PROBABILITY ELEMENTS OF COST ESTIMATING
FOR BUILDINGS

by

RODNEY JOSEPH ALBERTS
SB, Massachusetts Institute of Technology
(1970)

Submitted in partial fulfillment of
the requirements for the degree of

Master of Science

at the

Massachusetts Institute of Technology
(February, 1972)

Signature of Author

RODNEY J. ALBERTS
Department of Civil Engineering, January 21, 1972

Certified by

ROBERT J. HANSON
Thesis Supervisor

Accepted by

CHARLES C. LEE
Chairman, Departmental Committee on Graduate Students of the
Department of Civil Engineering
ABSTRACT

PROBABILISTIC ELEMENTS OF COST ESTIMATING
FOR BUILDINGS

by

RODNEY JOSEPH ALBERTS

Submitted to the Department of Civil Engineering
on January 21, 1972, in partial fulfillment of the
requirements for the degree of Master of Science.

Methods of introducing probabilities in the traditional ways
of estimating costs of buildings are investigated. The investiga-
tion covers the input required and how it can best be handled to
arrive at a total cost distribution.

It is concluded that, in most cases, a minimum amount of data
will be available to describe each component in a cost estimate—
usually one, two or three data points. The input data can best be
described by a normal or a shifted lognormal distribution which is
a compromise between flexibility of the distribution to describe
empirical data and ease of computation.

Three methods of computing the total cost distribution are
investigated: (1) Assume the total cost to be normally distributed
and then derive the mean and variance of the distribution; (2) Per-
form a Monte Carlo simulation; and (3) Numerically evaluate the
convolution integral. It is concluded that if certain general con-
ditions are met, the method of assuming a normal distribution will
provide satisfactory results. Otherwise, it will be necessary to
perform a Monte Carlo simulation or a numerical integration of the
convolution integral.

Thesis Supervisor: Robert J. Hansen
Title: Professor of Civil Engineering
ACKNOWLEDGEMENTS

This research was made possible through grants by the National Science Foundation (Grant No. GK-25510X) and the Perkins and Will Partnership.

The research was supervised by Professors Robert J. Hansen and José M. Roesset.

Special thanks is due to Mrs. Jessica Malinofsky for her help in preparing and typing the final draft.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Page</td>
<td>1</td>
</tr>
<tr>
<td>Abstract</td>
<td>2</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>3</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>4</td>
</tr>
<tr>
<td>List of Figures</td>
<td>6</td>
</tr>
<tr>
<td>List of Tables</td>
<td>6</td>
</tr>
<tr>
<td><strong>Chapter 1</strong></td>
<td></td>
</tr>
<tr>
<td>Introduction</td>
<td>7</td>
</tr>
<tr>
<td>1.1 Nature of Cost Estimating</td>
<td>7</td>
</tr>
<tr>
<td>1.2 Decision Analysis</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Outline of Thesis</td>
<td>11</td>
</tr>
<tr>
<td><strong>Chapter 2</strong></td>
<td></td>
</tr>
<tr>
<td>State of the Art in Cost Estimating</td>
<td>12</td>
</tr>
<tr>
<td>2.1 Simulation Techniques</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Numerical Integration Techniques</td>
<td>14</td>
</tr>
<tr>
<td><strong>Chapter 3</strong></td>
<td></td>
</tr>
<tr>
<td>Nature of Probability in Cost Estimating</td>
<td>17</td>
</tr>
<tr>
<td>3.1 Sources of Uncertainty in Cost Estimating</td>
<td>17</td>
</tr>
<tr>
<td>3.1.1 Price and Wage Uncertainty</td>
<td>18</td>
</tr>
<tr>
<td>3.1.2 Uncertainties in Amount of Material and Labor</td>
<td>19</td>
</tr>
<tr>
<td>3.1.3 Uncertainties in Total Quantity of an Item</td>
<td>19</td>
</tr>
<tr>
<td>3.2 Correlations</td>
<td>20</td>
</tr>
<tr>
<td><strong>Chapter 4</strong></td>
<td></td>
</tr>
<tr>
<td>Possible Model Forms</td>
<td>23</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>23</td>
</tr>
<tr>
<td>4.2 Form of the Probability Distribution</td>
<td>24</td>
</tr>
<tr>
<td>4.2.1 Normal Distribution</td>
<td>25</td>
</tr>
<tr>
<td>4.2.2 Lognormal Distribution</td>
<td>28</td>
</tr>
<tr>
<td>4.2.3 Beta Distribution</td>
<td>34</td>
</tr>
<tr>
<td>4.2.4 Conclusions</td>
<td>36</td>
</tr>
</tbody>
</table>
Table of Contents Continued

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3 Form of the Input</td>
<td></td>
</tr>
<tr>
<td>4.3.1 One Input Value</td>
<td>37</td>
</tr>
<tr>
<td>4.3.2 Two Input Values</td>
<td>38</td>
</tr>
<tr>
<td>4.3.3 Three Input Values</td>
<td>39</td>
</tr>
<tr>
<td>4.3.4 Four or more Input Values</td>
<td>40</td>
</tr>
<tr>
<td>4.3.5 The Choice of the Central and Extreme Values</td>
<td>40</td>
</tr>
<tr>
<td>4.4 Form of the Algorithm</td>
<td></td>
</tr>
<tr>
<td>4.4.1 Method of Assuming a Normal Distribution</td>
<td>45</td>
</tr>
<tr>
<td>4.4.2 Monte Carlo Simulation</td>
<td>51</td>
</tr>
<tr>
<td>4.5 Use of Cost Models</td>
<td></td>
</tr>
<tr>
<td>4.5.1 Limitations</td>
<td>58</td>
</tr>
<tr>
<td>4.5.2 Data Necessary</td>
<td>60</td>
</tr>
<tr>
<td>Chapter 5 Summary and Conclusions</td>
<td>65</td>
</tr>
<tr>
<td>5.1 Summary and Conclusions</td>
<td>65</td>
</tr>
<tr>
<td>5.2 Recommendations for Further Study</td>
<td>67</td>
</tr>
<tr>
<td>List of References</td>
<td>76</td>
</tr>
<tr>
<td>Appendix 1 Examples of Possible Model Forms</td>
<td>78</td>
</tr>
</tbody>
</table>
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 1.1</td>
<td>PDF's of Two Alternative Projects</td>
<td>69</td>
</tr>
<tr>
<td>Fig. 2.1</td>
<td>CDF of Input for Simulation Model</td>
<td>69</td>
</tr>
<tr>
<td>Fig. 2.2</td>
<td>PDF of Input for Simulation Model</td>
<td>69</td>
</tr>
<tr>
<td>Fig. 3.1</td>
<td>PDF's and CDF's Showing Possible Effects of Correlation</td>
<td>70</td>
</tr>
<tr>
<td>Fig. 4.1</td>
<td>Normal Distribution</td>
<td>71</td>
</tr>
<tr>
<td>Fig. 4.2</td>
<td>Two-Parameter Lognormal Distribution</td>
<td>71</td>
</tr>
<tr>
<td>Fig. 4.3</td>
<td>Three-Parameter Lognormal Distribution</td>
<td>71</td>
</tr>
<tr>
<td>Fig. 4.4</td>
<td>Lognormal Distribution Skewed to the Right</td>
<td>72</td>
</tr>
<tr>
<td>Fig. 4.5</td>
<td>Graph of $f(s^2)$ for the Central Value Equal to the Mode</td>
<td>72</td>
</tr>
<tr>
<td>Fig. 4.6</td>
<td>Graph of $f(s^2)$ for the Central Value Equal to the Mean</td>
<td>73</td>
</tr>
<tr>
<td>Fig. 4.7</td>
<td>Beta Distribution</td>
<td>73</td>
</tr>
<tr>
<td>Fig. 4.8</td>
<td>Examples of Fitting a Normal Distribution through Three Points</td>
<td>74</td>
</tr>
<tr>
<td>Fig. 4.9</td>
<td>Graph of $D$ versus $V$ for Choice of Central Value Corresponding to the Mode or the Median</td>
<td>74</td>
</tr>
<tr>
<td>Fig. 4.10</td>
<td>Graph of $q_\alpha$ versus $\alpha$</td>
<td>74</td>
</tr>
<tr>
<td>Fig. 4.11</td>
<td>Graph of $D$ versus $V$ for the Normal Distribution</td>
<td>75</td>
</tr>
<tr>
<td>Fig. 4.12</td>
<td>Graph of $D$ versus $V$ for the Lognormal Distribution</td>
<td>75</td>
</tr>
<tr>
<td>Fig. 4.13</td>
<td>Graph of $D$ versus $V$ Showing Effect of Skewness</td>
<td>75</td>
</tr>
<tr>
<td>Fig. A1.1</td>
<td>CDF of Output of Example 1</td>
<td>81</td>
</tr>
<tr>
<td>Fig. A1.2</td>
<td>CDF of Output of Example 2</td>
<td>82</td>
</tr>
<tr>
<td>Fig. A1.3</td>
<td>CDF of Output of Example 3</td>
<td>83</td>
</tr>
</tbody>
</table>

### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table A1.1</td>
<td>Input Data for Example 1</td>
<td>84</td>
</tr>
<tr>
<td>Table A2.2</td>
<td>Input Data for Example 2</td>
<td>88</td>
</tr>
<tr>
<td>Table A2.3</td>
<td>Input Data for Example 3</td>
<td>89</td>
</tr>
</tbody>
</table>
Chapter 1
Introduction

1.1 Nature of Cost Estimating

From beginning to end, the building process is a very complex and involved procedure. Many people and disciplines are included—financiers, developers, architects, engineers, planners, contractors, etc. There is, however, an item that transcends the whole process and that is money. From just about any point of view, costs become one of the major items in the whole design and construction process. The owner is working within a budget, alternative designs have cost implications, and construction must be financed. Along with the actual finances involved, it is important to be able to estimate the financial situation at the various stages.

In theory, cost estimating is a simple process, which involves three steps:

1) Determine what is to be costed;
2) Estimate the unit price and the quantity;
3) Estimate the cost by multiplying the quantity by the unit price.

For the actual mechanical procedure represented by the above three steps, this is a true statement. However, the difficulty enters when trying to supply the information necessary for the mechanical process. In other words, how are the quantities and unit prices obtained?

There are a number of books and handbooks available that describe the methods and procedures for doing a quantity take-off (15, 17).
Along with this, there are probably innumerable techniques and shortcuts that have been developed by innumerable individuals to aid them in developing a cost estimate that have never been disclosed or published. In the same way, there are a number of references available for determining unit prices (20,21). Also many firms keep extensive historical data on unit costs for the purpose of using them in cost estimates.

Even with the above aids, the game of cost estimating almost by definition has many uncertainties associated with it. Prices in a capitalistic economy, such as we have in the United States, are determined in a highly complex and dynamic process. Along with this, there are uncertainties associated with the estimating of the quantities of materials involved in the estimate. The uncertainties here are probably not as great as in the prices; nevertheless, they do exist and should be recognized.

Up to the present time, the uncertainties involved with costs seem to have been only implicitly recognized in the sense that everybody concerned with cost estimating realizes that there are uncertainties involved, but very little has been done in the way of developing formal mechanisms for dealing with the uncertainties. For example, cost estimating books (see Reference 17) often give a range of costs for a particular item. The estimator must decide where in this range the true cost for a particular project lies. He then uses this single value in his estimate. Along the same lines, a cost estimate is usually given in the form of a single value which would probably be a mean or an ex-
pected cost. This single value gives no indication of the uncertainty or confidence associated with this expected value. In many circumstances it would be more meaningful to make a statement to the effect that there is an 80% chance that the cost of a building will be less than a given amount, or that the total cost has an expected value of $X$ and a standard deviation of $Y$ rather than stating that the cost will be $X$.

