Error*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monte Carlo 5000</td>
<td>8.8036 E-3</td>
<td>3.9560 E-3</td>
<td>0</td>
</tr>
<tr>
<td>Monte Carlo 2000</td>
<td>8.912 E-3</td>
<td>4.053 E-3</td>
<td>1.2%</td>
</tr>
<tr>
<td>Monte Carlo 1000</td>
<td>8.552 E-3</td>
<td>3.669 E-3</td>
<td>-2.9%</td>
</tr>
<tr>
<td>LHS-10</td>
<td>1.143 E-2</td>
<td>5.661 E-3</td>
<td>29.8%</td>
</tr>
<tr>
<td>RASCAL-10</td>
<td>8.792 E-3</td>
<td>3.601 E-3</td>
<td>-0.1%</td>
</tr>
<tr>
<td>LHS-31</td>
<td>1.030 E-2</td>
<td>6.432 E-3</td>
<td>17.0%</td>
</tr>
<tr>
<td>RASCAL-31</td>
<td>8.820 E-3</td>
<td>3.843 E-3</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

\* The Relative Error is R.E. = 100(X - T)/T where X is the estimated value and T is the true value. T is taken as the Monte Carlo 5000 results.

The sampling method developed in reference [38] are also given in Table 4.5 and are identified as MOMOD. For the 10th and 50th percentile calculation, all of the methods give approximately the same accuracy while the 90th percentile is very poorly predicted by the LHS analysis with the RASCAL and MOMOD methods giving approximately the same accuracy. However, while the MOMOD method gives a reduction in the Monte Carlo running time of 16.8, the RASCAL-31 run represents a reduction of 34.5. Thus, for approximately the same accuracy, the RASCAL method provides a reduction of two or more over the computation time of the MOMOD technique. Figures 4-2 and 4-3 present the results of the calculation of
Figure 4.2 Small Fault Tree Analysis—MOMOD and RASCAL
Figure 4.3 Small Fault Tree Analysis—LHS and RASCAL
the empirical CDF by each of the methods selected for study. It is important to note that, if the Monte Carlo result is taken to be the exact solution, then results lying above this line are nonconservative since the result being estimated is an adverse effect. Thus, if such a high degree of accuracy were not exhibited by the RASCAL method, its use would be suspect. However, given the

<table>
<thead>
<tr>
<th>Method</th>
<th>Percentile</th>
<th>CPU Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10th</td>
<td>50th</td>
</tr>
<tr>
<td>Monte Carlo</td>
<td>4.684 E-3</td>
<td>8.049 E-3</td>
</tr>
<tr>
<td>RASCAL-10</td>
<td>4.086 E-3</td>
<td>8.469 E-3</td>
</tr>
<tr>
<td>LHS-10</td>
<td>5.857 E-3</td>
<td>1.026 E-2</td>
</tr>
<tr>
<td>RASCAL-31*</td>
<td>4.591 E-3</td>
<td>7.846 E-2</td>
</tr>
<tr>
<td>LHS-31*</td>
<td>4.865 E-3</td>
<td>8.531 E-3</td>
</tr>
</tbody>
</table>

* Linear interpolation of percentiles used

required accuracy of most risk analysis calculations, the use of the RASCAL technique does not represent an overly nonconservative result, and it may be used in place of the much more expensive Monte Carlo analysis.
The second fault tree to be studied was first presented in "A Comparison of Methods for Uncertainty Analysis of Nuclear Power Plant Safety System Fault Tree Models" by H. F. Martz, et al. [37]. In this study, several different types of distributions were used to represent the basic event probabilities. However, in this study, only a comparison to the lognormal study will be made. Martz correctly points out for his study that the DPD method is not capable of analyzing the fault tree defined in Table 4.2. However, the RASCAL version of the standard DPD algorithm does allow the calculations to be performed.

Table 4.5 gives the results of the three Monte Carlo, the two RASCAL, and two LHS analyses of the larger fault tree. What this table shows is the same trend seen in the smaller fault tree study, i.e., the RASCAL procedure consistently outperforms the LHS analysis while providing a close approximation to the Monte Carlo result for a fraction of the cost. The empirical CDF calculated by the RASCAL method is compared to the 5000 simulation run of the Monte Carlo code in Figure 4-4. The LHS results are not included because of their poor performance.

The final topic to be addressed by this study is the estimation of extreme tails probabilities. In many risk analyses it is important to calculate the probability of very rare events, Chapter 3 discussed how the RASCAL method can be viewed as an importance sampling scheme for Monte Carlo analysis. The next section describes a method for using the RASCAL technique to estimate upper tail probabilities. The application being made is for the smaller fault tree study.

4.4 Estimating Upper Tail Probabilities

In the previous analyses, using the RASCAL method, the value of the discrete points were chosen such that these were equal
Figure 4.4 Emergency Feedwater System Fault Tree Analysis
<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Result</th>
<th>Standard Deviation Result</th>
<th>Mean Error</th>
<th>Relative Error Standard Deviation Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monte Carlo 5000</td>
<td>1.635 E-3</td>
<td>6.490 E-4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Monte Carlo 2000</td>
<td>1.640 E-3</td>
<td>6.740 E-4</td>
<td>0.3%</td>
<td>3.9%</td>
</tr>
<tr>
<td>Monte Carlo 1000</td>
<td>1.609 E-3</td>
<td>5.487 E-4</td>
<td>-1.6%</td>
<td>-15.5%</td>
</tr>
<tr>
<td>LHS</td>
<td>2.676 E-3</td>
<td>1.699 E-3</td>
<td>68.8%</td>
<td>161.8%</td>
</tr>
<tr>
<td>RASCAL-10</td>
<td>1.646 E-3</td>
<td>5.355 E-4</td>
<td>0.7%</td>
<td>-17.5%</td>
</tr>
<tr>
<td>LHS-31</td>
<td>1.860 E-3</td>
<td>6.480 E-4</td>
<td>13.8%</td>
<td>-0.2%</td>
</tr>
<tr>
<td>RASCAL-31</td>
<td>1.647 E-3</td>
<td>6.047 E-4</td>
<td>0.7%</td>
<td>-6.8%</td>
</tr>
</tbody>
</table>

Mathematically, the procedure used for N discrete points is

\[
P_i = \frac{1.0}{N} \quad \text{for all}
\]

\[
\int_{b_{i-1}}^{b_i} f(x) \, dx = P_i = P \quad \text{for all } i
\]

(4-6)

where \( b_0 \) is equal to negative infinity or the minimum value of the PDF, \( f(x) \). The \( b_i \) thus determined defined intervals of equal probability. The value of the DPD used is taken as the conditional mean of the interval and is denoted \( Z_i \). Thus, each DPD is given by

\[
\text{DPD} = [(Z_i, P)]
\]
where $Z_i$ is the conditional mean of $(b_{i-1}, b_i)$ and $b_i$ is defined by Equation (4-6). For this procedure, the largest value of the CDF which can be estimated is $1 - P$. Thus, if the 99.99 percentile value is needed, then

$$1 - P = 0.999$$

$$1/N = 1 - 0.999 = 0.001$$

$$N = 1,000$$

Obviously, using a value of $N$ equal to 1,000 is not cost efficient and probably not possible from the consideration of computer storage limits. To estimate the 99.99 percentile value, it is necessary to use unequal values of $P_i$ in Equation (4-6).

The first test case uses the distribution of probability masses shown in Table 4-6. The first 12 points of 20 covered 95% of the distribution while the last 8 covered the remaining 5%. The results are compared in Figure 4-2. As this figure shows there is not a great deal of accuracy in the result of this 20 discrete point RASCAL analysis, although the upper tails are not dramatically off. The hypothesized reason for the inaccuracy is because of the relatively large statistical inaccuracy in estimating the lower 95% of the input distribution. Thus, while the upper tails look reasonably good, they are not as accurate as the RASCAL-31 case (assuming the Monte Carlo result is exact) for estimating the 97th percentile. Thus, the hypothesis is that the statistical inaccuracy of estimating lower percentile
values is being propagated to the higher percentile values in the combination process.

To test this hypothesis, a second run of the RASCAL code was made using the probability intervals given in Table 4-7. In this case, the change between probability values was less drastic than in the first case. That is, more statistical accuracy was required throughout the range of the DPD. The result of this calculation is shown in Figure 4-5 with the upper tails shown in more detail in Figure 4-6. Table 4-8 gives a comparison of the upper tail values by the RASCAL-50 and Monte Carlo analyses and shows that the RASCAL method provides the value of the CDF up to the 99.9% value within 6%. Even with 50 points, the running time of the RASCAL method was 25.8 seconds or an order of magnitude less than the Monte Carlo analysis. Therefore, the RASCAL technique can be used as an alternative to importance sampling for Monte Carlo.
Figure 4.5 Unequal Probability Interval Comparison
Figure 4.6 Upper Tail Calculation Using Unequal Probability Intervals
### Table 4-7. Probability of Masses for Each of 50 Intervals of Individual RASCAL for Fault Tree Study I

<table>
<thead>
<tr>
<th>Intervals</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>0.001</td>
</tr>
</tbody>
</table>

### Table 4-8. Comparison of Upper Tails Calculation by Monte Carlo and RASCAL for Fault Tree Study I

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Monte Carlo</th>
<th>RASCAL-50</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>1.64E-2</td>
<td>1.57E-2</td>
<td>-4.3%</td>
</tr>
<tr>
<td>96</td>
<td>1.70E-2</td>
<td>1.63E-2</td>
<td>-4.1%</td>
</tr>
<tr>
<td>96</td>
<td>1.80E-2</td>
<td>1.75E-2</td>
<td>-2.8%</td>
</tr>
<tr>
<td>98</td>
<td>1.95E-2</td>
<td>1.88E-2</td>
<td>-3.6%</td>
</tr>
<tr>
<td>99</td>
<td>2.26E-2</td>
<td>2.14E-2</td>
<td>-5.3%</td>
</tr>
<tr>
<td>99.5</td>
<td>2.50E-2</td>
<td>2.36E-2</td>
<td>-5.6%</td>
</tr>
<tr>
<td>99.8</td>
<td>2.76E-2</td>
<td>2.69E-2</td>
<td>-2.5%</td>
</tr>
<tr>
<td>99.9</td>
<td>2.85E-2</td>
<td>2.91E-2</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

### 4.5 Summary

The analysis of the smaller fault tree with only six basic events demonstrated that the RASCAL method provides a good estimate (within 5%) of a Monte Carlo analysis requiring approximately 35 times more computer time. The basic LHS analysis gives poor results overall when a few discrete strata are used. When on
the order of 30 strata are used, reasonable estimates of the mean are obtained while the estimate of the standard deviation is crude. The prediction of the empirical CDF of the top event probability is accurate when the RASCAL method is employed and poor when the LHS design is used. Of course, if replicate LHS designs are made, the accuracy will increase, as discussed in Chapter 4. However, this requires additional levels of sophistication for the user of LHS techniques while only basic procedures are being compared in this study.

The analysis of the larger fault tree, consisting of 33 basic events, was performed to confirm the results of the smaller tree study and to demonstrate that the RASCAL modification of the DPD algorithm does allow the evaluation of this tree which could not be performed with the standard DPD algorithm. The results of this study did coincide with the results of the previous study in that the RASCAL method provided an accurate solution while the LHS method performed poorly.

Finally, the use of the RASCAL method as an alternative to Monte Carlo importance sampling was examined. As long as the discrete probability mass distribution did not change drastically between points, it was found that the RASCAL method can predict up to the 99.9 percentile value within 6%. This has important implications since standard Monte Carlo importance sampling schemes require changes in the coding (see Chapter 2) while the RASCAL method allows the sampling procedure to be changed by input, i.e., no coding changes are required.