1.2 Decision Analysis

In many situations concerned with the building process a selection must be made from among two or more alternatives. This choice is often based, at least in part, on the relative costs of the various alternatives. If the costs associated with the alternatives are deterministic, that is, the probability of the cost actually being equal to the given cost is unity,* the decision will usually be relatively straightforward. However, this will seldom be the case in cost estimating—that is, there will usually be uncertainties involved with the alternative costs. As a result, the decision process is not as straightforward as in the deterministic case. An example will illustrate this point.

* Throughout the text the following notation will be used to describe random variables. The Probability Density Function (PDF) for a continuous random variable is given by $f_X(x)$, where $f_X(x)$ is the probability that the random variable $X$ will take on a value between $x$ and $x + dx$. The Cumulative Distribution Function (CDF),

$$F_X(x) = \int_{-\infty}^{x} f_X(x) \, dx,$$

is the probability that the random variable $X$ will take on a value less than or equal to $x$. 
Suppose that there are two alternative designs for a building and a choice must be made as to which one to use. Let Alternative 1 represent a conventional solution for which the expected cost is $1,000,000 and Alternative 2 involves an innovative process for which the expected cost is equal to $900,000. If this is the only information available, then Alternative 2 seems like the more attractive design. Suppose further, though, that there is more uncertainty associated with the total cost of Alternative 2 than with Alternative 1. In other words, let the standard deviation of Alternative 1 be $50,000, and the standard deviation of Alternative 2 is $190,000. This situation is illustrated in Fig. 1.1, which gives the PDF's for each of the alternatives. In this situation, Alternative 2 may no longer be the most attractive design. As can be seen, even though the expected cost of Alternative 2 is less than Alternative 1, there is a high probability that the cost of Alternative 2 will exceed that of Alternative 1.

The above example points out a situation that is at least implicitly recognized in many decisions. That is, an innovative scheme may be rejected because it is too risky. In other words, the decision maker assumes that the variance associated with the innovative scheme is very large. However, he may make this decision without actually knowing the true value of the variance, and so he does not really know how risky the innovative scheme is. As a result, since recognizing uncertainties may alter the decision process, it seems that there should be a formalized procedure for handling these uncertainties and the associated variances.
1.3 Outline of Thesis

The primary purpose of this thesis is to explore in the context of cost estimating for buildings the nature of the uncertainties and the associated probabilities and how these can be used to arrive at better and more meaningful cost estimates. In a thesis such as this, it is impossible to examine all the aspects of the problem and arrive at definite conclusions. Rather, what follows is what this author considers to be some of the basic questions that must be considered in developing a cost estimating model.

With this in mind, the thesis is composed of the following: Chapter 2 presents a summary of what has been done thus far in developing techniques and models to deal with the uncertainties in cost estimating. Chapter 3 considers the nature of cost estimating and how these can be included in an assessment of the uncertainty associated with cost estimating. Chapter 4 investigates possible ways of arriving at a total cost distribution, once the individual cost components are known. Chapter 5 presents conclusions and recommendations for further study and research. Finally, as an appendix, examples of various possible models are presented and compared that are based on the results of the thesis.
Chapter 2
State of the Art in Cost Estimating

The probabilistic nature of cost estimating seems to be at least implicitly recognized in most discussions of and techniques developed for cost estimating in the context of the building process. Most books on the subject talk about ranges for both costs and labor productivity for the various items in a building. Also, they usually point out that the given costs are applicable only in a given section of the country or under certain "average" conditions or that factors must be used to account for less than ideal weather—cold weather concreting, for example. Unfortunately though, very little seems to have been done in the way of developing formal mechanisms for including uncertainties in cost estimating. There have been a few developments along these lines in the last few years which include a simulation model for assessing the preliminary cost of a building (11, 16), work done in the management field for assessing the risk associated with investments (12, 13, 23), development of bidding strategies based on probabilities (6, 7), and a model developed for preparing bids on government contracts (3).

The above models are based on one of two methods for deriving the probability distribution of the total cost. These are a Monte Carlo simulation and a numerical integration of the convolution integral. In all but very simple situations, the calculus needed to derive the distribution of function of more than one random variable may be intractable. The multivariate forms of a normal distribution and the binomial distribution can be derived in an exact form (5), but for most other
distributions it is not possible to derive an exact form for the multivariate distribution. As a result it is necessary to use either simulation or numerical integration techniques to derive the multivariate distribution for a given situation.

2.1 Simulation Techniques

The techniques of Monte Carlo simulation have been developed over the past twenty years to cover two types of problems called deterministic or probabilistic depending on whether the method is used to describe a phenomenon involving uncertainty or to approximate a deterministic problem, i.e., evaluate an integral (10). In the present context we are concerned with the first type which in its simplest form is "...to observe random numbers chosen in such a way that they directly simulate the physical random processes of the original problem and to infer the desired solution from the behavior of these random numbers (10)." In other words, if we wish to determine the frequency distribution of the total cost of a building when the frequency distribution of the component costs are known, it is possible to approximate the total cost through a series of experiments. It is begun by generating a series of random numbers that correspond to the individual cost components, which can then be summed to arrive at a total cost. The process is repeated many times, the results of which can be displayed as a histogram which is an approximation of the actual PDF of the total building cost.

This technique was used in the method developed first by the Standard Oil of Indiana Company (11) and later extended by the Perkins
and Will Partnership (16). In this model, the total building cost is broken down into several components which are assumed to be described by three-parameter lognormal probability distributions. The component distributions are derived on the basis of three input values which reflect the cost spread expected by the estimator. The three input values are the most likely cost, an optimistic cost and a pessimistic cost. The most likely cost corresponds to the mode of the lognormal distribution and the optimistic and pessimistic correspond to the 10 and 90 percentiles of the CDF, i.e.,

\[ F_X(x_1) = 0.10 \]
\[ F_X(x_h) = 0.90 \]

where \( x_1 \) is the optimistic cost and \( x_h \) is the pessimistic cost. This situation is illustrated in Figs. 2.1 and 2.2. After a probability distribution has been assigned to each cost component, a Monte Carlo simulation is performed to derive the total cost distribution. This can be done directly since, at least in the original model, independence is assumed among the various cost components.

2.2 Numerical Integration Techniques

The derivation of the PDF of a sum of random variables involves the evaluation of a convolution integral of the form,

\[ F_Z(z) = \int \int f_Y(z-x) f_X(x) \, dx \, dx \]

for \( Z = X + Y \). Similarly, for the product of two random variables,
\[ Z = XY, \]

\[ F_Z(z) = \int_{\text{all } x} F_Y(z/x) f_X(x) \, dx. \]

As was pointed out earlier, except for a few cases it is impossible to obtain an exact, analytical solution to the convolution integral. However, it is possible to discretize the CDF \( F_Z(z) \) and then use a numerical integration technique to approximate the final distribution. This technique has been used in IGRAM (Interactive Graphic Risk Analysis Method), which is a model that is being developed at the Sloan School of Management at the Massachusetts Institute of Technology to assess the risk involved in investment decisions (22). In this model the user describes the probability distributions of the variables of interest by giving values that correspond to several percentiles of the CDF—the model in its present form will accept up to 200 points. After all the variables have been described, the model numerically evaluates the convolution integral following a pairwise method. Even though the IGRAM model was originally developed for investment evaluation, it would clearly be possible to use it in place of a Monte Carlo simulation for assessing the total cost distribution of a building. In the current version of IGRAM it is necessary to assume either independence or perfect correlation among the various contributing random variables. The developers of the model plan to incorporate into future versions of the model a method for introducing partial correlations.

Both of the techniques described in this chapter are based on cer-
tain assumptions and approximations. In the following chapters these assumptions and approximations are examined for the purpose of determining how the uncertainties inherent in cost estimating can best be handled.
Chapter 3

Nature of Probability in Cost Estimating

3.1 Sources of Uncertainty in Cost Estimating

The world in many cases is probably deterministic; however, due to a lack of a complete understanding of the processes involved, there are many uncertainties in describing a process. As a result of this, it may be helpful to use probability theory to describe the process. This is quite obvious when the process of interest is the occurrence of earthquakes. If the total process that results in an earthquake were completely understood it would be possible to predict precisely the occurrence of an earthquake. Since this is not the case, the occurrence of earthquakes is treated as a random process. Even though it may not be so obvious, the same is true in cost estimating. Consequently, it is worthwhile to look at the sources of uncertainty in cost estimating. Basically, uncertainties in the cost of a building fall into three broad categories: (1) Uncertainties in the market prices of material and the labor wage rates; (2) Uncertainties in the quantities of material and labor involved in a process; and (3) Uncertainties in the total quantity of an item in the building. Depending on the type and use of the cost estimate, these three sources of uncertainty may be lumped together or it may be desirable to separate them. For example, in a preliminary estimate that would include the principal components of a building, the cost of the structural frame, say concrete, might be given as one number with an associated variance that would reflect all three of the uncertainties listed above. On
the other hand, in a more detailed estimate, it might be desirable to estimate the total cost of formwork for concrete columns in a building in terms of the square feet of surface area of the columns. To do this it would be necessary to know the amount of lumber necessary for the formwork, the labor necessary to assemble and erect the forms, the cost of lumber and labor wage rates, and the total square feet of column surface in the building. In this estimate each of the items could have a variance that would reflect the uncertainty in just that particular item.

3.1.1 Price and Wage Uncertainty

Uncertainties in wages and prices for the most part are the result of two phenomena of the market system such as exists in the United States. The first is fluctuation in prices and wages due to the interaction of the supply and demand sectors of the economy. This phenomenon will affect different items in different ways—it would probably affect wages less than material prices, and it would probably affect structural steel less than it would affect cement, since the supply of cement is more elastic than that of structural steel. Of course, this may not always be true since there are many factors that will affect prices—availability, construction activity, number of bidders, and so on.

Inflation is the second phenomenon that causes uncertainty in wages and prices. Over the long run, most wages and prices will increase, but the exact amount of the increase will in general not be known. This becomes particularly important when a cost estimate is
being made for a building that will not be constructed for a year or so in the future.

3.1.2 Uncertainties in Amount of Material and Labor

For a given job, the amount of labor per unit of production will not be constant for the duration of a single job or from job to job. For example, at the start of a job the productivity of a labor crew will in general be less than towards the middle or end of the job. In the same way, since no two jobs are identical in all respects, the respective productivities will be different. This situation is brought out in most books on cost estimating. Pulver (17) gives ranges of labor required for many items. For example, he gives the approximate labor-hours required to construct and erect 100 square feet of form surfaces as ranging from 6 to 12 hours.

In much the same way, the amount of material that is necessary for a given job will vary. Taking again an example from Pulver, he estimates that from 190 to 320 feet board measure of lumber are required to form 100 square feet of column surface. The spread is due to such factors as the column shape, the type of forms and a certain amount of inherent wastage.

Even though in all cases the spread may not be as great as in the examples given above, nevertheless the spread does exist, and there should be a way of introducing this into the field of cost estimating.

3.1.3 Uncertainties in the Total Quantity of an Item

In most cases, the total quantity of an item in a building is
determined by some type of quantity take-off which, almost by definition, will have uncertainties associated with it. The magnitude of the variance associated with a quantity take-off would most likely be a function of the estimator and the type of estimate. An experienced estimator conceivably can do a more exact take-off than an inexperienced estimator. Furthermore, in a preliminary estimate, there would probably be a greater uncertainty than in a final estimate since the final sizes and forms of many of the components are not known exactly. For example, in a concrete building at the preliminary stage structural member sizes may not be known exactly or may be subject to future change. In this case, there is obviously a range associated with the quantity of concrete. On the other hand, in a final estimate performed by an experienced estimator based on a complete set of working drawings there might be a very small or even negligible variance associated with the various component quantities in a building.

3.2 Correlations

The two examples of methods developed to introduce uncertainties in cost estimating given in Chapter 2 are based on the assumption of independence among the various cost components. This is obviously not strictly true for all items in a building. For example the quantity of reinforcing steel is strongly correlated to the quantity of concrete; conversely, the number of window units in a building is very likely loosely correlated to the quantity of roofing material. It is difficult to make general statements regarding the importance of including correlations in a cost estimate; however, a good start would
be to look at how correlations affect the mean and variance of a random variable. For this, assume that the total cost of a building is a linear function of individual cost components, i.e., \( C = \sum_{i=1}^{n} X_i \) where \( C \) is the total cost and \( X_i, \ldots, X_n \) represent the cost of various components in the building. The mean and variance of \( C \) is given by (2),

\[
E(C) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} m_{X_i}
\]

Correlation has no effect on the total mean. It does influence the variance, however,

\[
\text{Var}(C) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{Cov}(X_i, X_j),
\]

where \( \text{Cov}(X_i, X_j) = \rho_{X_i X_j} s_{X_i} s_{X_j} \)

and \( m_{X_i} \) and \( s_{X_i} \) are the mean and standard deviation of \( X_i \) and \( \rho_{X_i X_j} \) is the coefficient of correlation between \( X_i \) and \( X_j \).

In general since coefficients of correlation can vary between -1 and +1, correlation among the various components can either increase or decrease the variance of the total cost \( C \). However, in the context of cost estimating it seems reasonable to expect that the various cost components will be positively correlated. It follows that the effect of assuming independence or neglecting correlations is to predict a variance that is less than the actual variance. This situation is illustrated in Fig. 3.1, which shows possible PDF's and CDF's of the case of assuming independence and of taking correlations into account. As can be seen, an estimator may be incorrect in making a statement to
the effect that there is an 80% probability that cost will be less than \( y \) dollars, since in fact there is an 80% probability that cost will be less than \( y' \) dollars where \( y' > y \).

There seems to have been very little work done in an attempt to assess the importance of correlations in cost estimating. Some work along these lines is included in later sections of this thesis. Chapter 4 presents methods of including correlations in a Monte Carlo simulation and some possible approaches to assessing correlations. Also, the examples in Appendix 1 indicate for a limited case the effect of correlations on the variance of the total cost.
4.1 **Introduction**

In deciding on the possible form of a cost estimating model it is necessary to consider several factors:

1. From a given set of input data, what kind of probability distribution can be inferred, with what effort, and how sensitive is the final result to this decision? For example, a normal distribution offers simplicity of computation, a shifted lognormal offers more flexibility, and a beta distribution provides maximum flexibility.

2. What kind of input data can be expected from a user? Would this take the form of minimum, expected, and high values as in the case of the model used for the Standard Oil Building, a number of percentile points or possibly a mean and a standard deviation or coefficient of variation?

3. For a given set of cost components, what is the best practical way to arrive at the total cost distribution? Should this be done analytically by assuming the final distribution to be normal, or by using a Monte Carlo simulation, or by a numerical integration of the convolution integral?

These three factors are discussed separately in this chapter even though in fact they are not entirely independent. For example, the method used for arriving at the total cost distribution will have an
effect on the way the distribution of each of the cost components is assessed.

4.2 **Form of the Probability Distribution**

In order to use either a Monte Carlo method or a numerical integration method to arrive at a total cost distribution it is necessary to assign a CDF to each of the cost components. The CDF can be arrived at in a number of ways. Probably the most obvious way would be to give percentile values which could be connected by straight lines to produce a discrete form of a CDF. This could be a very satisfactory process if the estimator has enough information about a random variable to make a meaningful CDF. This type of procedure is called Subjective Probability Assessment and has been studied fairly extensively in regard to applications in risk analysis of investment policies. Since there are a number of studies along these lines (Reference 18 discusses Subjective Probability Assessment and contains a good bibliography of work on the subject) this subject will not be covered in this thesis.

From the viewpoint of cost estimating, it is probably more beneficial to look at the case where an estimator has a limited amount of data about a given cost component. For this case, in order to arrive at a meaningful CDF, it is necessary to assume that the cost variable can be described by a continuous form of a PDF and a CDF, and that the data is sufficient to derive the parameters of the distribution. In order to consider a given distribution, it should possess the following three characteristics:

1. The PDF of the proposed distribution should be unimodal. That is,
it seems reasonable to assume that for a given cost component there should be only one "most likely" cost.

2. Since, in the present context, we are dealing with costs, it seems reasonable to assume the distribution should be defined only for values greater than zero, which reflects the fact that negative costs are meaningless. However, as will be seen in later discussions, a distinction should be made between the mathematical properties of a distribution and its reasonable physical interpretation. For example, the range of the normal distribution is from \(-\infty\) to \(+\infty\). However, in real life, it is possible to set up the normal distribution in such a way that there is a negligible probability that the random variable will take on a value less than zero.

3. The distribution should be relatively easy to use. This would include the ease with which the parameters of the distribution can be derived and also the ease with which a random variate can be generated for use in a Monte Carlo simulation.

There are several distributions that satisfy these requirements, which will be discussed in the following section.

4.2.1 Normal Distribution

The form of the PDF of the general normal distribution is

\[
f_X(x) = \frac{1}{s_X \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x - m_X}{s_X} \right)^2 \right] \text{ for } -\infty \leq x \leq \infty.
\]

This PDF is shown in Fig. 4.1 (2). The PDF and the CDF of the standardized normal distribution where \(m_X = 0\) and \(s_X = 1\) are widely tabu-
lated and can be used to evaluate the general case. The PDF and CDF
of the standardized normal variate are denoted by $f_{U}(u)$ and $F_{U}(u)$,
where

$$f_{U}(u) = \frac{1}{\sqrt{2\pi}} \exp \left[ - \frac{1}{2} u^2 \right], \quad -\infty < u < \infty$$

and

$$F_{U}(u) = \int_{-\infty}^{u} f_{U}(u) \, du$$

Using these, the PDF and CDF of the general normal variate become,

$$f_{X}(x) = \frac{1}{s_{X}} f_{U} \left( \frac{x - m_{X}}{s_{X}} \right), \quad -\infty < x < \infty,$$

and

$$F_{X}(x) = F_{U} \left( \frac{x - m_{X}}{s_{X}} \right), \quad -\infty < x < \infty. \quad (4.1)$$

As can be seen, in the general case, the normal is a two-parameter
distribution. As a result, two data points are necessary to estimate
the parameters $m_{X}$ and $s_{X}$. This can be accomplished by making use of
Equation 4.1 in the following ways, depending on the form of the input
data.

Mean and One Extreme Value Known

In this case we have the mean, $m_{X}$, and the value of the variable,
$x$, that corresponds to the $\alpha(100)$ percentile of the cumulative distri-
bution, and we want to determine the standard deviation, $s_{X}$. Let $q_{\alpha}$
be the value of the standardized normal variate corresponding to the
$\alpha(100)$ percentile. Then,

$$F_{U} \left( \frac{x - m_{X}}{s_{X}} \right) = \alpha,$$

and

$$\frac{x - m_{X}}{s_{X}} = q_{\alpha},$$
which gives
\[ s_X = \frac{x - m_X}{q_\alpha} \]

**Two Extreme Values Known**

For this case we have the values of the variate, \( x_1 \) and \( x_2 \), corresponding to two percentiles of the cumulative distribution and we want to determine the mean, \( m_X \), and the standard deviation, \( s_X \). If \( \alpha(100) \) and \( \beta(100) \) are the percentiles for which we know the values, \( x_1 \) and \( x_2 \), and \( q_\alpha \) and \( q_\beta \) are the corresponding values of a standardized normal variate, then

\[ F_U \left( \frac{x_1 - m_X}{s_X} \right) = \alpha \]

and

\[ F_U \left( \frac{x_2 - m_X}{s_X} \right) = \beta \]

which gives
\[ \frac{x_1 - m_X}{s_X} = q_\alpha \]

and
\[ \frac{x_2 - m_X}{s_X} = q_\beta \]

Solving these two equations for \( m_X \) and \( s_X \) gives,

\[ s_X = \frac{x_1 - x_2}{q_\alpha - q_\beta} \]

and
\[ m_X = x_2 + q_\beta s_X. \]

For simulation purposes a standardized normal variate can be generated from a uniform variate,\(^*\) as a direct consequence of the Central Limit Theorem.

\(^*\) Most computer installations have programs for generating a uniform random variate in the interval 0 to 1.
Central Limit Theorem (2), which states that under rather general conditions the distribution of a sum of random variables approaches the normal distribution. For the uniform distribution the convergence is rapid. For example, the IBM supplied subroutine (called GAUSS) for generating a standardized normal variate uses 12 uniform variates.

The derived standardized normal variate can be transformed into a general normal variate with mean, $m_X$, and standard deviation, $s_X$, by making use of the following relationships,

$$ u = \frac{x - m_X}{s_X} $$

which gives

$$ x = u \cdot s_X + m_X $$

where $u$ is a standardized normal variate and $x$ is the desired normal variate with mean, $m_X$, and standard deviation, $s_X$.

4.2.2 Lognormal Distribution

The lognormal distribution is related to the normal distribution in that the logarithm of the random variable is normally distributed, i.e., given the equation $X = \ln Y$, if $X$ is normally distributed, then $Y$ is lognormally distributed. The PDF of the lognormal distribution has the following form,

$$ f_Y(y) = \frac{1}{y \sigma_X \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{\ln y - m_X}{\sigma_X} \right)^2 \right], \quad y \geq 0. $$

The CDF and PDF of the lognormal distribution are most easily evaluated in terms of the standardized normal variate $u$. This becomes,
\[ f_Y(y) = \frac{1}{y \sigma_X} f_U \left( \frac{\ln y - \mu_X}{\sigma_X} \right) \]

and

\[ F_Y(y) = F_U \left( \frac{\ln y - \mu_X}{\sigma_X} \right) \]

The general shape of the lognormal distribution is shown in Fig. 4.2. It should be noted that the lognormal parameters \( \mu_X \) and \( \sigma_X \) in the above equations are scale and shape parameters. The location of the median, mean and mode and the magnitude of the standard deviation can be derived from the lognormal parameters as follows,

\[
\bar{m}_Y = \exp (\mu_X)
\]

\[
m_Y = m_Y \exp \left( \frac{1}{2} \sigma_X^2 \right),
\]

\[
\bar{m}_Y = \exp (\mu_X - \sigma_X^2),
\]

and

\[
s_Y^2 = m_Y^2 \left[ \exp (\sigma_X^2) - 1 \right],
\]

where \( \bar{m}_Y \) = median, \( m_Y \) = mean, \( \bar{m}_Y \) = mode and \( s_Y^2 \) = variance.

In the usual form as given above, the lognormal distribution is a two parameter distribution. However, it is possible to add a third parameter, \( t \), to the distribution which is simply a shift of the origin. This situation is shown in Fig. 4.3. The PDF and CDF of the shifted lognormal distribution are given by

\[ f_Y(y) = \frac{1}{(y-t)\sigma_X} f_U \left( \frac{\ln(y-t) - \mu_X}{\sigma_X} \right), \quad t \leq y \leq \infty, \]
and \[ F_Y(y) = F_U\left(\frac{\ln(y-t) - \mu_X}{\sigma_X}\right), \quad t \leq y \leq \infty. \]

The location parameters become
\[ \bar{m}_Y = t + \exp(\mu_X), \]
\[ m_Y = t + \exp\left(\frac{1}{2} \sigma_X^2\right), \]
\[ \hat{m}_Y = t + \exp(\mu_X - \sigma_X^2), \]

and the variance and standard deviation are unchanged by the transformation. In general, as is indicated in Fig. 4.3, the lognormal distribution is skewed to the left; however, it is possible to derive a lognormal distribution that is skewed to the right by a substitution of the form \( y' = t - y \), and the PDF becomes
\[ f_Y(y') = \frac{1}{(t-y)\sigma_X} f_U\left(\frac{\ln(t-y) - \mu_X}{\sigma_X}\right), \quad -\infty \leq y' \leq t. \]

This results in the distribution shown in Fig. 4.4. It is obvious from Fig. 4.4 that the shape and the scale parameters are unchanged and the location parameters become,
\[ \bar{m}_Y = t - \exp(\mu_X) \]
\[ m_Y = t - \exp\left(\frac{1}{2} \sigma_X^2\right), \]
and
\[ \hat{m}_Y = t - \exp(\mu_X - \sigma_X^2). \]

For the shifted lognormal distribution it is necessary to give three data points in order to determine the three parameters, \( \mu_X, \sigma_X \).
and t. In the context of cost estimating, the three data points will probably take the form of specifying one of the location parameters* and two percentile values of the CDF. For these cases, the Method of Percentiles (1) can be used to estimate $\mu_X$, $\sigma_X$ and t. Also in the following discussion, it will be assumed that the input values will correspond to symmetrical percentiles, that is, the percentiles will be $\alpha(100)$ and $(1-\alpha)(100)$. This should not be a serious limitation from the point of view of the estimator; however, it does make the following computation much more simple.