This initial study of the RASCAL method has demonstrated that RASCAL works well for linear, algebraic, and time independent modeling systems. The next chapter examines the use of RASCAL for a model which is time dependent and modeled by a linear, differential equation model.
5.1 Introduction

This chapter presents the results of the second case study using each of the probabilistic methods selected for analysis. The goal of this study is to determine each method's ability to predict the uncertainty in the corrosion depth in a nuclear waste canister. The corrosion model selected for use in the current analysis is one suggested by Sastre and Pescatore [39]. In their study there is probabilistic data provided for the inputs to the model. Thus, this model provides a good test of the Monte Carlo, LHS, and RASCAL uncertainty analysis techniques since it represents a physically realistic model and some data is available for its evaluation. It is important to note that the results presented here, represented by mean values, do not necessarily correspond to the "engineering" best estimate value of the corrosion depth of a nuclear waste canister. Similarly, the upper and lower bounds of the corrosion depth given below only are meaningful relative to each other—the absolute value is not to be taken as representing reality. The reasons for this is, that while these physical models are based on the physical phenomena, they are not complete and they are idealized. Thus, while the model is appropriate for assessing the applicability and performance of uncertainty analysis methods, it is not appropriate to attach significance to the actual numeric values which result.
The remainder of this chapter is structured as follows. First, the corrosion model is described after which the necessary input variables to this model are defined. The results of the time dependent uncertainty analysis are given next. The evaluation of the performance of the Monte Carlo, LHS, and RASCAL methods is then presented. The next section, in anticipation of the following chapter, presents a qualitative discussion on the use of these methods in probabilistic and sensitivity analysis. Finally, a brief summary is given.

5.2 Corrosion Model for Waste Canister Evaluation

The corrosion model used for this study is one suggested by Sastre and Pescatore [42]. The general form is given by

\[ C_D = (K_u + K_p) \exp(a/T) O^b C_l^c t^n \]  

(5-1)

where

- \( C_D \): corrosion (mm)
- \( K_u \): uniform corrosion factor
- \( K_p \): pitting corrosion factor
- \( T \): temperature (K)
- \( O \): oxygen concentration (ppm)
- \( C_l \): chlorine concentration (ppm)
- \( t \): time (years)
- \( a, b, c, n \): empirical constants

The constants in Equation (5-1) are determined by a regression analysis of data developed by Westinghouse [40]. The calculation of the temperature, \( T \), is performed using a one-dimensional coordinate system in which the canister, overpack, and backfill regions are represented by concentric circles. The near field
temperature, \( T_n(t) \), is calculated from
\[
T_n(t) = T_f(t) + T
\]  
(5-2)
where \( T \) is the sum of terms of the form
\[
T = D_1/(2 KL) \log_2(D_2/D_1)
\]  
(5-3)
and the far field temperature is given by
\[
T_f(t) = 1/K(a/\pi)^{1/2} \sum_{i=1}^{n} a_i/r_i Da(r_it) + T_0
\]  
(5-4)

The expression for \( T \) can be obtained from reference [41]. The terms in Equations (5-3) and (5-4) represent the following quantities:

- \( D_i \) : diameter of physical region (M)
- \( a \) : thermal diffusivity (M²M/sec)
- \( K \) : thermal conductivity (W/M/K)
- \( a_i, r_i \) : empirical constants
- \( Da(x) \) : Dawson's integral evaluated at \( x \).

In order to check the computer model developed for this study, the 0.1 to 99.9 percentile spread for the uniform corrosion facts, \( K_u \), were calculated and compared to the results of reference [39]. The range computed in this reference for \( K_u \) was (0.00147, 676) while the computer model developed for this study gave (0.00157, 636). This range was found to be within the needed accuracy for this study since the ultimate goal is the evaluation of uncertainty analysis methods not an accurate prediction of the corrosion depth.

5.3 Input Variable Definition

For the corrosion model presented in the previous chapter, there are a total of 16 inputs to the model. Of these 16 inputs, 10 are considered to be random variables. Eight of the random variables are described by uniform distributions, one by a normal distribution, and the last is described by a special form of a Weibull distribution, called a Rayleigh distribution in the probabilistic literature. Each of these variables, both constant
and stochastic, are shown in Table 5.1 with their associated values or distribution type and parameters.

In Table 5.1, the first parameter listed is the lower limit for a uniform distribution, the minimum value for the Rayleigh distribution, and the variables value for a constant. The second parameter is the range of the uniform distribution and the spread of the Rayleigh distribution, (see Table 2.1). The discretization of each of the random variable-input distributions was done in the same way for both the LHS and RASCAL analyses. The user supplies the number of discrete points which is to be used for the analysis, and the program selects the points from the distribution so that the intervals between points (or strata) have the same probability weight.

In the RASCAL analysis, the conditional mean of each distribution is calculated and the probability mass is lumped at this point. For the LHS design, no conditional mean calculation is required since the strata must be sampled for the actual value of the random variable to be used in the calculation of the response. For the 10 random variables under consideration, the discretization shown in Table 5.2 was used in both analyses. Obviously, for the RASCAL method the associated probability in the vector of ordered pairs is always equal to 1/N, or 0.1, for the case shown.

The RASCAL analysis has an option for using unequal probability intervals. If this option is selected, the user supplies the probability value that each stratum should contain, and the program calculates where the discrete points should be located. This option does not exist for the LHS analysis since, by definition, LHS must use equal probability intervals. When the number of discrete points is different from 10, the same procedure is used to generate results corresponding to Table 5.2.
### TABLE 5.1 CORROSION MODEL INPUT DATA DESCRIPTION

<table>
<thead>
<tr>
<th>Variable</th>
<th>Distribution</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rock Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Geothermal Temperature (°C)</td>
<td>Uniform</td>
<td></td>
</tr>
<tr>
<td>2. Thermal Conductivity (W/M/K)</td>
<td>Uniform</td>
<td></td>
</tr>
<tr>
<td>3. Rock Density (Kg/M**3)</td>
<td>Uniform</td>
<td></td>
</tr>
<tr>
<td>4. Specific Heat (J/Kg/K)</td>
<td>Uniform</td>
<td></td>
</tr>
<tr>
<td><strong>Waste Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Waste Age (Years)</td>
<td>Constant</td>
<td></td>
</tr>
<tr>
<td>6. Initial Power (W)</td>
<td>Constant</td>
<td></td>
</tr>
<tr>
<td>7. Backfill Outer Diameter (M)</td>
<td>Constant</td>
<td></td>
</tr>
<tr>
<td>8. Backfill Thermal Conductivity (W/M/K)</td>
<td>Uniform</td>
<td></td>
</tr>
<tr>
<td>9. Overpack Outer Diameter (M)</td>
<td>Constant</td>
<td></td>
</tr>
<tr>
<td>10. Canister Wall Thickness (M)</td>
<td>Constant</td>
<td></td>
</tr>
<tr>
<td>11. Canister Length (M)</td>
<td>Constant</td>
<td></td>
</tr>
<tr>
<td><strong>Corrosion Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Pitting Corrosion Factor</td>
<td>Uniform</td>
<td></td>
</tr>
<tr>
<td>13. Chlorine Concentration (ppm)</td>
<td>Uniform</td>
<td></td>
</tr>
<tr>
<td>14. Oxygen Concentration (ppm)</td>
<td>Uniform</td>
<td></td>
</tr>
<tr>
<td>15. Exponent of Time</td>
<td>Normal</td>
<td></td>
</tr>
<tr>
<td>16. Uniform Corrosion Parameter (mM/year)</td>
<td>Rayleigh</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable Number</th>
<th>Parameter 1</th>
<th>Parameter 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54.0</td>
<td>60.0</td>
</tr>
<tr>
<td>2</td>
<td>1.25</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>390.0</td>
<td>2410.0</td>
</tr>
<tr>
<td>4</td>
<td>820.0</td>
<td>1160.0</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>2100.0</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>0.686</td>
<td>NA</td>
</tr>
<tr>
<td>8</td>
<td>0.40</td>
<td>1.4</td>
</tr>
<tr>
<td>9</td>
<td>0.325</td>
<td>NA</td>
</tr>
<tr>
<td>10</td>
<td>0.053</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>4.10</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>6.0</td>
</tr>
<tr>
<td>13</td>
<td>1.00</td>
<td>101.0</td>
</tr>
<tr>
<td>14</td>
<td>0.01</td>
<td>3.0</td>
</tr>
<tr>
<td>15</td>
<td>0.469</td>
<td>0.0349</td>
</tr>
<tr>
<td>16</td>
<td>0.01</td>
<td>118.643</td>
</tr>
</tbody>
</table>
5.4 Time Dependent Uncertainty Analysis Results

The first effect to be examined is that of pitting corrosion. Because of the variation associated with the uniform pitting corrosion factors, rate equations were used for both the uniform and pitting corrosion models. Thus,

\[ U(T + \Delta t) = U(T) + \Delta t \frac{U(t)}{\Delta t} \]  \hspace{1cm} (5-5)

\[ P(T + \Delta t) = P(T) + \Delta t \frac{P(t)}{\Delta t} \]  \hspace{1cm} (5-6)

where \(U(t)\) and \(P(t)\) are the predictive equations for uniform and pitting corrosion, respectively, and \(U(0)\) and \(P(0)\) are both equal to zero. In Equation (5-5), the exponent of time, i.e., variable 15 from Table 5.1, is held constant at its mean value, while it is normally distributed, with parameters given in Table 5.1, for use in Equation (5-6).

Two runs of the model were made using the model described and tested as presented in the previous sections. In the first run the pitting corrosion factor (variable 12) was set equal to zero. In the second run, the pitting corrosion was set equal to a uniformly distributed variable between 1.0 and 6.0. This is important to the subsequent analysis to indicate whether or not one form of corrosion or another is more important to the overall uncertainty in corrosion rates.

Figure 5.1 presents the results of the two runs of the model. In this figure, the mean corrosion depth is plotted together with a one standard deviation bound about the mean value. The scale shown is between 100 and 1,000 years. As can be seen from Figure 5.1, the pitting corrosion model increases the mean corrosion depth by approximately a factor of 5 over the case when only uniform corrosion is used. Additionally, the standard
FIGURE 5.1 MEAN CORROSION DEPTH AS A FUNCTION OF TIME BY MONTE CARLO ANALYSIS
deviation is slightly larger, relatively, in the pitting corrosion case than in the uniform corrosion case. If a relative measure of the uncertainty is defined as

\[ Q = \frac{\text{Standard deviation}}{\text{mean}} \]

then the pitting corrosion analysis produces values of \( Q \) ranging from 65% at year 200 to 30% at year 1,000, while the uniform corrosion analysis gives corresponding values of 45% and 20%. Because of the difference in the mean values, this implies that the standard deviations differ by slightly more than a factor of seven on a relative scale.

The indication is that the addition of the pitting corrosion model produces an uncertainty slightly larger than anticipated from the simple fact that the pitting corrosion factor is uniformly distributed between one and six. However, it must be remembered that the exponent of time is also allowed to vary in the pitting corrosion while it is fixed in the uniform corrosion model. This cannot account for the entire increase in the uncertainty in the data, since if this were the case, then the mean values obtained from the two models would only be about a factor of 3.5 different rather than the factor of almost 5.0 that was found. The conclusion that one must come to is that an additional source of uncertainty has been introduced due to the combination of the two models. This source of uncertainty is often overlooked in many analyses but must at least be estimated, as has been done here, to complete the description of the contributors to the response uncertainty.