**Two Extreme Values and the Median Known**

Let $y_\alpha$ and $y_{1-\alpha}$ be the values of the random variable that correspond to the $\alpha(100)$ and $(1-\alpha)(100)$ percentiles, then we can write the following equations:

$$y = t \pm \exp \left( \mu_X \pm q \sigma_X \right),$$

$$y_{1-\alpha} = t \pm \exp \left( \mu_X \pm q \sigma_X \right),$$

and

$$m_Y = t \pm \exp \left( \mu_X \right)$$

where q is the value of the standardized normal variate that corresponds to the $\alpha(100)$ percentile of the CDF, i.e., if $F_Y(u) = \alpha$, then $u = q$; and the sign ($\pm$) is dependent on the skewness of the distribution. The solution of these equations results in

* Section 4.3.5 will discuss the implications of choosing the different location parameters.
\[
\sigma_X = \frac{1}{q} \ln \left[ \frac{y_{1-\alpha} - m_Y}{m_Y - y_{\alpha}} \right],
\]
\[
\mu_X = \ln \left[ \frac{m_Y - y_{\alpha}}{1 - \exp(-q \sigma_X)} \right],
\]
and
\[
t = m_Y - \exp(\mu_X),
\]
for the distribution skewed to the left. For a distribution skewed to the right, the solution results in,
\[
\sigma_X = \frac{1}{q} \ln \left[ \frac{m_Y - y_{\alpha}}{y_{1-\alpha} - m_Y} \right],
\]
\[
\mu_X = \ln \left[ \frac{m_Y - y_{\alpha}}{\exp(q \sigma_X) - 1} \right],
\]
and
\[
t = m_Y + \exp(\mu_X).
\]

**Two Extreme Values and the Mode Known**

The procedure for this case begins as in the previous case, and the following three equations result,
\[
y_{\alpha} = t \pm \exp(\mu_X \pm q \sigma_X),
\]
\[
y_{1-\alpha} = t \pm \exp(\mu_X \pm q \sigma_X),
\]
and
\[
m_Y = t \pm \exp(\mu_X - \sigma_X^2).
\]
These equations cannot be solved directly for \( \mu_X, \sigma_X \) and \( t \). Rather, we can define the following function \((1)\),
\[ f(\sigma^2) = \frac{m_Y - y_\alpha}{y_{1-\alpha} - m_Y} = \frac{\exp \left( -\sigma_X^2 \right) - \exp \left( -q \sigma_X \right)}{\exp \left( q \sigma_X \right) - \exp \left( -\sigma_X^2 \right)} \]

This function is graphed in Fig. 4.5. This graph is entered with a value of \( f(\sigma^2) = \frac{m_Y - y_\alpha}{y_{1-\alpha} - m_Y} \) and the corresponding value of \( \sigma_X^2 \) is determined. We can then determine \( \mu_X \) and \( t \) by the following equations,

\[ \mu_X = \ln \left[ \frac{y_{1-\alpha} - m_Y}{\exp \left( q \sigma_X \right) - \exp \left( -\sigma_X^2 \right)} \right], \]

and

\[ t = y_{1-\alpha} - \exp \left( \mu_X + q \sigma_X \right). \]

**Two Extreme Values and the Mean Known**

This procedure results in the following three equations,

\[ y = t + \exp \left( \mu_X + q \sigma_X \right), \]

\[ y_{1-\alpha} = t + \exp \left( \mu_X + q \sigma_X \right), \]

and

\[ m_Y = t + \exp \left( \frac{1}{2} \sigma_X^2 \right). \]

Again these three equations cannot be solved directly and so we define,

\[ f(\sigma^2) = \frac{m_Y - y_\alpha}{y_{1-\alpha} - m_Y} = \frac{\exp \left( \frac{1}{2} \sigma_X^2 \right) - \exp \left( -q \sigma_X \right)}{\exp \left( q \sigma_X \right) - \exp \left( \frac{1}{2} \sigma_X^2 \right)} \]

The graph of this function is shown in Fig. 4.6. The resulting equations for \( \mu_X \) and \( t \) are,

\[ \mu_X = \ln \left( y_{1-\alpha} - m_Y \right) - \ln \left[ \exp \left( q \sigma_X \right) - \exp \left( \frac{1}{2} \sigma_X^2 \right) \right] \]

and

\[ t = y_{1-\alpha} - \exp \left( \mu_X + q \sigma_X \right). \]
Because of the relationship between the lognormal distribution and the normal distribution, it is relatively easy to transform a standardized normal variate into a lognormal variate with parameters $\mu_X$, $\sigma_X$ and $t$. The transformation is given by $y = t + \exp(\sigma_X u + \mu_X)$ where $u$ is the standardized normal variate and $y$ is the desired lognormal variate. This is the case for a lognormal distribution skewed to the left. For a distribution skewed to the right the transformation is similar, i.e., $y = t - \exp(\sigma_X u + \mu_X)$.

4.2.3 Beta Distribution

The most general form of the beta distribution involves four parameters and as a result is a very flexible distribution for describing empirical data (2). The PDF is given by

$$f_Y(y) = \frac{1}{B(b-a)t^{r-1}} (y-a)^{r-1} (b-y)^{t-r-1}$$

where $B = \frac{(r-1)!}{(t-1)!}$ for integer values of $r$ and $t-r$ or

$$B = \frac{\Gamma(r) \Gamma(t-r)}{\Gamma(t)}$$

for $r$ and $t-r$ not being restricted to integer values.

The mean and variance of the beta distribution are given by,

$$m_X = \frac{r}{t} (b-a) + a$$

and

$$s_X^2 = (b-a)^2 \frac{r(t-r)}{t^2(t+1)}$$

Fig. 4.7 gives an indication of the great variety of shapes the distribution can be made to assume by varying the parameters $r$ and $t.$
Since four parameters are involved it is necessary to input four data points. However, it is not possible to use the Method of Percentiles to derive the distribution parameters as was the case in the normal and lognormal distributions. The great value of the beta distribution lies in its ability to approximate the form of a great number of data points, but in general it does not seem to be applicable to estimation procedures that involve a few data points.

The beta distribution has been used in PERT (Program Evaluation and Review Technique) (14) which is an extension of CPM (Critical Path Method) that assigns a probability distribution to the job durations. Three values are given by the user—the most likely duration (mode of the distribution), an absolute minimum duration, \( a \), and an absolute maximum duration, \( b \). In order to obtain the fourth piece of data necessary, the method assumes that the standard deviation is equal to \( \frac{1}{6} \) of the interval \( (b - a) \) and that the mean is given by, \( m_\chi = \frac{1}{6} (a + 4m + b) \). The distribution parameters can then be determined from these equations and the equations for the mean and standard deviation given previously.

A beta variate with integral values for \( r \) and \( t - r \) can be generated in terms of a standardized normal variate. First of all a beta variate in the interval 0 to 1 is obtained through the following transformation (9),

\[
y' = \frac{\frac{2r}{\sum_{i=1}^{2r} u_i^2} \sum_{i=1}^{2r} u_i^2}{\frac{2r}{\sum_{i=1}^{2r} u_i^2} + \frac{2t}{\sum_{i=2r-1} u_i^2}}
\]
where $y'$ is a beta variate in the interval 0 to 1 and the $u_i$ are standardized normal variates. The desired variate, $y$, in the interval $a$ to $b$ is then obtained from the transformation, $y = a + (b - a) y'$.

4.2.4 Conclusions

There are a number of other continuous distributions that satisfy the three criteria given in Section 4.2. These include the gamma distribution, the extreme value distributions, the chi-square, the t and F distributions and so on. It does not appear that these distributions would offer any further advantages over the three distributions described earlier.

The normal distribution has been widely used in both engineering and non-engineering applications. This is due to several reasons: (1) The distribution has been extensively studied and its standardized CDF is widely tabulated, (2) From the Central Limit Theorem, the sum of many effects are approximately normally distributed, and (3) It is often assumed to describe the underlying process when there is no compelling or logical reason to assume some other distribution.

The lognormal distribution has properties very similar to those of the normal distribution. In addition, the three-parameter lognormal distribution offers a higher degree of flexibility in describing empirical data or observations. The beta distribution offers an additional degree of flexibility, but the form of the PDF is more complex than the PDF's of both the normal and the lognormal distributions. The result is that there is a definite trade-off between flexibility and ease of use. With this in mind, this author has reached the conclusion that
the normal and lognormal distributions offer a good compromise between flexibility and ease of use and should prove to be satisfactory in most applications in cost estimating models.

4.3 Form of the Input

It is not possible to cover all possible situations that might arise concerning the input data that is necessary to arrive at the final cost distribution. It is visualized by this author that in most situations the estimator will have a minimum amount of data available for describing the distribution of the cost components of a building. Consequently, this chapter discusses the possible forms that limited input data might take and how this data can be handled. Also the sensitivity of the final output to the assumptions that are made in regard to the input is discussed. Again, it should be pointed out that the handling of input data is not entirely independent of the method that will be used for arriving at the final distribution.

4.3.1 One Input Value

With a single value input, which would probably be the average, the problem is essentially deterministic. As a result, if a constant is included among a sum of random variables, it would contribute to the mean and variance in the following way. Let \( Y = a + bX \), where \( a \) and \( b \) are constants and \( X \) and \( Y \) are random variables. Then,

\[
\mu_Y = a + bm_X,
\]

and

\[
s_Y^2 = b^2 s_X^2.
\]

Also, if the PDF of \( X \) is given by \( f_X(x) \), then since \( Y \) is a linear func-
tion of \( X \), the PDF of \( Y \) is \( f_Y(y) = f_X[(y-a)/b] \). The effect of this is that the basic shape of the PDF of \( X \) is retained but is shifted by an amount equal to \( a \) and is scaled by an amount equal to \( 1/b \). As a result, regardless of the method to be used in deriving the total cost distribution, constants can be easily handled if they are of the form \( Y = a + bX \) which would probably cover just about all situations that might be encountered in cost estimating.

4.3.2 Two Input Values

There are at least three forms that two input values can take:

1) The mean and standard deviation (or coefficient of variation);

2) A central value and one extreme value; and

3) Two extreme values.

For Case 1 if the method of assuming a normal distribution is used to arrive at the total cost distribution, the information can be used directly. However, if other methods are used and for Cases 2 and 3 it is necessary to assume a form for the underlying distribution. Furthermore, since only two pieces of information are given, the underlying distribution would have to be described by at most two parameters—in general, either the normal or the two-parameter lognormal distribution. The choice between the normal or the lognormal would for the most part be the result of subjective evaluation on the part of the estimator.

If only two pieces of data are given, this would probably indicate that a complete description of the variable is not available or
is unknown. For this it might be satisfactory to assume a normal distribution for the variable, particularly if the variable describes such things as material quantities or labor productivity. Therefore, for most cases where only two input values are given, a normal should be assumed; however, there remains the option on the part of the estimator to assume a lognormal distribution if he feels that a skewed distribution better describes the variable.