The next step in the analysis is to examine the predicted corrosion depth by each of the three uncertainty analysis methods. For the sake of comparison, a Monte Carlo analysis was performed using 5,000 sample runs for the calculations. Since it is straightforward to calculate confidence bands for the mean value of the corrosion
TABLE 5.2 DISCRETIZATION USED FOR THE RASCAL AND LHS ANALYSIS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Discrete Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54.60 55.20 55.80 56.40 57.00 57.60 58.20 58.80 59.40 60.00</td>
</tr>
<tr>
<td>2</td>
<td>1.375 1.5 1.625 1.75 1.875 2.0 2.125 2.25 2.375 2.5</td>
</tr>
<tr>
<td>3</td>
<td>2449.0 2488.0 2527.0 2566.0 2566.0 2566.0 2566.0 2566.0 2566.0 2566.0</td>
</tr>
<tr>
<td>4</td>
<td>854.0 888.0 922.0 956.0 990.0 1024.0 1058.0 1092.0 1126.0 1160.0</td>
</tr>
<tr>
<td>8</td>
<td>0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4</td>
</tr>
<tr>
<td>12</td>
<td>1.5 2.0 2.5 3.0 3.5</td>
</tr>
<tr>
<td>13</td>
<td>10.9 20.8 30.7 40.6 50.0</td>
</tr>
<tr>
<td>14</td>
<td>0.309 0.608 0.907 1.206 1.505</td>
</tr>
<tr>
<td>15</td>
<td>0.41157 0.43284 0.44545 0.45554 0.45460</td>
</tr>
<tr>
<td>16</td>
<td>0.010 54.463 79.256 100.2 119.91</td>
</tr>
</tbody>
</table>

depth, this analysis is assumed to represent "truth", i.e., the correct value. Figure 5-2 gives the mean corrosion depth as calculated by the Monte Carlo method for years 100 to 1,000, while the 95% confidence limits are given in Table 5.3.
FIGURE 5.2 MEAN CORROSION DEPTH CALCULATED BY EACH UNCERTAINTY ANALYSIS METHOD.
### TABLE 5.3 MEAN CORROSION DEPTH AS PREDICTED BY MONTE CARLO SIMULATION

<table>
<thead>
<tr>
<th>Year</th>
<th>Lower Bound (mM)</th>
<th>Mean (mM)</th>
<th>Upper Bound (mM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>15.4</td>
<td>15.9</td>
<td>16.3</td>
</tr>
<tr>
<td>200</td>
<td>25.1</td>
<td>25.7</td>
<td>26.5</td>
</tr>
<tr>
<td>300</td>
<td>33.1</td>
<td>33.8</td>
<td>34.4</td>
</tr>
<tr>
<td>400</td>
<td>39.8</td>
<td>40.2</td>
<td>40.7</td>
</tr>
<tr>
<td>500</td>
<td>45.5</td>
<td>46.0</td>
<td>46.6</td>
</tr>
<tr>
<td>600</td>
<td>50.8</td>
<td>51.6</td>
<td>52.3</td>
</tr>
<tr>
<td>700</td>
<td>55.3</td>
<td>56.2</td>
<td>56.7</td>
</tr>
<tr>
<td>800</td>
<td>60.2</td>
<td>60.8</td>
<td>61.5</td>
</tr>
<tr>
<td>900</td>
<td>64.3</td>
<td>64.9</td>
<td>65.6</td>
</tr>
<tr>
<td>1,000</td>
<td>68.4</td>
<td>68.9</td>
<td>69.9</td>
</tr>
</tbody>
</table>

The RASCAL method suffers from the drawback that for the analysis currently being performed there are 16 variables and a minimum of 10 discrete points. Since many of these variables are dependent and no condensation can be performed, this implies that at least $2.0\times10^{10}$ storage points are required. Obviously, this is outside the limits of most fixed computer storage requirements. However, the RASCAL algorithm is capable of analyzing this problem and is used in this case.

The results of the RASCAL and LHS calculations are shown in Table 5.4 and 5.5, respectively. The result of the RASCAL case when 10 discrete points are used, denoted RASCAL-10, are very good but were actually too good to believe. As can be seen from the RASCAL run when 20 points are used, denoted RASCAL-20, the results when 10 points are used were obviously a result of fortuitous circumstances. However, as was expected, the RASCAL method is always better than the LHS method. The reason
for this is that no replicates of the LHS design were used. In this case, there are less sample points taken in the LHS analysis, and, thus, less accuracy is expected. This is an important point since the evaluation of any method should include the ability of the method to perform when the user is not an expert in the particular technique. Since the LHS analysis was only a factor of two less costly, in terms of computer time, than the RASCAL analysis, this implies only two replicates could be performed for the same amount of computer time. Yet in this case, the RASCAL analysis would still use approximately a factor of five more sample points (as opposed to a factor of 10.0 in the actual analysis) and, therefore, it is expected that the RASCAL method would still outperform the LHS method with these restrictions.

In the LHS analysis, there appears to be some sort of bias being introduced into the results. For the 10 strata cases, the mean prediction gets closer to the Monte Carlo result as time increases, while for the 20 strata analysis, the reverse is true. One possible explanation of this result is that the uniform corrosion factor is causing the conditional mean of each strata to be skewed to the right, i.e., towards higher values. Because the RASCAL analysis is providing an estimate of the distribution within each strata while the LHS method is not, the bias will be introduced until enough combinations have been made for the distribution to approach a normal distribution or until enough strata have been used to reduce the bias. Alternatively, replicate designs using LHS can be used to assimilate the effects of such skewness.
### TABLE 5.4 MEAN CORROSION DEPTH (MM) AS CALCULATED BY THE RASCAL METHOD

<table>
<thead>
<tr>
<th>Year</th>
<th>Monte Carlo</th>
<th>RASCAL-10</th>
<th>RASCAL-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>15.9</td>
<td>15.6</td>
<td>15.7</td>
</tr>
<tr>
<td>200</td>
<td>25.7</td>
<td>25.1</td>
<td>25.4</td>
</tr>
<tr>
<td>300</td>
<td>33.8</td>
<td>33.3</td>
<td>33.0</td>
</tr>
<tr>
<td>400</td>
<td>40.2</td>
<td>40.4</td>
<td>39.3</td>
</tr>
<tr>
<td>500</td>
<td>46.0</td>
<td>46.5</td>
<td>45.2</td>
</tr>
<tr>
<td>600</td>
<td>51.6</td>
<td>51.3</td>
<td>50.6</td>
</tr>
<tr>
<td>700</td>
<td>56.2</td>
<td>56.3</td>
<td>55.3</td>
</tr>
<tr>
<td>800</td>
<td>60.8</td>
<td>60.6</td>
<td>59.7</td>
</tr>
<tr>
<td>900</td>
<td>64.9</td>
<td>64.7</td>
<td>63.8</td>
</tr>
<tr>
<td>1,000</td>
<td>68.9</td>
<td>68.7</td>
<td>67.9</td>
</tr>
</tbody>
</table>

### TABLE 5.5 MEAN CORROSION DEPTH AS CALCULATED BY THE LHS METHOD

<table>
<thead>
<tr>
<th>Year</th>
<th>Monte Carlo</th>
<th>LHS-10</th>
<th>LHS-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>15.9</td>
<td>14.1</td>
<td>11.6</td>
</tr>
<tr>
<td>200</td>
<td>25.7</td>
<td>22.6</td>
<td>32.9</td>
</tr>
<tr>
<td>300</td>
<td>33.8</td>
<td>29.7</td>
<td>32.9</td>
</tr>
<tr>
<td>400</td>
<td>40.2</td>
<td>35.5</td>
<td>39.4</td>
</tr>
<tr>
<td>500</td>
<td>46.0</td>
<td>41.1</td>
<td>44.5</td>
</tr>
<tr>
<td>600</td>
<td>51.6</td>
<td>46.4</td>
<td>49.2</td>
</tr>
<tr>
<td>700</td>
<td>56.2</td>
<td>52.1</td>
<td>53.8</td>
</tr>
<tr>
<td>800</td>
<td>60.8</td>
<td>57.2</td>
<td>58.2</td>
</tr>
<tr>
<td>900</td>
<td>64.9</td>
<td>60.4</td>
<td>62.5</td>
</tr>
<tr>
<td>1,000</td>
<td>68.9</td>
<td>65.1</td>
<td>66.3</td>
</tr>
</tbody>
</table>

While some indication of the RASCAL and LHS methods effectiveness can be obtained from an examination of the mean values, the ultimate goal is to assess each method's ability to indicate the uncertainty in the model. Figure 5-3 presents a plot of
the standard deviation calculated by Monte Carlo, RASCAL, and LHS analysis. From this figure, it is obvious that the RASCAL method of analysis does considerably better than the LHS method in predicting the standard deviation which is one measure of the uncertainty in the response model. For the reasons discussed above, this result is to be expected since the DPD method uses many more sample points than the LHS method. Finally, Table 5.6 gives the results of the calculation of each of the first four moments of the distribution of the corrosion depth at year 1,000.

**TABLE 5.6 MOMENTS OF THE DISTRIBUTION OF CORROSION DEPTH AT YEAR 1,000 BY MONTE CARLO SIMULATION**

<table>
<thead>
<tr>
<th>First Moment</th>
<th>Second Moment</th>
<th>Third Moment</th>
<th>Fourth Moment</th>
<th>Beta₁</th>
<th>Beta₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.9</td>
<td>424.1</td>
<td>5881.8</td>
<td>6.75e5</td>
<td>0.6733</td>
<td>3.7513</td>
</tr>
</tbody>
</table>

It is important to keep in mind during the assessment of the three uncertainty analysis methods the distinctions among probabilistic, uncertainty, and sensitivity analysis. For probabilistic analysis, it is important to define the PDF of the response. Finally, sensitivity analysis is concerned with determining how the response can change due to small changes in the input variables around the design points, usually the mean values. Since the primary focus of this study is on the performance of the methods for uncertainty analysis, each method will be ranked according to its performance to estimate the range of possible responses. However, since each of the methods can also be used in probabilistic or sensitivity analysis, some
discussion of their performance in these areas will also be provided.

5.5 Uncertainty Analysis Performance Evaluation

Because of the complex nature of the function being evaluated, there is no analytic solution to predict what the possible spread in the distribution of the corrosion depth may be. Therefore, the Monte Carlo results will be assumed to represent "truth" in the estimate of the spread in the distribution. This has the added advantage that, for the Monte Carlo analysis, estimates of how much confidence one can place in the estimates of certain statistical measures can be.

This immediately raises the question of, "What measure should be used to represent the uncertainty in the spread of the distribution of corrosion depths?" In many uncertainty analyses (see, for example, reference [41]), the measure of the range of possible responses is the variance. In this evaluation, variances are not used for two reasons. First, the Monte Carlo estimate is known to be less accurate for estimating higher order moments; and since this is to represent truth, a more reasonable measure should be examined. Secondly, since so few points are used in the LHS method, it would automatically be at a disadvantage in such an evaluation since the number of points used to estimate the variance would not be sufficient to ever produce a reasonable estimate of the variance. Therefore, the measure of the uncertainty in the response to be used in this evaluation is percentiles. The percentile measure of uncertainty offers two distinct advantages: (1) confidence level bounds can be obtained for the Monte Carlo results, and (2) it is known that for both the DPD and LHS methods as the number of discrete points is increased, the percentile values must approach the Monte Carlo values.
discussion of their performance in these areas will also be provided.

5.5 Uncertainty Analysis Performance Evaluation

Because of the complex nature of the function being evaluated, there is no analytic solution to predict what the possible spread in the distribution of the corrosion depth may be. Therefore, the Monte Carlo results will be assumed to represent "truth" in the estimate of the spread in the distribution. This has the added advantage that, for the Monte Carlo analysis, estimates of how much confidence one can place in the estimates of certain statistical measures can be.

This immediately raises the question of, "What measure should be used to represent the uncertainty in the spread of the distribution of corrosion depths?" In many uncertainty analyses (see, for example, reference [41]), the measure of the range of possible responses is the variance. In this evaluation, variances are not used for two reasons. First, the Monte Carlo estimate is known to be less accurate for estimating higher order moments; and since this is to represent truth, a more reasonable measure should be examined. Secondly, since so few points are used in the LHS method, it would automatically be at a disadvantage in such an evaluation since the number of points used to estimate the variance would not be sufficient to ever produce a reasonable estimate of the variance. Therefore, the measure of the uncertainty in the response to be used in this evaluation is percentiles. The percentile measure of uncertainty offers two distinct advantages: (1) confidence level bounds can be obtained for the Monte Carlo results, and (2) it is known that for both the DPD and LHS methods as the number of discrete points is increased, the percentile values must approach the Monte Carlo values.
Percentile Measure of Uncertainty and Confidence Levels. The result of a Monte Carlo analysis will be a vector of \( N \) responses where \( N \) is the number of Monte Carlo simulations. If this vector of responses, denoted \( V \), is ordered then an empirical cumulative distribution function (CDF) can be constructed. The first value of \( V \), \( v(1) \), is the value of the 1.0/\( N \) percentile value. For example, in the Monte Carlo analysis performed for this study, \( N \) is equal to 1,000. Thus, \( v(1) \) is the estimate of the 0.1 percentile value of the distribution of \( V \), in this case the corrosion depth. Similarly, \( v(10) \) is the estimate of the first percentile value, \( v(100) \), the estimate of the tenth percentile value, and so on. If the \( M \) percentile value is desired, then it can be shown that, at a confidence level of \( g \), this value is bounded by

\[
M_u = \text{int}[m + Z((1 + g)/2)(m^2(1 - m/N))^{0.5}] + 1
\]

\[
M_l = \text{int}[m - Z((1 + g)/2)(m^2(1 - m/N))^{0.5}] + 1
\]

\[
m = MN/100
\]

where \( \text{int}[ ] \) is the integer part of the term inside the square brackets, and \( Z(x) \) is the value of the standard normal distribution at the point \( x \). For a confidence level of 95%, i.e., we are 95% confident that the actual value of the \( M \) percentile value lies between \( M_l \) and \( M_u \), \( Z(0.975) \) is equal to 0.83525. Given these parameters, the values and confidence levels of the Monte Carlo analysis for the 5th, 10th, 90th, and 95th percentile values are given in Table 5.7.
Because the RASCAL and LHS analyses were done only for the cases when 10 and 20 discrete points or strata were used, it is only possible to estimate the percentile values in increments of five. This can be altered in the RASCAL scheme if unequally weighted intervals are used. It is an important feature of the RASCAL scheme that such a change is possible since this allows very low probability calculations to be made. In the LHS method, the only way to estimate low probability events is to make some assumption about the distributional form of the response PDF. However, for the purposes of uncertainty analysis, it is not necessary to make such estimates since the concern is the range of possible responses. The crucial word is possible, since, theoretically almost any positive value is possible for the corrosion depth. In order to give a reasonable estimate, it is usually sufficient to examine the range between the 5th and 95th percentile values (this range is used since it corresponds to a four standard deviation spread in the normal distribution). Additional discussion of this point is provided in the following section.

Table 5-8 gives the results of the percentile calculations for the DPD-20 and LHS-20 cases, together with the Monte Carlo result,
at year 1,000. As the figures show, the RASCAL over estimates the spread. The RASCAL analysis does predict the 10th and 90th percentile value within the accuracy of the Monte Carlo result while the LHS result only falls within the range of the 95th percentile value. Given the poor performance of the LHS prediction for the other values examined, it is not believed that this result is an indication of anything other than good fortune. Additionally, in the analysis, the response is the corrosion depth and, therefore, an over estimate of the spread represents a conservative result.

On method for gauging the effectiveness of each method's ability to predict the uncertainty is to assume one is interested in estimating the spread between the 10th and 90th percentile values. Table 5-9 gives the predicted values of each percentile obtained from each of the five runs. As can easily be seen from this table, the RASCAL calculation reaches the "correct" value much more quickly than the LHS calculation. The indication is that if the 5th and 95th percentile values are important, then a RASCAL calculation using 40 discrete points would be more appropriate.

<table>
<thead>
<tr>
<th>Percentile Value</th>
<th>Value</th>
<th>Value</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LHS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentile</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>38.07</td>
<td>33.22</td>
<td>49.15</td>
</tr>
<tr>
<td>10</td>
<td>43.45</td>
<td>43.62</td>
<td>50.14</td>
</tr>
<tr>
<td>90</td>
<td>96.16</td>
<td>96.45</td>
<td>88.68</td>
</tr>
<tr>
<td>95</td>
<td>103.9</td>
<td>116.7</td>
<td>104.6</td>
</tr>
</tbody>
</table>

TABLE 5.8 PERCENTILE ESTIMATES FROM EACH OF THE UNCERTAINTY METHODS
TABLE 5.9 PREDICTION OF SELECTED PERCENTILES USING DIFFERENT NUMBERS OF DISCRETE POINTS OR STRATA

<table>
<thead>
<tr>
<th>Method Used</th>
<th>10th Percentile</th>
<th>90th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monte Carlo</td>
<td>43.45</td>
<td>96.16</td>
</tr>
<tr>
<td>RASCAL-10</td>
<td>42.86</td>
<td>108.16</td>
</tr>
<tr>
<td>RASCAL-20</td>
<td>43.62</td>
<td>96.45</td>
</tr>
<tr>
<td>LHS-10</td>
<td>50.06</td>
<td>83.43</td>
</tr>
<tr>
<td>LHS-20</td>
<td>50.14</td>
<td>88.68</td>
</tr>
</tbody>
</table>

The final consideration for recommending one method over another is the amount of computer time taken for each of the analyses. Table 5-10 gives the execution time used for each analysis on a CDC-855 series computer. The actual cost of the runs will vary from these figures because of different input and output requirements, plotting, etc. However, the values in Table 5-9 are expected to represent the fairest comparison of the computational efficiencies. The results indicate that the RASCAL analysis could use up to 45 discrete points before it would become as expensive as the Monte Carlo analysis while the LHS method could use up to 81 strata. However, given the increased accuracy of the RASCAL method at the lower number of discrete intervals examined, i.e., 20, and its greater flexibility to estimate lower percentile values, it is believed that the DPD method is a better method for performing uncertainty analysis.

5.6 Probabilistic and Sensitivity Analysis

The ability of the Monte Carlo analysis to predict classic statistical quantities of interest, e.g., mean, variance, confidence levels, etc., gives it a great appeal. However, it must be
TABLE 5.10  COMPUTER TIME REQUIRED FOR EACH UNCERTAINTY ANALYSIS

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (Seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monte Carlo</td>
<td>572</td>
</tr>
<tr>
<td>RASCAL-10</td>
<td>96</td>
</tr>
<tr>
<td>RASCAL-20</td>
<td>254</td>
</tr>
<tr>
<td>LHS-10</td>
<td>51</td>
</tr>
<tr>
<td>LHS-20</td>
<td>69</td>
</tr>
</tbody>
</table>

recalled that the corrosion model being studied has been greatly simplified and is relatively cheap to evaluate. When more accurate or better physical descriptions become available, it is expected that each evaluation will require a substantial amount of computer time. At that point, it may become prohibitively costly to perform a Monte Carlo analysis even using stratified or importance sampling algorithms in conjunction with the Monte Carlo scheme. Then it will be crucial to have in place a method to not only examine the uncertainty of these new models, but also to use for probabilistic modeling and/or sensitivity studies.

The sensitivity analysis of a function or computer program, as was discussed above, is not overly concerned with the prediction of the PDF of the response but rather is concerned with assessing the variability of the response due to changes in the input variables when these changes are small and near the design points. For such instances, the LHS method provides the best technique for determining the sensitivity, for two reasons. First, the sample space which the LHS scheme uses is relatively small about
the design points; and, therefore, the information obtained will be most accurate near this point. Secondly, the substantially lower cost of performing a LHS analysis makes it very attractive to use whenever possible.

The probabilistic analysis of a system or subsystem should be performed with a Monte Carlo analysis if it is possible from a cost standpoint. If it is not economically possible to use a Monte Carlo method then, since one is interested in determining the PDF of the response, the RASCAL method provides the best alternative of those examined in this study. In using the RASCAL method, one must insure that either enough discrete points are being used to adequately describe the range of the PDF in which there is interest or the intervals must be chosen so that the input PDFs are weighted to cover the region of interest.

5.7 Summary

The inclusion of stochastic variation or uncertainty in parameters or models is best performed with Monte Carlo simulation methods, whether the analysis is probabilistic, uncertainty, and/or sensitivity. If the Monte Carlo is too costly then the preferred methods are RASCAL for probabilistic and uncertainty analysis, and LHS for sensitivity analysis. It was found during this study that the mean corrosion depth could be predicted within 2% of the Monte Carlo results throughout a 1,000 year calculation, using RASCAL at one-half the computer cost. The LHS mean calculation came within 5% of the Monte Carlo prediction at one-fifth the cost. The prediction of the standard deviation, one measure of the uncertainty, showed an even larger discrepancy with the RASCAL method predicting it within 5% while the LHS prediction only came within 30%. Finally, the RASCAL method obtained the same accuracy as the Monte Carlo method in predicting the 10th
and 90th percentile values of the corrosion distribution at 1,000 while the LHS was 10-15% off in its predicted value.
FIGURE 5.3  STANDARD DEVIATION PREDICTED BY EACH UNCERTAINTY ANALYSIS METHOD
6.1 Introduction

The previous chapter presented the results of a study in which the governing physical equations were first order, linear, differential equations. The purpose of this chapter is to examine a physical process which is governed by a first order, nonlinear differential equation. An important problem of this type is the growth of cracks in metals due to fatigue loading. In order to maintain a consistency with the literature, time will be measured in cycles, denoted \( N \), throughout this chapter. Since fatigue loading is the cyclic application of a relatively high load followed by its relaxation, this measure of time makes more sense for the subsequent analysis.

These types of studies are important applications of a growing field usually identified as Probabilistic Fracture Mechanics (PFM). PFM studies have been performed to assess the degree of conservatism in piping designs by Kurth, Leis, and Cox [42], and to assess earthquake loading effects on nuclear pipes by Harris, Dedhia, and Lim [43]. The fracture mechanics model presented in this study, while simple, forms the basis for the fracture mechanics models in each of these studies. Of course, PFM analysis is not only important in the nuclear industry but is also beginning to be applied in the aerospace, petroleum, and chemical industries.
The remainder of the chapter is divided as follows. First, the basic fracture mechanics model is presented. Secondly, the input model applications are given. Finally, a comparison of the models is made.

6.2 Fracture Mechanics Model

The PFM models which currently exist are all based on linear elastic fracture mechanics (LEFM) theory. LEFM theory assumes that the plastic zone, which forms around the crack tip and is the region of the material which has been deformed, is small compared to the crack dimension. As shown in Figure 6-1, the radius of the plastic zone, which forms around the crack tip and is the region of the material which has been deformed, is small compared to the crack dimension. As shown in Figure 6-1, the radius of the plastic zone, \( r_p \), must be much less than the half length of the crack, \( a \), i.e.,

\[ r_p \ll a. \]

For this assumption and using the analysis provided by Irwin [44] and Westergaard [45], it can be shown that the stresses in the vicinity of the crack tip can be expressed

\[
\begin{bmatrix}
0_{xx} \\
0_{yy} \\
\tau_{xy}
\end{bmatrix}
= \frac{K \cos \theta/2}{(2\pi r)^{1/2}} \begin{bmatrix}
1 - \sin \theta/2 \sin 3\theta/2 \\
1 + \sin \theta/2 \sin 3\theta/2 \\
\sin \theta/2 \cos \theta/2
\end{bmatrix} + E(r^0) \quad (6-1)
\]

where \( r \) and \( \theta \) are the polar coordinates of a point with respect to the crack tip, as shown in Figure 6-2. The factor, \( K \), is called the stress intensity factor, and it is a measure of the stress "amplification" at the crack tip.