Once the type of distribution is assumed, the distribution parameters can be determined by the methods described in Section 4.2.1 for use in a Monte Carlo simulation or in the method of assuming a normal distribution.

4.3.3 Three Input Values

It is visualized by this author that three input values would take the form of a central value and two extreme values. For this, the most obvious situation would be to assume the three values describe a three-parameter lognormal distribution, for which the parameters can be determined by the methods outlined in Section 4.2.2. Of course it is possible that the estimator might want to use a normal distribution. In general since there are three input values, it is not possible to derive a unique normal distribution. Rather it would be necessary to define an operating rule to decide on the form of the normal distribution. An example will illustrate this point. Assume that two extreme values which correspond to symmetrical percentiles and a central value are given as follows: $x_\alpha = 10$, $x_{1-\alpha} = 25$ and $x_c = 15$, where $\alpha = 0.10$ and $x_\alpha$ and $x_{1-\alpha}$ are values that correspond to the 10 and 90 percentiles
and $x_c$ is a central value, say the mean. It is obvious that a normal distribution can not be fitted through these three points. Rather, it would be necessary to decide on a scheme for arriving at a "best fit" normal distribution which could take any number of forms—ignore the mean and use the extreme values, use the mean and one extreme value, or possibly a "least square" fit. This situation is shown in Fig. 4.8. In general, then, it seems reasonable to assume that three input values describe a three-parameter lognormal distribution unless the estimator for some reason desires to use a normal distribution and is willing to devise a rule for deciding on a unique distribution.

4.3.4 **Four or More Input Values**

In general, if an estimator is able to provide four or more input values, a discrete form of a CDF should be constructed. The reason for this is that a four-parameter distribution (the beta distribution, for example) is extremely difficult to work with and that in general it would not be possible to derive a two or three parameter distribution that would include all of the input values. Of course it would be possible as in the preceding section for the estimator to devise a method for deciding on a "best fit" two or three parameter distribution.

4.3.5 **The Choice of the Central Value**

Since the normal PDF is symmetrical and unimodal, the mean, the mode and the median are coincident; consequently, there is no controversy interpreting the meaning of a given central value or "expected cost." This is not the case for the skewed lognormal distribu-
tion, where the mode, the median and the mean have distinct values. A problem arises as to whether a central input value should be interpreted as corresponding to the mean, median or mode. The choice will depend somewhat on the interpretation of the estimator or in the way in which the problem is presented to the user. The mode is the most probable value of the distribution, the median is the value of the variable which has a 50% chance of being exceeded, and the mean is analogous to the center of gravity. As can be seen, the mode and median perhaps have a more obvious physical interpretation in the context of cost estimating than the mean. As a result the following will concentrate on the mode and the median.

It is possible to get at least a feeling for the sensitivity of a lognormal distribution to the choice of the median or the mode. This can be done by looking at the difference in the value of the variable at a given level of the CDF by assuming the central value corresponds to the median or the mode. The median is given by $\bar{m}_Y = e^{\mu}$, and the mode by $m_Y = e^{\mu} - \sigma^2$. Let $c$ correspond to the central value; then we have,

$$\mu_1 = \ln c \quad \text{and} \quad \mu_2 = \ln c + \sigma^2,$$

where $\mu_1$ and $\mu_2$ are the parameters of the lognormal distribution derived by assuming $c$ corresponds to the median and the mode respectively. Assume further that we are interested in determining the difference between $c_1'$ and $c_2'$, where $c_1'$ and $c_2'$ correspond to the values of the random variable associated with the 0.80 probability level. $c_1'$ and $c_2'$ can be derived as follows:
\[ F_U \left( \frac{\ln c_1' - \mu_1}{\sigma} \right) = 0.80 , \]

which gives
\[ c_1' = c \exp (0.842 \sigma) . \]

Also
\[ F_U \left( \frac{\ln c_2' - \mu_2}{\sigma} \right) = 0.80 , \]

which gives
\[ c_2' = c \exp (0.842\sigma + \sigma^2) . \]

The percent difference, \( D \), can be expressed as
\[
D = c_2' - c_1' \left( \frac{100}{c_1'} \right) = \frac{\exp (0.842\sigma + \sigma^2) - \exp (0.842\sigma)}{\exp (0.842\sigma)} (100)
\]
\[
= [\exp (\sigma^2 - 1)](100) = V^2(100) ,
\]

where \( V \) is the coefficient of variation. A graph of \( D \) versus \( V \) is shown in Fig. 4.9. As can be seen, for relatively small values of \( V \), \( D \) is of the order of 1 to 2%, which would indicate the results should be relatively insensitive to whether the mode or the median is chosen.

When the input data is in the form of extreme values, it is necessary to specify the percentile of the CDF to which the value corresponds. Conceivably, the value could correspond to an absolute minimum or maximum, or possibly the 5, 10 or 15 percentile of the CDF. For the normal distribution there is no absolute maximum or minimum, since \( f_X(x) \to 0 \) as \( x \to \pm \infty \). The lognormal distribution is similar in that \( f_X(x) \to 0 \) as \( x \to +\infty \). However, for most practical problems it should be possible to assign a maximum or minimum value for which the probability of the variable taking on a value less than this is negli-
gible; in other words, an absolute maximum or minimum. This might be, say, the 1 and 99 percentiles or the 0.1 and 99.9 percentiles. The case for the extreme value not being an absolute minimum or maximum would be a statement to the effect that the estimator has some knowledge of a low or a high value, but is not prepared to give an absolute maximum or minimum value. This could be interpreted as being along the lines of the odds being 1 in 10 that the cost will be less than a given value or that the estimator is 90% sure that the cost will be less than a given value.

It seems reasonable to expect that the extreme value might correspond to the 5, 10 or 15 percentiles of the CDF. As in the case for choosing the mode or median of the lognormal distribution, it is possible in this situation to assess, at least approximately, the sensitivity of the cost distribution to the choice of different percentiles of the CDF. For the normal distribution, we have

\[ F_U \left( \frac{x - \mu}{\sigma} \right) = \alpha, \]

and

\[ x = q_\alpha \sigma + \mu, \]

where \( \alpha(100) \) is the given percentile of the CDF, \( q_\alpha \) is the value of the standardized normal variate corresponding to the \( \alpha(100) \) percentile and \( x \) is the value of the variable at the specified percentile. For the lognormal distribution there is a similar situation, i.e.,

\[ F_U \left( \frac{\ln x - \mu}{\sigma} \right) = \alpha, \]

and

\[ x = \exp \left( q_\alpha \sigma + \mu \right). \]
Fig. 4.10 is a plot of \( q_\alpha \) versus \( \alpha \) which indicates the change, \( \Delta q_\alpha/\Delta \alpha \), is relatively small for values of \( \alpha \) greater than approximately 0.05. For the normal distribution, Fig. 4.11 is a plot of the percent difference of \( x \) for different values of \( q_\alpha \) as a function of the coefficient of variation, i.e.,

\[
D = \frac{x_{\alpha_1} - x_{\alpha_2}}{x_{\alpha_1}} \times 100,
\]

where \( \alpha_1 \) and \( \alpha_2 \) equal 0.5 and 0.10 or 0.10 and 0.15 respectively.

Fig. 4.12 gives a similar situation for the case of the lognormal distribution. The result is that for values of the coefficient of variation that are likely to be encountered in most applications of cost estimating (\( V < 0.5 \), say) the percent difference would be less than 10% and possibly less than 5%. This would indicate that the cost distribution would be relatively insensitive to the choice of the percentile of the CDF.

Reference 18 gives a good account of the subject of deriving subjective probability assessments which has been studied fairly extensively in the area of risk analysis for investment possibilities. The general conclusion seems to be that the area of subjective probability assessment is at best a fuzzy area and is extremely difficult to apply successfully in practice. Considering this with the rather crude sensitivity analysis given above, this author has reached the conclusion that the choice of the central value corresponding to the mode or the median, or the extreme values corresponding to the 5, 10 or 15 percentiles, is relatively unimportant in an absolute sense. Rather, it is
much more important for the estimator to clearly indicate his assumptions and then to be consistent. This is not to say that the estimator should not give the median for one component and the mode for another, but that, rather, he should be clearly aware of the distinction and proceed accordingly.

4.4 Form of the Algorithm

There are at least three methods for arriving at an estimate of the PDF of the total cost. These are:

1) Assume that the total cost distribution is normal;
2) Perform a Monte Carlo simulation; and
3) Perform a numerical integration of the convolution integral.

Each of these three methods has certain advantages and disadvantages, so that it is not possible to make a statement to the effect that one of the methods is generally superior to the other methods. On the other hand, it might be possible to decide that under a given set of conditions one method might be better than one of the other methods.

This section examines in some detail the first two methods, but due to time and facility constraints it was not possible to examine the third method in any detail.

4.4.1 Method of Assuming a Normal Distribution

Of the three methods this is the easiest to apply. The basis for this method is the Central Limit Theorem, which says that under very general conditions, as the number of variables in the sum becomes large, the distribution of the sum of random variables will approach the normal distribution. The keys to this theorem are the "very gen-
eral conditions." According to Benjamin and Cornell (Reference 2, page 251) this can be loosely interpreted to mean, "... even if the number of variables involved is only moderately large, as long as no one variable dominates, and as long as the variables are not highly dependent, the distribution of their sum will be very near normal." This could have important implications for a cost estimating model. Namely, that it would be possible to arrive at a total cost distribution without performing a Monte Carlo Simulation or resorting to numerical approximations and, consequently, the model could be used without access to a computer.

If it is assumed that the total cost distribution is normal, then it is possible to derive the distribution by determining the mean and standard deviation of the total cost. If the total cost is of the form

$$C = g(X_1, X_2, \ldots, X_n)$$

it is possible to approximate its mean, $m_C$, and its variance, $s_C^2$ by expanding the equation for $C$ in a multi-dimensional Taylor series about the respective means, retaining the linear terms and taking the expectations of $C$ (2). This method will give reasonably good approximations for $m_C$ and $s_C^2$ if the coefficients of variation are not large. This method results in the following equations for $m_C$ and $s_C^2$,

$$m_C = g(m_{X_1}, m_{X_2}, \ldots, m_{X_n}) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{\partial^2 g}{\partial X_i \partial X_j} \right) m \text{ Cov} (X_i, X_j),$$

\* For the examples in Appendix 1, it is assumed that the total cost is of the form, $C = Q_i (L_i + M_i)$, where $Q_i$, $L_i$, and $M_i$ are the quantity, unit labor cost and unit material cost, respectively, of cost component $i$. 
and 
\[ s_C^2 \approx \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{\partial g}{\partial x_i} \right)_m \left( \frac{\partial g}{\partial x_j} \right)_m \text{Cov} (X_i, X_j), \]

where \( \left( \frac{\partial g}{\partial x_i} \right)_m \) denotes the partial derivative of \( g \) evaluated at \( m_{X_i} \). If the correlations among the \( X_i \)'s are neglected, the equations can be simplified to the following,

\[ m_c = g(m_{X_1}, m_{X_2}, \ldots, m_{X_n}) + \frac{1}{2} \sum_{i=1}^{n} \left( \frac{\partial^2 g}{\partial x_i^2} \right)_m \text{Var} (X_i), \]

and

\[ s_C^2 \approx \sum_{i=1}^{n} \left( \frac{\partial g}{\partial x_i} \right)^2_m \text{Var} (X_i). \]

It should be noted that these expressions for \( m_c \) and \( s_C^2 \) may be simplified even more, depending on the exact form of \( c \). For example, if \( c \) is of the form,

\[ c = \sum_{i=1}^{n} c_i \]

where \( c_i = Q_i (L_i + M_i) \), then we have,

\[ m_c = \sum_{i=1}^{n} m_{Q_i} (m_{L_i} + m_{M_i}), \]

regardless of whether or not the variables are correlated.