Examination of Equation (6-1) shows that at the crack tip, i.e., \( r = 0 \), the stresses become infinite. This is a result of LEFM
Figure 6.1 LEFM Theory for a Crack in an Infinite Plate

Figure 6.2 Coordinate System for Stress Calculation
theory and, obviously, in real problems the stresses are not infinite.

In examining crack growth, there are three types of loading considered: tension (opening) denoted Mode I; shear (sliding) denoted Mode II; and out-of-plane shear (tearing) or Mode III. Each mode of loading has its own associated stress intensity factor denoted $K_I$, $K_{II}$, and $K_{III}$. In general, the stress intensity factor can be written

$$K_N = f(\text{geometry})(a)^{0.5}, \quad N = I, II, \text{ or } III$$

where

$$f(\text{geometry}) : \text{factor dependent on the material shape}$$

$$\sigma : \text{stress determined by the loading}$$

$$a : \text{one-half crack length.}$$

For simplicity of discussion, only Mode I loading perpendicular to the crack in a large sheet under uniform tension is considered. In this case, the subscript is deleted since it is understood that $K_I$ is being calculated and

$$K = \sigma (\pi a)^{0.5}$$

$$\sigma : \text{far field stress} \quad (6-2)$$

If the stress is now cycled between an upper, $\sigma_{\text{max}}$, and lower, $\sigma_{\text{min}}$, value, then the change in $K$ per cycle is

$$\Delta K = \Delta \sigma (\pi a)^{0.5} = K_{\text{max}} - K_{\text{min}} \quad (6-3)$$
Paris [47] found that the crack growth per fatigue cycle, \( \frac{da}{dN} \), can be correlated over a wide range of values for \( K \) by

\[
\frac{da}{dN} = C(\Delta K)^n
\]  

(6-4)

This is the Paris law equation for fatigue crack growth. It is the simplest form of LEFI theory yet it contains all of the essential features of calculating stable crack growth.

To use Equation (6-4), Equation (6-3) is substituted and the result is integrated:

\[
\int_{a_0}^{a_f} a^{-n/2} da = \int_{0}^{N_F} C(\Delta a)^n \frac{n}{2} dN
\]  

\[
a_f^{(2-n)/2} = a_o^{(2-n)/2} + \frac{2\pi n/2}{2-n} C(\Delta a)^n N_F
\]  

(6-5)

Equation (6-5) is the predictive equation to be used in the subsequent studies. Before proceeding to the probabilistic portion, it is necessary to examine this equation in more detail.

Assume that a critical crack size which will cause failure, denoted \( a_c \), is known. Then, rather than integrating from \( a_0 \) to \( a_f \) and solving for \( a_f \), it is more useful to integrate from \( a_0 \) to \( a_c \) and solve for \( N_F \). Thus,

\[
N_F = (a_c^P - a_o^P)/(p \pi n/2 C(\Delta a)^n)
\]  

(6-6)

\[
p = (2-n)/2
\]
For example, suppose

\[\begin{align*}
  a_c &= 0.5 \\
  m &= 4.0 \\
  c &= 9.11 \times 10^{-11}
\end{align*}\]

then \(N_F = 72,600\) cycles. This is an important consideration in performing since it would not make any sense to carry calculations out to, say 250,000 cycles, since the material would have already failed.

The second consideration arises from asking what is the crack size for \(N_F\) equal to 250,000 cycles with all other parameters the same. The answer is that \(a_f\) is equal to -237.311! Obviously, a negative crack size has no meaning. The solution comes from examining Figure 6-3. At some point, defined as \(N_c\), the crack growth becomes unstable and goes to infinity. After cycle \(N_c\) is reached, then the crack size is forever negative. It must be noted that this is because of the fortuitous choice of \(m=4\). If \(m\) is equal to 6, then the crack size to the positive side of \(N_c\) decreases from positive infinity to zero. However, the critical point can always be determined from

\[
P = \frac{(m - 2) a_o}{2 \pi m/2_c (\Delta \sigma)^m}, \quad P = (2 - m)/2, \quad m > 2 \quad (6-7)
\]

6.3 Random Variable Definitions

In Equation (6-5), there are three random variables which should be included in the analysis. First is the initial crack size denoted by \(a_o\). Second is the stress range, \(\Delta \sigma\). Finally, the parameter \(C\) is allowed to vary while \(m\) is set equal to 4.0 to account for material property variations as discussed by Harris [47].
Figure 6.3 Stability Region for Crack Growth
It is standard practice in the nuclear industry to assign lognormal distributions to almost all variables. In this study, Rayleigh distributions are used to describe the variables. The PDF is given by

\[ f(x) = \frac{x}{\delta^2} e^{-x^2/2\delta^2} \]

and the CDF is

\[ F(x) = -e^{-x^2/2\delta^2} \]

For these studies, shifted Rayleigh distributions are used since none of the three variables is ever allowed to reach zero. Thus,

\[ f(x) = \frac{(x - 1)^2}{(n - 1)} \frac{1}{e^{\frac{1}{2} (x - 1)^2}} \]

\[ F(x) = -\frac{1}{(n - 1)^2} e^{\frac{1}{2} \frac{(x - 1)^2}{(n - 1)^2}} \]

The parameters for the Rayleigh distribution for each variable is given in Table 6-1.

Finally, it should be noted how the critical cycle number, \( N_c \), is handled when \( a_0 \), \( c \), and \( \sigma \) have been set. For a general value of \( m \), it is necessary to use Equation (6-7) to calculate \( N_c \) and if the current cycle number is greater than \( N_c \), then \( a_f \) is set to a very large number or the critical crack size \( a_c \). However, since \( m = 4 \), Figure 6-3 gives the behavior of Equation (6-5) and at the end of the calculation of \( a_f \), if it is less than zero, it is set equal to \( a_c \).
### TABLE 6-1. RAYLEIGH PARAMETERS FOR PFM ANALYSIS

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_0)</td>
<td>0.005</td>
<td>0.01</td>
</tr>
<tr>
<td>(c)</td>
<td>(1.00 \times 10^{-14})</td>
<td>(9.11 \times 10^{-12})</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>10.0</td>
<td>35.0</td>
</tr>
</tbody>
</table>

6.4 PFM Analysis Results

The results of the mean and standard deviation calculation for each of the probabilistic methods is shown in Table 6.2. The Monte Carlo result represents a 10,000 run simulation of Equation (6-5). The runs identified as LHS-10 represent the LHS analysis using 10 discrete strata are used. Similarly, the results labeled LHS-20 are from the LHS analysis when 20 strata are used, LHS-30 when 30 strata are used and so on. The analogous labels are used for the RASCAL results. Additionally, in the RASCAL analysis, 150 points per interval were used. Therefore, the RASCAL-10 run calculated 1,500 crack sizes, the RASCAL-20 run 3,000; and so forth. Finally, the number in parentheses represents the computer time used on a CDC-6500 machine.

The results show that the LHS analysis is extremely poor. This is because only 0.5% of the points used in the Monte Carlo analysis are being used to estimate the mean and standard deviation. Because of the large deviations from the Monte Carlo results, which are also to represent "truth", the LHS analysis is not discussed further. The RASCAL results show that when 20 discrete points are used, the mean values are predicted within 4% of the Monte Carlo results.
TABLE 6.2. RESULTS OF PFM ANALYSIS BY EACH PROBABILISTIC TECHNIQUE

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (Cycles)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2000</td>
<td>4000</td>
<td>6000</td>
<td>8000</td>
<td>10,000</td>
</tr>
<tr>
<td>LHS-10 (1.2)</td>
<td>.012</td>
<td>.004</td>
<td>.061</td>
<td>.146</td>
<td>.014</td>
</tr>
<tr>
<td>LHS-20 (2.7)</td>
<td>.012</td>
<td>.004</td>
<td>.036</td>
<td>.011</td>
<td>.014</td>
</tr>
<tr>
<td>LHS-30 (5.5)</td>
<td>.012</td>
<td>.004</td>
<td>.013</td>
<td>.006</td>
<td>.013</td>
</tr>
<tr>
<td>LHS-40 (9.6)</td>
<td>.012</td>
<td>.005</td>
<td>.012</td>
<td>.003</td>
<td>.012</td>
</tr>
<tr>
<td>LHS-50 (13.0)</td>
<td>.011</td>
<td>.003</td>
<td>.012</td>
<td>.004</td>
<td>.021</td>
</tr>
<tr>
<td>RASCAL-10 (7.1)</td>
<td>.011</td>
<td>.003</td>
<td>.011</td>
<td>.003</td>
<td>.011</td>
</tr>
<tr>
<td>RASCAL-20 (13.8)</td>
<td>.011</td>
<td>.003</td>
<td>.011</td>
<td>.003</td>
<td>.012</td>
</tr>
<tr>
<td>RASCAL-30 (20.7)</td>
<td>.011</td>
<td>.003</td>
<td>.011</td>
<td>.003</td>
<td>.012</td>
</tr>
<tr>
<td>RASCAL-40 (28.9)</td>
<td>.011</td>
<td>.003</td>
<td>.011</td>
<td>.004</td>
<td>.012</td>
</tr>
<tr>
<td>RASCAL-50 (36.9)</td>
<td>.011</td>
<td>.003</td>
<td>.011</td>
<td>.004</td>
<td>.012</td>
</tr>
<tr>
<td>Monte Carlo (52.9)</td>
<td>.011</td>
<td>.007</td>
<td>.012</td>
<td>.011</td>
<td>.012</td>
</tr>
</tbody>
</table>

Going to 50 discrete points improves the mean value calculation to within 2% of the Monte Carlo result. Figure 6.4 depicts how the mean crack size production becomes more accurate as more discrete points are added. It is obvious from this figure
that the general trend is to approach the Monte Carlo result although at different points in time the convergence to the "true" result as predicted by Monte Carlo proceeds differently. Because of the nonlinear nature of the problem and the possibility of the crack size exceeding the computer limit for large numbers as time increases, there will be significantly larger scatter in what the predicted mean value is.

Figure 6.5 depicts the results of the standard deviation calculation by the Monte Carlo and RASCAL methods. Whereas the RASCAL method provides a reasonable estimate of the standard deviation and does show the general trend of the Monte Carlo result of the previous chapters. There are two considerations: (1) the crack growth calculation is a nonlinear problem and (2) the input variables were discretized into equal probability intervals.

The effect of the use of the unequal probability intervals is shown in Figure 6.6. Obviously, the range of the crack size is predicted much more accurately; and, therefore, a significant improvement in the standard deviation has been made.

At this point, it has been shown that the RASCAL method can be made to work well for linear and nonlinear problems when the number of variables is relatively few. The remaining question is: can an algorithm be derived for smoothing and optimizing the RASCAL discretization of the input variables? This question is answered in Chapter 7 where an optimization algorithm for the RASCAL technique is presented.
Figure 6.4 Mean Crack Size Versus Time
Figure 6.5 Standard Deviation of Crack Size Versus Time
Figure 6.6 Empirical CDF of Crack Size
CHAPTER 7

THE OPTIMIZATION OF THE RASCAL ALGORITHM

7.1 Introduction

In the previous three chapters, the Monte Carlo, LHS, and RASCAL methods of analysis were compared for three different physical problems. In all cases, the LHS design fared poorly. In each of the analyses, the RASCAL method always provided excellent agreement with the mean calculation but only good to poor agreement with the standard deviation. Additionally, it was found that increased accuracy could be obtained with the RASCAL method if more points were placed in the tails of the distribution but only up to an undetermined limit. If too many points were placed in the tails, then the accuracy decreases.

The purpose of this chapter is to examine the RASCAL algorithm mathematically and derive a procedure for choosing the optimum discretization scheme for the input variables. In order to do this, the next section provides a more detailed description of the RASCAL method followed by the optimization procedure. Finally, a summary of the results is given.

7.2 RASCAL Description

The DPD method provides a simple algorithm for performing a probabilistic calculation, however, as pointed out in the previous discussion, the number of calculations needed to characterize
the output, or response, DPD quickly becomes unwieldy. Consider the case in which a fracture mechanics analysis is being performed with four random variables each one of which is discretized into 50 points. This implies that the output final crack size vector is characterized by $6.25 \times 10^6$ points. Clearly, not all of these points are statistically significant. Recalling the Monte Carlo description, the final crack size vector can be characterized accurately by at least three orders of magnitude less points. Thus, the problem is posed: How can the sample space represented by the DPD of the response be sampled to reduce the number of calculations to be performed?

To focus the discussion, let's examine the Figure 7-1 and 7-2. In the first figure, a CDF for a random variable is shown along with a five point discretization. If a Monte Carlo analysis is being performed, then the random number, shown as $r$ in the figure, would be used to determine the value of the random variable denoted as $x_r$. In the DPD analysis, the value is forced to assume the interval mean value, in this case, $x_m$. Therefore, an error between the Monte Carlo analysis and the DPD analysis, call it $e_i$, is introduced and is defined as

$$e_i = (x_{r_1} - x_m)$$

where $x_m$ is the mean value of the discrete interval into which $x_{r_1}$ falls.

Examining the second figure in which 20 discrete intervals have been used, we see that the maximum error, $e_i$, is substantially reduced from the 5 point discrete case. Obviously, in the limit as the number of data points approaches infinity, the DPD algorithm becomes the Monte Carlo algorithm. Thus, if the set of all possible discrete responses from the DPD algorithm is denoted as $D$ and the possible Monte Carlo responses as $\Omega$, then it is clear that $D$ is a subset of $\Omega$. The number of points in $D$ is
Figure 7.1 Five Point Discretization of X

Twenty Point Discretization of X
finite, and in fact, is equal to $N_D^{N_V}$, whereas there are an infinite number of points in $\Omega$. However, in any Monte Carlo analysis, only a subset of $\Omega$, which is finite, will be used to estimate the true statistical properties of $\Omega$. Normally, the size of this subset is on the order of $10^3$ to $10^4$. Thus, the RASCAL algorithm draws enough samples from the discrete space to be of a similar size to the Monte Carlo analysis if similar accuracy is desired. Therefore, if $N_D^{N_V}$ is less than $10^3$, the standard DPD algorithm can be used. If not, then the following algorithm is used.

Given that there are $N_D$ discrete points for each of the $N_V$ random variables in the analysis, divide the input PDF into $N_D$ intervals of equal probability, i.e., $P_i = 1/N_D$ for all $i$. Generate $N_V$ random integers, denoted $I_k$, between 1 and $N_D$. For the $j$th random variable, select $I_j$ discrete point from its corresponding DPD. These set of points define the first response. If the final crack size, $a_f$, is a function of the Paris parameter, $c$, the load, $\sigma$, the initial crack size, $a_0$, and the load ratio, $r$,

$$a_f = f(a_0, c, \sigma, r)$$

then

$$V_1 = f(A_0(I_1), C(I_2), S(I_3), R(I_4))$$

and

$$P = (1/N_D)^4$$

where $A_0$, $C$, $S$, and $R$ are the DPDs for $a_0$, $c$, $\sigma$, and $r$, respectively; $V_1$ is the first value obtained for the final crack size and
P is the probability that the crack size is equal to $V_1$. Because each input DPD is divided into the same number of intervals and equal probability intervals are used, $P$ is invariant. The entire process is repeated, say 1,000 times. Thus, the DPD for the response, which is the final crack size vector, is approximated by

$$A_F = (V_1, V_2, \ldots, V_{1,000}; (1/N_D)^4) .$$

The reason that the DPD algorithm requires the use of all possible combinations is to insure that the response DPD, $A_F$, for the example, remains a PDF. Assume that $N_D$ is equal to 5. In order for $A_F$ to be a discrete PDF, the following must be true:

$$S_p = \sum_{i=1}^{1,000} (1/N_D)^4 = 1 .$$

For $N_D$ equal to 5, $S_p$ is equal to 1.6. However, for $N_D$ equal to 6, $S_p$ is approximately 0.772. Since the input DPDs are uniformly sampled, the probability is simply renormalized. For example, if the probability of each discrete value of the response is $P$, then a PDF is obtained by simply dividing $P$ by $S_p$. Therefore, this algorithm produces the following approximation for $a_F$:

$$A_F = \{V_1, V_2, \ldots, V_{1,000}; [((1/N_D)^4/(\sum_{i=1}^{1,000} 1/N_D^4))]} .$$

This is statistically correct since the "missing" probability mass was uniformly removed from all parts of the input PDFs. To summarize the statistics, the same condensation procedure described for the DPD method in Chapter 2 is used. Since random sampling of the discrete representations of the input variables and a condensation procedure is used, the method is denoted RASCAL (RAndom Sampling Condensation ALgorithm). The subset
of \( D \) from which the \textit{RASCAL} results are obtained is denoted \( R \). If the random sampling scheme were used regardless of the number of responses being calculated, then

\[
\lim_{N^R \to \infty} R = D
\]

where \( N^R \) is the number of responses is the only statement which can be made. For example, if there are 20 discrete points for each of the four variables in the analysis, then even if 20\(^4\) response are generated, there is still a \( 1/e \) (approximately 0.368) chance that any specific value of \( D \) will not be contained in \( R \). Even if \( N^R \) is increased by a factor of 5, there is a 0.7\% probability that any specific value of \( D \) will not be contained in \( R \) yet \( R \) still remains a subset of \( D \), i.e., no value of \( R \) is not also in \( D \). Therefore, the further restriction is placed on the \textit{RASCAL} algorithm that it is used only when \( N^R \) is less than \( N_D N^D / 2 \). This insures that the calculation time can never exceed that of the DPD algorithm and once the number of points in the \textit{RASCAL} method is within a factor of 2 of the DPD method, then the resulting loss of accuracy from replicate discrete responses does not justify the relatively small reduction in computer time.

### 7.3 Optimization Algorithm

In Chapters 3 through 5, indications were given that equal probability interval discretization of the input variables did not provide the most accurate answer if the Monte Carlo result is accepted as truth. The previous section described one method of measuring the accuracy which will be used for the remainder of this discussion. If a variable is truly continuous and accurately described by a continuous PDF, denoted \( f \), and its associated
CDF, \( F \), then
\[
e(r) = |F^{-1}(r) - X(J)|
\]
\[
J = \text{INT}(N_D + r) + 1
\]
\[
\text{INT}(Z) = \text{integer part of } z
\]
\[
r = \text{uniformly distributed random variable between 0 and 1}
\]
\[
X(J) = \text{DPD for input variable } X.
\]

As Figures 7-1 and 7-2 indicate, the error \( e(r) \) must approach zero as \( N_D \) goes to infinity. However, when \( N_D \) is finite, how are the \( X(J) \) selected so that
\[
E = \max[e(r)]
\]
is minimized? Clearly, this is minimized when
\[
X(J + 1) = X(J) + \Delta
\]
for all \( J \). In this case, all of the intervals are equal and thus \( e \) is invariant. For distributions which are unbounded, e.g., the normal, the determination of \( \Delta \) is not possible, i.e., \( \Delta \) is infinite. In these cases, artificial bounds are supplied. For example, Table 7.1 shows the optimum interval size for a Rayleigh distribution, \( \sigma = 1.0 \), in the first column. The maximum interval size for the equal probability interval scheme with the location in terms of the discrete interval in which the maximum occurs is shown in the second and third columns. The last two columns give the similar information for the minimum interval size. As can easily be seen from this table, the maximum error size, \( e \), can range from a factor of 1.6 times greater than the optimum interval case to a factor of 7.6. However, it is still desired to see how the error changes throughout the distribution range. This is shown in Figure 7-3. To obtain both the data in Table 7.1 and Figure 7-3, it is assumed that the upper bound of the distribution is equal to \((1.0 - 1.0/N_D)\).
Figure 7.3 Comparison of Interval Discretization Schemes
This is done in order to make the comparison fair since the equal probability interval discretization is bounded above by the \((N_D - 1)/N_D\) percentile value.

Examining Figure 7-3, an explanation of the results seen in the previous three chapters can now be provided. Since the RASCAL method is one-half of the interval size, it is seen that the mid-range of the PDF is represented very accurately when equal probability intervals are used. Specifically, for the 20 interval case shown in Figure 7-3, the error is less than the optimum intervals size\(^1\) from about the 11th to 79th percentile.

\[\text{TABLE 7.1. ERROR ASSOCIATED WITH DIFFERENCE DISCRETIZATIONS FOR RASCAL}\]

<table>
<thead>
<tr>
<th>Number of Discrete $N_D$</th>
<th>Optimum Scheme Interval Size</th>
<th>Equal Probability Interval Size</th>
<th>Maximum Size</th>
<th>Maximum Location</th>
<th>Minimum Size</th>
<th>Minimum Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>.35882</td>
<td>.66805</td>
<td>4</td>
<td>.34272</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.21460</td>
<td>.45904</td>
<td>9</td>
<td>.16617</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.12239</td>
<td>.32029</td>
<td>19</td>
<td>.08256</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>.05594</td>
<td>.25988</td>
<td>1</td>
<td>.03298</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>.03035</td>
<td>.23770</td>
<td>1</td>
<td>.01649</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Values which are less than the 11th percentile or greater than the 79th percentile value introduce errors which increase expo-

\(^1\)Recall optimum means that the maximum error is minimized. Therefore, other schemes can be found which will have lower errors at some point in the range but not throughout the range.
nentially away from these values. For a monotonically increasing function (which all of the examples in this study are) this implies that the response is very accurately represented in the mid-range but poorly represented as the tail values are approached. Therefore, while the mean values are found very accurately with the RASCAL method, the variance (or measure of the spread) was relatively inaccurate. As the input variable DPD was forced to cover a wider range of the distribution, the answers became more accurate provided that not "too much" of the probability mass was moved from the lower tail to the upper tails.\(^2\) The reason for this result now becomes obvious. If too large of a probability is placed in the upper tails to increase accuracy, then the error in the lower and mid-range values becomes poor. For example, with a 5 discrete point case, if the 90th, 95th, and 99th percentile values are used (i.e., 2.146, 2.448, and 3.035, respectively) then only 2 points remain to describe the rest of the distribution. The best that can be done is to have two equal intervals of approximately 1.073 increments. Therefore, the maximum error is approximately 60\% greater than equal probability interval discretization case and 200\% greater than the optimum case while there is only one value less than the equal probability discretization and it is only 10\% less while it is only 15\% less than the optimum interval size. On the other hand, it does cover the three values not covered by the previous calculations. This can be compared to the optimum interval case by calculating \( \sigma \) for the Rayleigh distribution by

\[
\sigma = \left[ (-2 \ln(100 - P)/100) \right]^{1/2}/N_D
\]

\(^2\)Recall from basic probability theory that the probability mass must be conserved. Thus, if 10 discrete points are used, each having equal probability, then if it is desired to use a point having probability 0.05, the remaining probability of 0.5 must either be divided among the remaining points or another point must be added to account for this probability mass.
where $P$ is the percentile value of concern. In this case, the error is still 77% greater than the optimum interval error.