Also

\[ s_{C_i}^2 = (m_{L_i} + m_{M_i})^2 s_{Q_i}^2 + m_{Q_i}^2 (s_{L_i}^2 + s_{M_i}^2) \]

if the \( Q_i, L_i \) and \( M_i \) are assumed independent of each other. Then we have for the variance of \( c \),

\[ s_C^2 = \sum_{i=1}^{n} s_{C_i}^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{Cov} (C_i, C_j). \]

Because of the relative ease with which this method could be used, it would be desirable to investigate the conditions and the lim-
its on the individual terms for which it would be possible to assume a normal distribution. There are four factors which must be considered: (1) The number of variables included in the model, (2) The variance of the variables, (3) The skewness of the individual distributions, and (4) The correlations of the variables.

In general, it is not possible to discuss each of the factors separately, since the total effect depends on the relative effect of each of the factors. Examples of this would help to illustrate this point. If all the individual distributions have coefficients of skewness very nearly equal to zero (which indicates a nearly symmetrical distribution), then the sum would be very nearly normal regardless of the variations and correlations among the individual distributions. If the variance of one variable is dominant, which would be the case if the $\text{Var} (X_j) = \sum_{i \neq j} \text{Var} (X_i)$, then the final distribution would be very close to that of the dominant variable, and as a result the accuracy in assuming a normal distribution would depend on the distribution of the dominant variable. Another extreme case would occur if all the variables are perfectly correlated, for which the final distribution would be that of the correlated variables.

As can be seen, all the factors must be considered. It is difficult to decide on general criteria, since as in most engineering approximations, the validity or accuracy of the approximations very often depends on subjective evaluation. However, it is possible to develop some guidelines which should prove useful in assessing the accuracy of this method.

We can probably disregard the first factor, since 10 variables
can be considered a moderately large number and 25 is essentially infinite (10). It seems likely that for purposes of cost estimating for buildings, the number of variables will be more than 10. It also seems likely that in general for cost estimating for buildings no one variable would completely dominate, particularly if the number of variables becomes moderately large. With this in mind, it is safe to say that as long as the variance of any one variable is less than 25% of the total variance, we can expect reasonably good results.

It is possible to get a feel for the relative importance of the skewness of a distribution in using this method. This can be done by looking at a case where the components are perfectly correlated and have identical lognormal distributions, which would result in a total cost distribution that is also lognormally distributed. We want to look at the "error" associated with assuming that this distribution is normal. The error, D, can be computed in the following way. Assume that we want to know the value of C associated with the 80 percentile of the CDF, i.e., \( F_C(c) = 0.80 \). If \( m_C \) and \( s_C \) are the mean and standard deviation of C, it is possible to derive the value of C for the case that \( m_C \) and \( s_C \) are the parameters of a normal distribution and for the case where \( m_C \) and \( s_C \) are the location parameters of a lognormal distribution. Let \( c_L \) and \( c_N \) denote the 80 percentile values of C for the lognormal and normal cases respectively. The error term, D, between assuming a normal distribution when in fact the distribution is lognormal is given by

\[
D = \frac{c_L - c_N}{c_L} \quad (100).
\]
For the normal case we have,

\[ F_U \left( \frac{c_N - m_C}{s_C} \right) = 0.80, \]

which gives

\[ c_N = 0.842 \ s_C + m_C. \]

Substituting the coefficient of variation, \( V = \frac{s_C}{m_C} \), in the expression for \( c_N \), we obtain,

\[ c_N = m_C \ (1 + 0.842 \ V). \]

For the lognormal case it is necessary first to derive the parameters \( \mu \) and \( \sigma \) in terms of \( m_C \) and \( s_C \). These become,

\[ \sigma^2 = \ln \left( \frac{s_C^2}{m_C^2} + 1 \right) = \ln \left( V^2 + 1 \right), \]

and

\[ \mu = \ln m_C - \frac{1}{2} \ \sigma^2. \]

With these, \( c_L \) can be determined.

\[ F_U \left( \frac{\ln c_L - \mu}{\sigma} \right) = 0.80, \]

which gives

\[ c_L = \exp (0.842 \sigma + \mu). \]

Substitution of the expressions for \( \mu \) and \( \sigma \) gives,

\[ c_L = m_C \ \exp \left\{ 0.842 \left[ \ln (V^2 + 1) \right]^{1/2} - \frac{1}{2} \ln (V^2 + 1) \right\}. \]

D, the error term then becomes,

\[ D = \frac{\exp \left\{ 0.842 \left[ \ln (V^2 + 1) \right]^{1/2} - \frac{1}{2} \ln (V^2 + 1) \right\} - (1 + 0.842 \ V)}{\exp \left\{ 0.842 \left[ \ln (V^2 + 1) \right]^{1/2} - \frac{1}{2} \ln (V^2 + 1) \right\}}. \]

D as a function of \( V \) is shown in Fig. 4.13. As can be seen
when \( V \) is small, \( D \) is small, but that \( D \) increases rapidly with increasing \( V \).

The preceding analysis indicates an extreme case—that of assuming a symmetrical PDF (the normal) when in fact the distribution is skewed (the lognormal). In reality the true PDF will lie somewhere between these two extremes, and as a result, the true error should be less than \( D \). Therefore if the estimator knows the coefficient of variation of the total cost, then he can use the chart in Fig. 4.13 to determine \( D \). With this and his knowledge of the other three factors given at the beginning of this section, he can determine if he is willing to accept the assumption that the total cost is normally distributed.

4.4.2 Monte Carlo Simulation

A description of the Monte Carlo simulation technique was given in Section 2.1. As was pointed out, a Monte Carlo simulation results in a histogram of the total building cost. Since the Monte Carlo method is a series of experiments, we can say in general that the greater the number of experiments (or iterations in the Monte Carlo simulation) the better the approximation. The method provides an estimate, \( \bar{x} \) and \( s^2 \), of the mean, \( m_X \), and the variance, \( s^2_X \), respectively, of a random variable \( X \) by randomly sampling a number of variables, \( x_i \). We have,

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i ,
\]

and

\[
s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

The error associated with the Monte Carlo method can then be expressed
in terms of the differences between $\bar{x}$ and $m_X$ and $s^2$ and $s^2_X$, i.e.,

$$E_1 = |\bar{x} - m_X|$$

and 

$$E_2 = |s^2 - s^2_X|.$$

We will look first at the case of $E_1$. By the Central Limit Theorem, we would expect the distribution of $\bar{x}$ to be approximately normal. As a result it is possible to determine confidence levels* on the approximations of $\bar{x}$ as follows.

Let $(1 - \alpha)$ be the desired confidence level and $z$ be the value of the $(1 - \alpha/2)100$ percentile of a standardized normal distribution. Then we have,

$$E_1 = |\bar{x} - m_X| \leq \frac{z s_X}{\sqrt{n}},$$

where $s_X$ is the standard deviation of $X$ and $n$ is the number of iterations in the simulation. This can also be put in the form of,

$$n \geq \left( \frac{z s_X}{E_1} \right)^2,$$

(4.2)

which gives the number of Monte Carlo experiments necessary for a specified confidence level and a maximum acceptable error in the estimation of $m_X$.

An important aspect of the above result is that the error, $E_1$, is a function of $\frac{1}{\sqrt{n}}$, which indicates that the Monte Carlo method involves a slow rate of convergence to the exact solution. However, in the context of cost estimating, the absolute rate of convergence is probably unimportant. Rather, it is of more interest to determine if it is possible in using a Monte Carlo method to expect relatively accur-

* See Reference 2, Chapter 4 for a discussion of confidence levels.
ate results for a reasonable number of experiments. The accuracy can be assessed by assuming values of \(s_X\), \(z\) and \(E_1\) and then by using Equation (4.2) to obtain the value of \(n\). For example, assume that \(s_X = 0.10\ m_X\), that the error should be no greater than the order of 0.01 \(m_X\), and we want to be 95% confident of the results. These values give \(n = 384\). However, the estimate of \(s_X\) might be on the lower side. If we assume that \(s_X = 0.20\ m_X\), then \(n = 1536\).

In general, it is not possible to do the same type of analysis to assess the accuracy of using \(\overline{s}^2\) as an estimate of \(s_X^2\). However, it is possible to show (see for example Reference 2, page 380) that if the underlying variable, \(X\), is normally distributed, then

\[
\text{Var} (\overline{s}^2) = \frac{2(n - 1)}{n^2} s_X^4.
\]

For reasonable large values of \(n\), this shows that the variance of \(\overline{s}^2\) is approximately proportional to \(1/n\). Since the variance of \(\overline{X}\) is also proportional to \(1/n\), we might expect the accuracy of a Monte Carlo simulation in assessing \(s_X^2\) would be of the same order of magnitude as the error in assessing \(m_X\).

The general conclusions of this section are that even though a Monte Carlo method has a slow rate of convergence to the exact solution (i.e., is proportional to \(\sqrt{n}\)), it is possible to expect good results with a reasonable number of iterations, say 1000 to 2000. Furthermore, this range of iterations can be easily handled on a computer.

**Monte Carlo Simulations with Correlated Variables**

In its usual form a Monte Carlo method is used to simulate a
process in which the variables are assumed to be independent. This assumption of independence makes it possible to use a random number generator which by the nature of the process generates independent random variables, which through a suitable transformation are related to the variable of interest.

It is possible to use a Monte Carlo technique to simulate correlated random variables if the correlated random variables are first transformed into uncorrelated random variables through a suitable linear transformation. This process is called orthogonalization (4). For example, consider two random variables, $X$ and $Y$, which are not necessarily independent, and are subjected to the following linear transformation,

$$Z = aX + bY$$

and

$$W = cX + dY.$$ 

The problem then becomes one of determining the values of the coefficients $a$, $b$, $c$, and $d$ such that $Z$ and $W$ will be uncorrelated. Once this is done the equations can then be solved for $X$ and $Y$ in terms of $Z$ and $W$, i.e.,

$$X = a'Z + b'W,$$

and

$$Y = c'Z + d'W.$$ 

When this is done, we can proceed with the Monte Carlo simulation by generating independent, random values for $Z$ and $W$, which by definition are uncorrelated, and with the above transformation determine $X$ and $Y$, which will be correlated.

The method of determining the values of the coefficients in the original equation is essentially an eigenvalue problem. The method of
solution is shown below.

Starting with a system of equations of the form,

\[ Y_i = \sum_{j=1}^{n} a_{ij} X_j, \quad i = 1, 2, \ldots, n \]

in which the \( X_j \)'s are the not necessarily independent variables of interest, the \( Y_i \)'s are the uncorrelated variables and the \( a_{ij} \)'s are the coefficients that must be determined. The \( a_{ij} \)'s are eigenvectors that are determined as follows. Starting with a matrix of the coefficients of correlation of the \( X_j \)'s, \( R \), it is necessary to solve the following determinant for the eigenvalues, \( \lambda \),

\[ |R - \lambda I| = 0, \]

where \( I \) is the identity matrix. The values of \( \lambda \) are then used to determine the eigenvectors \( \overline{a}_{ij} \) in the following equation,

\[ [R - \lambda_i I] [\overline{a}_{ij}] = 0. \]

The eigenvectors must then be normalized to obtain the values of the \( a_{ij} \)'s, i.e.,

\[ \sum_{j=1}^{n} (\overline{a}_{ij})^2 = 1, \quad i = 1, 2, \ldots, n. \]

The significance of the \( \lambda_i \)'s is that they correspond to the variances of the \( Y_i \)'s, or in equation form, \( \text{Var} \ (Y_i) = \lambda_i \). Now that the \( a_{ij} \)'s are known, it is possible to solve the original equation for the \( X_j \)'s in terms of the \( Y_i \)'s. In matrix notation this becomes,

\[ \overline{X} = \overline{A}^{-1} \overline{Y} \]

where \( \overline{A} \) is the matrix containing the \( a_{ij} \)'s.
The procedure then for performing the Monte Carlo simulation is to first of all for each \( Y_i \) generate a normal random variable with a mean of 0 and a variance of \( \lambda_i \). When the \( Y_i \)'s have been determined, the \( X_j \)'s are determined with the following moments,

\[
E(X_j) = 0 \quad \text{and} \quad \text{Var} \ (X_j) = \sum_{i=1}^{n} \text{Var} \ (Y_i),
\]

since the \( Y_i \)'s are uncorrelated. It then remains to transform the \( X_j \)'s into the random variate corresponding to the distribution assumed for the cost components.

Another feature of this process that must be investigated is whether that in transforming the derived, correlated random variate, \( X \), into the random variates of interest (call them \( Z \)), the correlations remain the same. In other words, if \( Z = f(X) \), is \( \rho_{Z_i,Z_j} = \rho_{X_i,X_j} \)? In general, this is not true, but for the case where \( Z = \ln X \), the relationship is approximately true. This is shown as follows. Assuming the following relationship,

\[
Z_1 = \ln X_1, \quad (4.3a) \\
Z_2 = \ln X_2, \quad (4.3b)
\]

we want to determine if \( \rho_{Z_1Z_2} = \rho_{X_1X_2} \), where \( \rho \) is the coefficient of correlation. This can be shown to be approximately true by expanding equations (4.3a) and (4.3b) in a Taylor series about their respective means, retaining the linear terms and then taking the moments of the resulting equation. Expanding the \( \ln Y \) about \( m_Y \), we get,

\[
\ln Y = \ln m_Y + \frac{1}{m_Y} (Y - m_Y) + \ldots
\]

Retaining only the linear terms of the expansion, we get,
\[ Z = \ln Y = \ln m_Y + \frac{Y}{m_Y} - 1 = (\ln m_Y - 1) + \frac{1}{m_Y} Y. \]

This is of the form, \( Z = a + bY \). Using this relationship, we can write equations (4.3a) and (4.3b) as,

\[ Z_1 = a_1 + b_1 Y_1, \]  
\[ Z_2 = a_2 + b_2 Y_2 \]  

where \( a_i = \ln m_{Z_i} \) and \( b_i = \frac{1}{m_{Z_i}} \).

Taking the Cov \((Z_1, Z_2)\), we get,

\[ \text{Cov} (Z_1, Z_2) = E(Z_1 Z_2) - E(Z_1) E(Z_2). \]

Substituting equations (4.4a) and (4.4b) into this and expanding, we get,

\[ \text{Cov} (Z_1, Z_2) = b_1 b_2 [E(Y_1 Y_2) - E(Y_1) E(Y_2)] \]

\[ = b_1 b_2 \text{Cov} (Y_1, Y_2). \]

From the definition of covariance, this can be written as,

\[ \rho_{Z_1 Z_2} s_{Z_1} s_{Z_2} = \frac{1}{m_{Y_1} m_{Y_2}} \rho_{Y_1 Y_2} s_{Y_1} s_{Y_2}, \]  

In the same way we can find the variances of \( Z_1 \) and \( Z_2 \). These become,

\[ \text{Var} (Z_1) = b_1^2 \text{Var} (Y_1), \]

or

\[ s_{Z_1} = b_1 s_{Y_1} = \frac{s_{Y_1}}{m_{Y_1}}. \]  

Also

\[ \text{Var} (Z_1) = b_2^2 \text{Var} (Y_2), \]

or

\[ s_{Z_2} = b_2^2 s_{Y_2} = \frac{s_{Y_2}}{m_{Y_2}}. \]  


Substituting equations (4.7) and (4.8) into equation (4.5) yields the desired result,

\[ \rho_{Z_1Z_2} \frac{s_{Y_1}}{m_{Y_1}} \frac{s_{Y_2}}{m_{Y_2}} \approx \rho_{Y_1Y_2} \frac{s_{Y_1}}{m_{Y_1}} \frac{s_{Y_2}}{m_{Y_2}}, \]

or

\[ \rho_{Z_1Z_2} \approx \rho_{Y_1Y_2}. \text{Q.E.D.} \]

4.5 Use of Cost Models

4.5.1 Limitations

Under some general conditions it appears that the assumption of the total cost distribution being normal might be very satisfactory. These general conditions are:

1) A sufficient number of variables, say at least 15, are included in the model;

2) No one variable is dominant, that is, the \( \text{Var}(c_i) \leq \frac{1}{4} \text{Var}(C) \);

3) The individual component distributions are not strongly skewed;

4) The coefficient of variation of the total cost is relatively small—less than about 0.25; and

5) None of the dominant variables are highly correlated to each other.

Another argument in favor of the assumption of a normal distribution is that many distributions closely approximate the normal distribution in the area around their means. As in most situations in probability assessment, it becomes risky to attempt to predict the extreme cases. In other words, the further out in the tails of a PDF we go,
the less confident we become in the adequacy of the model. This seems to be particularly true in the area of cost estimating where, usually, we have data that covers average conditions and as a result is clustered around the mean. We probably know very little about the extreme conditions. For example, an estimator might be able to make a statement with a high degree of confidence to the effect that there is a 60% chance that the cost will be less than $X, but he might be less confident in making a statement to the effect that there is a 95% chance that the cost will be less than $Y. This argument suggests that, in general, the best that can be hoped for is that the model will adequately describe the process in the area about the mean where many distributions closely approximate the normal distribution. As a result the assumption of a normal distribution for the total cost might be valid.

It should be obvious that the interpretation of criteria and guidelines in this regard will be highly dependent on a subjective analysis by the estimator. Consequently, it becomes meaningless to provide anything but the general guidelines that have been given.

In the event that the general criteria for the assumption of normal distribution are not met, it may be desirable to go to the Monte Carlo simulation method, which is a very flexible method that can be adapted to cover many conditions. Even though the numerical integration method was not analyzed in this thesis, the author expects that closer study would show that the Monte Carlo method and the numerical integration method are somewhat interchangeable, particularly if
correlations are neglected. However, if the IGRAM system is extended so as to include partial correlations, then they may be completely interchangeable. A big factor, though, that is not known at this time is that of computational efficiency; that is, the relative costs of the two methods would have to be known in order to make a decision as to which method is the "better" method to use.

4.5.2 Data Necessary

As has been pointed out in several places in this thesis, uncertainties in cost estimating have always been at least implicitly recognized. Books on cost estimating often give ranges for prices (Reference 17, for example). Reference 7 has a listing of coefficients of variations that has been compiled to describe the productivity of various types of construction activity. Furthermore, most design offices probably have considerable quantities of historical cost data. It appears likely, then, that the data necessary for setting up a probabilistic cost estimating model is available, but it should be pointed out that for a researcher it is very hard to get. This is due in part to a natural reluctance of contractors, developers, architects and so on to release cost data.

The discussion in Section 4.4.1 indicated that correlations among the various components could influence the variance of the total cost. Even though the significance of cost component correlations has not been thoroughly explored in this thesis, it would be worthwhile to discuss methods of assessing correlations. There are at least two ways to get at component correlations. The first is a statistical analysis
of historical data, while the second, which might be considered the inverse of the first method, consists of a description of the causal relationships among the various components. The first method involves techniques of statistical regression analysis which will not be considered here, since there is a large volume of literature on the subject.

The second method of introducing correlation into a cost estimating model is a recognition that causal relationships exist among the various components. An attempt is made to identify these causal relationships which would be given in terms of what will be called correlation factors. It might be easier to first show the mathematical interpretation of this concept and then discuss the physical significance of the correlation factors. If we start with a system of equations as follows,*

\[ Y_1 = AX_1, \]

and

\[ Y_2 = AX_2 \]

where \( X_1, X_2, Y_1, Y_2 \) and \( A \) are all random variables and \( X_1 \) and \( X_2 \) are independent of each other and each is independent of \( A \), we can determine the correlation between \( Y_1 \) and \( Y_2 \) in the following way. From the definition of covariance we get,

\[ \text{Cov} (Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1) E(Y_2), \]

which upon substitution of \( Y_1 = AX_1 \) and \( Y_2 = AX_2 \) and expanding gives,

* It should be pointed out that the multiplicative form in this discussion is only one possible form. Others, such as an additive or exponential form, might be valid in certain circumstances.
\[ \text{Cov}(Y_1, Y_2) = m_{X_1} m_{X_2} s_A^2 \] 

This can also be written in terms of the coefficient of correlation,

\[ \rho = \frac{m_{X_1} m_{X_2} s_A^2}{s_{Y_1} s_{Y_2}} \]

where \( \rho \) is the coefficient of correlation between \( Y_1 \) and \( Y_2 \). Also we can derive the expressions for \( s_{Y_1} \) and \( s_{Y_2} \),

\[ s_{Y_1}^2 = s_A^2 s_{X_1}^2 + m_A^2 s_{X_1}^2 + m_{X_1}^2 s_A^2, \]

and

\[ s_{Y_2}^2 = s_A^2 s_{X_2}^2 + m_A^2 s_{X_2}^2 + m_{X_2}^2 s_A^2. \]

The result is the coefficient of correlation between \( Y_1 \) and \( Y_2 \) in terms of known or given parameters. A numerical example of this follows.

Assume that two items in a cost estimate are the excavation, with a mean of $446,600 and a standard deviation of $32,900, and the sheet piling, with a mean of $768,000 and a standard deviation of $50,200. Along with these we want to include an inflation factor on both that has a mean of 1.00 and a standard deviation of 0.11. The resulting situation is as follows:

\[ Y_1 = IX_1 \text{ and } Y_2 = IX_2 \] \quad \text{where } X_1 \text{ and } X_2 \text{ are the initial prices of excavation and shoring, } I \text{ is the Inflation Correlation Factor and } Y_1 \text{ and } Y_2 \text{ are the final prices. Using the equations developed above we get the following moments for } Y_1 \text{ and } Y_2, \]

\[ m_{Y_1} = $446,000 \quad \text{and} \quad s_{Y_1} = $55,000 \]
\[ m_{Y_2} = 768,000 \quad \text{and} \quad s_{Y_2} = 91,700 \]

and

\[ \text{Cov}[Y_1, Y_2] = 1,720,000. \]

From which we get,

\[ \rho_{Y_1 Y_2} = 0.34. \]

In order to include this technique in a cost estimating model, much more work needs to be done. Conceivably, it could be used in conjunction with regression analyses to arrive at a better understanding of component correlation. However, it is worthwhile to consider the physical significance of the correlation factors (I in the preceding example).

This author envisions that the correlation factors would be used to account for such things as inflation, wage increases, strike possibilities, material shortages, a bleak outlook for bidding and so on. To some extent all of these items could be considered random processes. For example, if an estimator is developing an estimate for a building that will be constructed sometime in the future, say two years from now, and his cost figures are in terms of present costs, he could include in his estimate a correlation factor that would account for inflation. Since inflation is not constant, it could be treated as a random variable with a given mean and standard deviation. The correlation factor, I, in the previous example might then have a mean of say, 1.05, and a standard deviation of say, 0.1. The resulting correlation factor for inflation would then multiply each cost component. Another example might be that the estimator knows that the contract for given trade will expire during the construction period and a new wage rate
will be negotiated. After looking at records of past wage settlements he might be able to determine an expected wage increase and an associated standard deviation. The resulting correlation factor would then multiply those items in the cost estimate which are affected by the given trade. Along these same lines, a family of correlation factors could be developed for a given cost estimate that would serve a dual purpose—(1) Correlations would be introduced into the model, and (2) The estimator is able to use his present knowledge of prices and conditions to project future prices and conditions.
Chapter 5

Summary and Conclusions

5.1 Summary and Conclusions

There are at least three major sources of uncertainty in cost estimating for buildings. These are:

1. Uncertainties in the market prices of materials and the labor wage rates;
2. Uncertainties in the quantities of material and labor involved in a process; and
3. Uncertainties in the total quantity of an item in a building.*

Depending on the type and use of the cost estimate, these three sources of uncertainty may be lumped together, or it may be desirable to separate them.