7.4 Summary

These calculations have shown that, while the initial inclination may be to adopt an equal probability interval rule for discretizing input variable distributions, this scheme is not optimum. Optimum in this context means the minimization of the maximum value of the sampling error which can be attained during the discrete calculations. The optimum interval choice is defined by

$$X(J) = X(J - 1) + \frac{[F^{-1}(P_U) - F^{-1}(P_L)]}{N_D}$$

$P_U$, $P_L$: upper and lower percentile bounds
$F$: CDF of $X$.

Once the interval endpoints have been so defined, the probability mass is lumped at the conditional mean of each interval. This procedure will remove artificial shifting of the distribution (by using the conditional means) and minimize sampling error (by using equal discrete space intervals).

To demonstrate the tremendous improvement available with the use of the optimum interval scheme, the PFM analysis was repeated for the 20 and 50 discrete point runs where the intervals are determined by the optimization scheme with the 99.9 percentile value used as the upper limit. The results of these runs are compared to the previously obtained Monte Carlo analysis results in Table 7.2. While the scatter for the first cycle increment is poor, the remaining results show very good agreement with the Monte Carlo result. Figure 7-4 plots these results for the standard deviation and when this figure is compared to Figure 7-3, dramatic increases in the accuracy are seen. Therefore,
the optimization scheme which has been derived in this chapter, works well in improving the accuracy of the RASCAL results.

**TABLE 7.2. PFM RESULTS USING THE OPTIMIZED RASCAL METHOD**

<table>
<thead>
<tr>
<th>Cycle</th>
<th>2,000</th>
<th>4,000</th>
<th>6,000</th>
<th>8,000</th>
<th>10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.01161</td>
<td>.01207</td>
<td>.01276</td>
<td>.01391</td>
<td>.01584</td>
</tr>
<tr>
<td></td>
<td>.00777</td>
<td>.01121</td>
<td>.01891</td>
<td>.02723</td>
<td>.03973</td>
</tr>
<tr>
<td></td>
<td>.01149</td>
<td>.01180</td>
<td>.01254</td>
<td>.01408</td>
<td>.01545</td>
</tr>
<tr>
<td></td>
<td>.00345</td>
<td>.07809</td>
<td>.01632</td>
<td>.02767</td>
<td>.03592</td>
</tr>
<tr>
<td></td>
<td>.01149</td>
<td>.01192</td>
<td>.01224</td>
<td>.01346</td>
<td>.01522</td>
</tr>
<tr>
<td></td>
<td>.00337</td>
<td>.00461</td>
<td>.01317</td>
<td>.02407</td>
<td>.03395</td>
</tr>
</tbody>
</table>
Figure 7.4 Standard Deviation of Crack Size Using Optimized RASCAL Method
CHAPTER 8
SUMMARY AND RECOMMENDATIONS FOR FURTHER RESEARCH

The ultimate goal of this dissertation was the development and application of a new probabilistic method for performing risk analysis calculations in the nuclear industry. The RASCAL method was developed and incorporates the following features:

(1) It is a discrete method
(2) It is capable of analyzing either discrete or continuous data
(3) Importance sampling can be performed without programming
(4) It reproduces the Monte Carlo results with reasonable accuracy at a fraction of the cost.
(5) It has been benchmarked against the Monte Carlo and LHS methods in three applications and was found to perform as well or better than either of these standard techniques.

During the applications the RASCAL method was found to work very well in the initial study to determine the uncertainty in the top event probability of a fault tree. The second study, to predict the corrosion depth in a nuclear waste canister, is a linear differential equation and the RASCAL method did a good job of predicting the uncertainty in the corrosion depth. The final analysis, the fatigue crack size predictions, found that the RASCAL method did not perform as well, relative to the previous analyses, and that many more discrete points were needed to obtain agreement with the expensive Monte Carlo analysis. In fact, only about a factor of two reduction of the Monte Carlo computational time could be achieved with the RASCAL method. It was noted that the fatigue crack growth model is a nonlinear differential equation. To determine the cause of the RASCAL method losing accuracy for the nonlinear problem an investigation
of the error introduced by the discretization of the input variables was made in the previous chapter. It was found that the use of equal probability intervals caused the error to be greatest in the tails of the variable distributions. Because of the nonlinear growth nature of the cracks this error was magnified as time increased. A scheme for optimizing the interval selection process was derived and subsequently applied. It was found that the use of the optimized interval scheme restored the advantages of the RASCAL method even for this nonlinear problem.

There are several important topics which still should be addressed in future research efforts. This research demonstrated that the RASCAL method works well for linear and nonlinear differential equations when all of the variables are statistically independent. The first question which comes to mind is: can RASCAL be modified to handle correlated or dependent variables? The important issue of showing that the RASCAL method can approximate the "true" results reasonably well was documented during this research by comparison to a Monte Carlo analysis. However, can confidence limits be estimated with the RASCAL method so that an estimate of the accuracy of the results and potential improvements in the accuracy can be made based only on the number of intervals and the number of points to be contained in each interval?

As a matter of convenience this work always used the same number of discrete intervals for each variable in the analysis, although it is not required by the algorithm. There are several important questions which this raises: Are there compelling reasons to change to unequal numbers of points; clearly the number of arithmetic operations which must be performed will decrease, however does the increased complexity of the logic outweigh this advantage?; How does the optimization of discrete intervals change when not all variables have an equal number of intervals?; Does the
lack of accuracy of a coarsely defined variable transmit to a very finely divided variable?.

A first step has been taken in this work to demonstrate the ability of the RASCAL method to perform probabilistic analysis in nuclear risk related problems. The RASCAL method is equally applicable to nonnuclear probabilistic calculations. In order to demonstrate the general applicability of the method these and other questions must be addressed.
There are several important topics related to the selection and use of probabilistic models which are not, of themselves, truly probabilistic methods but rather are statistical methods. Three such topics are discussed in this Appendix: (1) Data Analysis, (2) Fitting Distributions to Data, (3) Nonstandard Statistical Analysis.

### Statistical Analysis

There are four primary functions of a data analysis effort. The first is forming data sets based on expert opinion pooling. The second is methods for combining data with opinions or including new data with previous data sets. Regression analysis of data is important in many statistical analyses. And, finally, a method for fitting distributions to data must be available. Each of these topics is discussed below.

**Expert Opinion Polling.** The high costs and logistical difficulties associated with the procurement of pertinent experimental data many times make it necessary to rely on expert opinion in the absence of sufficient quantitative data. In developing the generic simulated load models, the experimental data available will fall into one of three categories:

1. No quantitative data (qualitative)
2. Sparse quantitative data
(3) Sufficient quantitative data.

If no quantitative experimental data is available, expert opinion will be the only source of quantitative information. Methods of obtaining and synthesizing expert opinion include Delphi method and a new method developed recently, denoted the multifactor method. These methods are described briefly.

If sparse quantitative experimental data are available, it will be necessary to supplement the data with expert opinion. Expert opinion will be obtained as for Category 1 and then combined with the experimental data to develop the simulated load model. Methods for combining expert and experimental information include simple weighting schemes and Bayesian analyses. Under Category 3, expert opinion polling would not be necessary.

The proposed method for obtaining and synthesizing expert opinion in this program is a combination of the Delphi method and the Battelle method. The proposed method is described briefly later.

The Delphi Method. The Delphi method was developed in the early 60s at the Rand Corporation. Although the method is heavily used in the pure forecasting area, Delphi is finding application in various areas where there is a need to incorporate subjective information into a quantitative model so that a group of experts may contribute information to the model.

The basic data-presentation and data-collection technique of the conventional Delphi method is typically a pencil and paper questionnaire generated by a director or a small monitor team. An initial questionnaire is used to present baseline information to the participants as well as to obtain information from them. Successive questionnaires summarize the results of earlier questionnaires and work towards convergence of opinion and a refinement of information. This iterative procedure of questionnaires with feedback is continued until a consensus is reached or no
further progress in this direction is being made.

The information obtained from the Delphi exercise is then refined further by the director or monitor team. The final produce is a quantitative model which reflects the consensus of expert opinion.

The Multifactor Method. A method has been developed for extracting and synthesizing expert opinion in the design of multifactor experiments. This method is equally applicable to the development of multifactor models.

With the conventional Delphi, participants do not discuss issues face to face and may be geographically remote from one another. The multifactor method, on the other hand, encourages face to face contact. A team of experts which includes a team statistician is assembled for the purpose of developing a model. Specific administrative agendas are used to guide the extraction, synthesis, and documentation of expert knowledge. Several statistical tools, including factorial tables, hierarchical trees, and their mathematical representations, are used by the team statistician to guide the generation and analysis of the numerical predictions required of team scientists.

Agenda #1 consists of carefully identifying the dependent variable and the physical factors that effect these dependent variables. Agenda #2 calls for a group consensus on suitable ranges for the physical factors identified in Agenda #1. A complete factorial table is formed consisting of low and high values for each factor, and each scientist then provides a prediction of the numerical value of the dependent variable that would result from each combination of levels of the physical factors and specific evidence supporting the predictions.

Agenda #3 consists of a statistical analysis of each team member's
predictions by the team statistician. The results are represented in the form of a hierarchical tree so that relationships between the physical factors and the dependent variable may be graphically displayed. Each member is allowed to revise his predictions until the hierarchical tree reflects the scientist's views concerning the relationships among the physical factors and the dependent variable. Finally, Agenda #4 calls for a synthesis of the team's predicted values, a review of the evidence supporting the predictions, and model selection. The model selected must be consistent with the synthesized predicted values previously obtained.

Sparse Data Analysis

In many engineered systems, one often encounters the situation where a relatively few number of data points have been obtained because of the cost and/or complexity of experimentally measuring such parameters. Additionally, throughout the system's service life, it is possible many times to obtain more data. Two problems must be addressed. First, given relatively few data points, how does one characterize the statistical properties of the data in the most accurate manner? Second, how is additional data, obtained after the original analysis, factored into new results?

The only reasonable, mathematically based approach is a Bayesian analysis. This method is briefly described below; however, it is assumed that the reader has some familiarity with probability theory.

Let X be a random variable representing the load distribution obtained from the expert opinion with the vector \( x = (x_1, x_2, ..., x_n) \) being \( n \) statistically independent observations of \( X \) (the sample data). There is also assumed to exist a vector of parameters denoted \( a = (a_1, a_2, ..., a_k) \) describing the vector
The joint prior PDF of $A$ (the prior model describes the current state of knowledge about $a$) is given by $g(a)$ while $f(x_i|a)$ is the conditional PDF at $x_i$ given $a$. The joint conditional PDF of $X$ given $a$, $f(x_i|a)$ is calculated by:

$$f(x|a) = \prod_{i=1}^{n} f(x_i|a) \cdot$$

This PDF represents the probability distribution of the load given that the parameters of $a$ are correct. The marginal PDF of $X$ is denoted $f(x)$ while the (joint posterior) PDF of $A$ given $x$ (the posterior model describes the state of knowledge after new data is added) is $g(a|x)$. Bayes' theorem states:

$$g(a|x) = \sum_{i=1}^{n} [f(x_i|a)] g(a)/f(x)$$

$$g(a|x) = f(x|a) g(a)/f(x)$$

where

$$f(x) = \int f(x|a) g(a)da : \text{a continuous}$$

$$\sum f(x|a) g(a) : \text{a discrete.}$$

Given the sample data $x$, $f(x|a)$ may be regarded as a function, not of $x$, but of $a$. When so regarded, we refer to the likelihood function of a given $x$, and write $L(a|x)$ to provide a distinct interpretation apart from $f(x|a)$. If we regard $f(x|a)$ in Bayes' theorem as the likelihood function, then

$$g(a|x) = g(a) L(a|x)$$

which states that the posterior distribution is proportional.
to the product of the prior distribution and the likelihood function. The constant of proportionality needed to insure that the posterior distribution integrates to one, is the marginal distribution of $X$. Therefore,

$$f(a|x) = \frac{g(a) L(a|x)}{\int f(x|a) g(a) da}.$$  

At this point, we can demonstrate how Bayes' theorem can be used during sparse data set analysis. To illustrate, suppose that we have an initial set of sample data $x_1$. For sparse data set analysis, this will be the result of the qualitative data assessment described previously. In addition, the prior distribution of $a$ is also known. Bayes' theorem states

$$g(a|x_1) = g(a) L(a|x_1).$$

Suppose a second set of sample data, i.e., the sparse data set $x_2$ is obtained. Since $x_1$ and $x_2$ are statistically independent,

$$g(a|x_1, x_2) = g(a) L(a|x_1) L(a|x_2) = g(a|x_1) L(a|x_2).$$

It is observed that the posterior distribution of the expert opinion data set now assumes the role of the prior distribution when the sparse data set is combined with expert opinion. This process can be repeated each time additional data is obtained. Through this process, knowledge about the parameters can be continually updated.

Quantitative Data Analysis. When sufficient data is available during the proposed program, at least one of two kinds of analysis are necessary, statistical analysis and distribution fitting.
The latter topic is discussed in the following subsection. This discussion will focus on how the optimum value of empirical parameters are best determined from a statistical standpoint. Standard statistical analysis is not discussed since previous experience in developing empirical relationships for experimentally determined data has indicated that standard statistical analysis packages are being used in many cases in which the basic assumptions made for the statistical method is violated. Therefore, a brief description of a nonlinear maximum likelihood technique is presented.

**Fitting Distributions to Data.** Given data (or a histogram summarizing data) generated by either N experimental tests or Monte Carlo simulations, the problem is to fit a suitable analytic distribution. For the purpose of illustration, formulas for three analytic distributions will be developed here: Normal, Lognormal, and Uniform. Formulas for additional distributional families are available, however, the inclusion of such formulas would only encumber the discussion.

The basic fitting procedure is moment matching, though, in some cases, matching of percentiles works better. Let X be a random variable with a probability density (PDF) denoted f(X), write X ~ f(x). The moments of X are

\[ \mu_i = E(X^i) = \int_{-\infty}^{\infty} x^i f(x) \, dx \quad i = 1, 2, 3, \ldots \]

while the central moments are

\[ m_i = E((X - \mu)^i) = \int_{-\infty}^{\infty} (x - \mu)^i f(x) \, dx, \quad i = 1, 2, 3, \ldots \]

Given a set of data \( X_1, X_2, \ldots, X_N \), sample statistics furnishing
estimates of the moments and central moments can be calculated by

\[ \mu_i = \frac{1}{N} \sum_{j=1}^{N} X_j^i \quad m_i = \frac{1}{N} \sum_{j=1}^{N} (X_j - \mu_i)^i \]

In fitting distributions to the data, it is sometimes necessary to shift the theoretical distribution by an amount \( \Delta \), i.e., to replace \( X \) by \( X + \Delta \).

The shifted density is \( f(X - \Delta) \), but there is no change in the central moments. As an example, the moments of three distributional families are given without derivation. Other distributional types are available, including Rayleigh and Weibull.

**Uniform.** For a uniform distribution between \( a \) and \( b \):

\[ \mu_1 = \frac{a + b}{2}, \quad m_2 = \frac{(b - a)^2}{12} \]

Thus,

\[ a = \mu_1 - \sqrt{3m_2}, \quad b = \mu_1 + \sqrt{3m_2} \]

**Normal.** Form a normal distribution with mean \( \mu \) and variance \( \sigma^2 \):

\[ \mu = \mu^1, \quad \sigma^2 = m_2^2 \]

**Lognormal.** For the unshifted case,

\[ \mu = e^{\frac{1}{2}\sigma^2}, \quad m = e^{\mu + \sigma^2} (e^{\sigma^2} - 1) \]
Thus,

\[ \mu = \ln \left( \frac{(\mu_1)^2}{\sqrt{\mu_2}} \right) \]

\[ \sigma = \ln \sqrt{\frac{\mu_2}{(\mu_1)^2}} \]

since

\[ \mu_2 = m_2 + (\mu_1)^2 \]

For the shifted case

\[ \Delta + e^{\mu + 1/2\sigma} = \mu_1 \]

\[ e^{2\mu + \sigma^2} (e^{\sigma^2} - 1) = m_2 \]

\[ e^{3\mu + 3/2\sigma^2} (e^{3\sigma^2} + 2) = m_3 \]

These equations do not have an explicit solution; it must be found numerically. It can be shown that if \( e^2 \) is greater than one, i.e., if \( m_3^* \) is greater than zero, then a unique solution must exist. It is known that the estimation of the moments of a lognormal by Monte Carlo simulation can be difficult. In these cases, the fitting of a lognormal can be done using selected percentiles.

In choosing the distributional family which should be used during the fitting procedure, higher order moments of the sample are used to guide the choice. Defining the coefficient of skewness, \( b_1 \), and the coefficient of kurtosis, \( b_2 \), as
\[
b_1 = \frac{m_3}{(m_2)^{3/2}}
\]
\[
b_2 = \frac{m_4}{(m_2)^2}
\]

Caution must be exercised in reading the literature because these definitions are not standardized. Skewness and kurtosis are dimensionless quantities, invariant under transformations of the form \(cX + d\) where \(c\) is greater than zero. The values of \(b_1\) and \(b_2\) are shown below for the three distributions used previously.

**Uniform:** \(b_1 = 0\) \hspace{1cm} \(b_2 = \frac{9}{5}\)

**Normal:** \(b_1 = 0\) \hspace{1cm} \(b_2 = 3\)

**Lognormal:** \(b_1 = (e^Z - 1)^{1/2}(Z + 2)\) \hspace{1cm} \(b_2 = Z^4 + Z^3 + Z^2 - 3\)

where \(Z = \sigma^2\).

The \((b_1, b_2)\) pairs for the lognormal family form a curve parameterized by \(Z\), while the other families have a single \((b_1, b_2)\) pair. The procedure for choosing the distributional family is to first estimate the skewness and kurtosis by replacing \(m_2\), \(m_3\), and \(m_4\) in their definitions by \(m_{\bar{2}}\), \(m_{\bar{3}}\), and \(m_{\bar{4}}\). The \(b_{\bar{1}}\) and \(b_{\bar{2}}\) points thus obtained are then plotted and the point (or line in case of the lognormal) closes to this point is chosen as the correct family. For cases in which the \((b_{\bar{1}}, b_{\bar{2}})\) point is in between two distributional families, the choice is based on two statistical tests: Chi-square and the Kalmogorov-Smirnov test. A plot of the \((b_{\bar{1}}, b_{\bar{2}})\) pairs for several standard distributions is shown in Figure A-1.
Figure A.1 Regions of Parametric Families
APPENDIX B

THE RASCAL ALGORITHM: A USER'S ALGORITHM MANUAL

The purpose of this appendix is to highlight the important features of the RASCAL algorithm, present a flowchart of the program logic, and provide a sample output. At the end of the appendix, a standard user's manual for executing the RASCAL program is provided.

The overall logic of the RASCAL algorithm is shown in Figure B-1. The first step in the program execution is the definition of the control parameters. There are nine control parameters which are defined in Table B.1. The next input is the DPD arrays. This is done in subroutine DEFINE. Figure B.2 provides a flowchart of this subroutine. The control variable for how the input is to be made is IG0. If IG0 is greater than zero, equal probability intervals for the distribution types as defined in Table B.2 are used. The parameters for the distributions are given by P1, P2, and P3. If IG0 is equal to zero, the discrete value and its associated probability are both user defined. If IG0 is less than zero then the probability of each discrete interval is input by the user, and the program calculates the interval breakpoints. The jth discrete value for variable Iy is stored in A(Iy, 1, j) while its associated probability is stored in A(Iy, 2, j).

The final step in the program set-up is to define the number of calculations to be performed, NEND. If RASCAL is used then NEND = NEND*NBIN unless this is greater than NBIN raised to the NV power in which case NEND is set equal to this value. If the LHS method is being used then NEND is set equal to NBIN.

The logic for obtaining the responses is shown in Figure B.1.
There are three functions shown which perform the following calculations.

**Rand (I,J).** Generates a random integer in the range I to J, inclusive.

**SAMPLE (IGO).** Samples from a distribution defined by IGO.
The distribution type is given in Table B.2 while the parameters are defined by input.

**COMBIN.** This is a user supplied function routine which defines the deterministic function being studied. The output of COMBIN is stored in the HOLD array as shown in the figure.

---

**TABLE B.1. RASCAL CONTROL PARAMETER DEFINITIONS**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>NV</td>
<td>Number of random variables</td>
</tr>
<tr>
<td>NBIN</td>
<td>Number of discrete intervals</td>
</tr>
<tr>
<td>NEND</td>
<td>Number of values to be calculated for each interval</td>
</tr>
</tbody>
</table>
| LHS      | 0 - Use RASCAL  
           | 1 - Use LHS |
| ICMB     | 1: Test case for RASCAL  
           | 2: Fatigue crack growth  
           | 3: Small fault tree  
           | 4: Large fault tree |
| NDEL     | Time increment |
| NF       | Final time |
| NPRT     | Printout every NPRT cycles |
| FAIL     | Upper bound for function COMBIN |
Once all of the NEND values of the response have been obtained, subroutine BINDET is called to define the discrete intervals of the response array (currently stored in HOLD(IV, 1, J) for variable IV) if RASCAL is being used. The logic for defining the bins is to order HOLD from smallest to largest using a bubble sorting algorithm. The NBINth value in HOLD is then used as the endpoint of the interval. This strategy insures that the same number of data points is used to characterize each interval and thus the statistical accuracy, whether it is good or poor, is consistent throughout the calculation. The probability is renormalized, and the expected value of the response is calculated in subroutine AVG. The response array is then condensed as
RASCAL MAIN PROGRAM FLOWCHART

1. Read Initial Parameters
2. Call DEFINE
3. If LHS < 0 Then Call CONDMN
   - NEND = NBIN * NEND
   - If LHS > 0 Then NEND = NBIN
   - If NBIN < NEND Then NEND = NBIN
4. NT = 0
5. A

Figure B.1 RASCAL Main Program Flowchart
Figure B.1 RASCAL Main Program Flowchart (continued)
Figure B.1 RASCAL Main Program Flowchart (continued)
described in Chapter 2 with a call to subroutine CONDNS. If output is requested at this time then subroutine RITE is called. At this time, NT is incremented and if it is less than NF the process is repeated. Otherwise, RITE is called and execution is terminated.

Table B.3 contains a sample run of the RASCAL algorithm.
<table>
<thead>
<tr>
<th>VARIABLE NUMBER</th>
<th>AT CYCLE 2000.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,4849E+03</td>
<td>6,2631E+00</td>
</tr>
<tr>
<td>6,4836E+00</td>
<td>6,4836E+00</td>
</tr>
<tr>
<td>7,1501E+01</td>
<td>7,1501E+03</td>
</tr>
<tr>
<td>1,4000E+02</td>
<td>1,4000E+02</td>
</tr>
<tr>
<td>8,9409E+03</td>
<td>8,9409E+03</td>
</tr>
<tr>
<td>6,8286E+04</td>
<td>6,8286E+04</td>
</tr>
<tr>
<td>7,1501E+01</td>
<td>7,1501E+03</td>
</tr>
<tr>
<td>1,4000E+02</td>
<td>1,4000E+02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AT CYCLE 2000.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
</tr>
<tr>
<td>STANDARD DEVIATION</td>
</tr>
</tbody>
</table>
LIST OF REFERENCES


[38] Chang, C., et al., "Monte Carlo Analysis Without Sorting", unpublished work, Korea


[40] Westinghouse, AESD-TME-3113