It is expected that in most cases a minimum amount of input data will be available to an estimator to describe the uncertainties in the cost components. The input data could take any one of several forms, but it would probably fall into one of the following categories:

1. One input value which would be an "average" cost. For this the probability is essentially deterministic.

* It is probably worth pointing out that the length of construction time can also be a major source of uncertainty in the total cost. Even though this is implied in Item 2 above, it might be worthwhile to consider the length of construction as a separate item.
2. Two input values which could be a mean and standard deviation (or coefficient of variation), two extreme (or percentile) values, or a central value (mean, median or mode) and an extreme value.

3. Three input values which would be a central value and two extreme values.

4. Four or more input values which would be a number of percentile values.

Once the input data is provided, it is necessary to decide on the form of the CDF that will be used to describe the data. It is concluded that for two input values, a normal or a two-parameter lognormal distribution should be assumed, and that for three input values a three-parameter lognormal distribution should be assumed. This conclusion is based on a compromise between flexibility of the distribution to describe empirical data and ease of computation. For four or more input values it would in general be necessary to use a discrete CDF to describe the data.

After the form of each of the cost components has been decided upon, there are at least three methods that can be used to derive the total cost distribution:

1. Assume that the total cost will be normally distributed and then derive the mean and standard deviation of the total cost.

2. Perform a Monte Carlo simulation.

3. Numerically evaluate the convolution integral.
It is concluded that if certain general conditions are met that the first method should provide satisfactory results. If these conditions are not met, or if the estimator has reason to believe the final distribution will not be normal, it is necessary to use either method 2 or 3.

Most of the work that has been done up to now in developing cost estimating models has been based on the assumption that the individual cost components are independent or that if correlation does exist, it can be neglected. It has been pointed out in the text and through the examples in Appendix 1 that correlations could have a significant effect on the final distribution and it may be unconservative to neglect them.

5.2 **Recommendations for Further Study**

This author has concluded that there are three major areas that deserve further consideration before a final form of a cost estimating model is decided upon.

The first area is that of subjective probability assessment. As has been pointed out previously, this area has been studied in the context of risk assessment in investment decisions and much of the work should be directly applicable to cost estimating. Along somewhat the same lines, it would be desirable to combine both objective data and subjective analyses. In other words, an estimator might have historical cost data available, but he might want to modify this data based on his analysis of the present market situation. This might be done with Bayesian statistics (Reference 2, Chapter 5).
This "theory" allows one to choose a "prior distribution" purely on the basis of judgment (for cost estimating this might be a subjective analysis of present market conditions) and then to update this distribution (it becomes a "posterior distribution") by incorporating new information (which might be historical cost data).

The second area of further study involves the numerical integration techniques that are being developed for IGRAM at the Sloan School of Management at M.I.T. It would be desirable to be able to compare a model based on Monte Carlo simulation with a model based on a numerical evaluation of the convolution integral to determine if under certain conditions one might be better than the other.

Finally, the subject of component correlation needs further investigation. Very little seems to have been done in this area, but it has been pointed out that component correlation can have a significant influence on the total cost distribution. It would be desirable first of all to know if there are significant correlations among various cost components and, second of all, to determine how the correlations can best be assessed and finally included in a probabilistic cost estimating model.
Figure 1.1 PDF's of Two Alternative Projects.

Figure 2.1 CDF of Input for Simulation Model.

Figure 2.2 PDF of Input for Simulation Model.
Figure 3.1 PDF's and CDF's Showing Possible Effects of Correlation.
Figure 4.1 Normal Distribution.

Figure 4.2 Two Parameter Lognormal Distribution.

Figure 4.3 Three Parameter Lognormal Distribution.
Figure 4.4 Lognormal Distribution Skewed to the Right.

Figure 4.5 Graph of $f(s^2)$ for the Central Value equal to the Mode.
Figure 4.6 Graph of \( f(s^2) \) for the Central Value Equal to the Mean.

Figure 4.7 Beta Distribution.
Figure 4.8 Example of Fitting a Normal Distribution through Three Points.

Figure 4.9 Graph of $D$ versus $V$ for Choice of Central Value Corresponding to the Mode or the Median.

Figure 4.10 Graph of $q_\alpha$ versus $\alpha$. 
Figure 4.11  Graph of $D$ versus $V$ for the Normal Distribution.

Figure 4.12  Graph of $D$ versus $V$ for the Lognormal Distribution.

Figure 4.13  Graph of $D$ versus $V$ Showing Effect of Skewness.
List of References


List of References Continued


Appendix I

Examples of Possible Model Forms

Three computer programs were written for use on the IBM 370 computer at the Information Processing Center at M.I.T. The first program covered the case of assuming a normal distribution with zero correlation; the second covered the case of assuming a normal distribution, but permitted correlations to be introduced; the third was a Monte Carlo simulation with zero correlations. At the outset, it was planned to include a fourth program—that of a Monte Carlo simulation that included correlation—but the program was not debugged in time to include in the examples.

Three examples are presented that illustrate the discussion in the main part of the thesis. In all the examples the total cost is of the form,

\[ C = \sum_{i=1}^{n} Q_i (L_i + M_i) \]

where \( C \) is the total cost, \( Q_i \), \( L_i \) and \( M_i \) are the quantity, labor unit wage rate and material unit price, respectively, of cost component \( i \). For the cases where correlation is included, it is assumed to be such that \( Q_i \)'s are correlated, the \( L_i \)'s are correlated, and the \( M_i \)'s are correlated, but that there is no cross correlation—that is, no \( Q_i \) is correlated to any \( M_i \) or \( L_i \) and so on.

Example 1

The 25 cost components in this example were taken from an actual estimate made by a professional estimator. The variances, however, were introduced by this author. The input data for each of the cost
components is shown in Table A1.1. This example is characterized by the fact that there is a moderately large number of variables, none of which are dominant. The results of the programs are shown in Fig. A1.1.* There are two interesting points in these results. The first is that there is no significant difference between the method of assuming a normal distribution and the Monte Carlo simulation. The second is that there is a significant difference between the case of assuming zero correlation and the case of assuming perfect correlation. For example, if we consider the 80 percentile of the CDF, we get a value of $818,000 for the case of zero correlation and a value of $866,000 for the case of perfect correlation. This represents a difference of approximately $48,000 or 6%.

Example 2

This example uses 7 cost components that represent the structural elements of a building. In contrast to example 1, this represents a rather gross estimate. The input data is shown in Table A1.2. This example is meant to show the effects of a relatively small number of variables. Also, as can be seen, two variables are dominant (components 4 and 6 account for approximately 92% of the total variance for the case of zero correlation). The results, which are shown in Fig. A1.2, show that even in this case the assumption of a normal distribution gives very good results. Also in this case, the effect of correlation is much less than that in Example 1. At the 80 per-

* Note that the results are shown on normal probability paper in which the ordinate is scaled such that if a distribution is normal the CDF will be a straight line.
centile we get a value of $47,350,000 for the case of zero correlation and a value of $48,350,000 for the case of perfect correlation.

**Example 3**

The data for Example 3 is the same as that for Example 2, except that the variance in component 6 has been increased so that it accounts for almost 97% of the total variance for the case of zero correlation. The input data is shown in Table A1.3 and the results are shown in Fig. A1.3. For this example, the assumption of a normal distribution begins to break down. This is to be expected, though, since there are a relatively small number of variables, one variable is dominant and strongly skewed to the left. Also, it can be seen that the effect of correlation is still relatively small, even though it is greater than in Example 2. At the 80 percentile we get a value of $53,250,000 for the case of zero correlation and a value of $54,600,000 for the case of perfect correlation, which is a difference of $1,350,000 or 2.5% as compared to $1,000,000 or 2.1% in Example 2.
Figure A1.1  CDF of Output of Example 1
Figure A1.2 CDF of Output of Example 2
Figure A1.3 CDF of Output of Example 3
### Table A1.1 Input Data for Example 1

<table>
<thead>
<tr>
<th>Component</th>
<th>Item</th>
<th>Type Distribution</th>
<th>Input Data and Parameters</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Footings:</td>
<td>Q</td>
<td>N</td>
<td>X_4</td>
<td>X_P</td>
</tr>
<tr>
<td>Concrete</td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>LN</td>
<td></td>
<td>15.00</td>
<td>18.00</td>
</tr>
<tr>
<td>Footings:</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forms</td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>LN</td>
<td></td>
<td>0.19</td>
<td>0.21</td>
</tr>
<tr>
<td>Footings:</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rebars</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>LN</td>
<td></td>
<td>0.15</td>
<td>0.17</td>
</tr>
<tr>
<td>Fdn. Walls:</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>LN</td>
<td></td>
<td>15.00</td>
<td>18.00</td>
</tr>
<tr>
<td>Fdn. Walls:</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forms</td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>LN</td>
<td></td>
<td>0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>Fdn. Walls:</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rebars</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>LN</td>
<td></td>
<td>0.15</td>
<td>0.17</td>
</tr>
<tr>
<td>Columns:</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conc.</td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>LN</td>
<td></td>
<td>15.00</td>
<td>19.00</td>
</tr>
<tr>
<td>Component</td>
<td>Item</td>
<td>Type Distribution</td>
<td>$X_b$</td>
<td>$X_p$</td>
</tr>
<tr>
<td>---------------</td>
<td>------</td>
<td>-------------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Columns:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forms</td>
<td></td>
<td></td>
<td>0.85</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
<td>0.17</td>
</tr>
<tr>
<td>Flat Slab:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15.00</td>
<td>18.00</td>
</tr>
<tr>
<td>Flat Slab:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forms</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
<td>0.17</td>
</tr>
<tr>
<td>Flat Slab:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rebars</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
<td>0.17</td>
</tr>
<tr>
<td>Slab on Gr:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15.00</td>
<td>18.00</td>
</tr>
<tr>
<td>Slab on Gr:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mesh, etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A1.1 (cont.)
<table>
<thead>
<tr>
<th>Component</th>
<th>Item</th>
<th>Type</th>
<th>Distribution</th>
<th>$X_L$</th>
<th>$X_P$</th>
<th>$X_H$</th>
<th>$m$</th>
<th>$S$</th>
<th>$M$</th>
<th>$\sigma$</th>
<th>$t$</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cure Slabs</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>187,000</td>
<td>1870</td>
<td></td>
<td></td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.015</td>
<td>0.0003</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft/$ft^2$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>N</td>
<td></td>
<td>0.10</td>
<td>0.20</td>
<td>0.015</td>
<td>0.004</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft/$ft^2$</td>
</tr>
<tr>
<td>Precast Panels</td>
<td>Q</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10,440</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>4.75</td>
<td>5.00</td>
<td>5.50</td>
<td>5.10</td>
<td>0.301</td>
<td>0.177</td>
<td>0.241</td>
<td>3.87</td>
<td>ft/$ft^2$</td>
</tr>
<tr>
<td>Precast Tees</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5800</td>
<td>29</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>2.25</td>
<td>2.50</td>
<td>3.30</td>
<td>2.73</td>
<td>0.442</td>
<td>0.012</td>
<td>0.397</td>
<td>1.60</td>
<td>ft/$ft^2$</td>
</tr>
<tr>
<td>2&quot; Conc. Topping</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5800</td>
<td>29</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.16</td>
<td>0.016</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft/$ft^2$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>0.13</td>
<td>0.14</td>
<td>0.17</td>
<td>0.15</td>
<td>0.017</td>
<td>-3.222</td>
<td>0.374</td>
<td>0.11</td>
<td>ft/$ft^2$</td>
</tr>
<tr>
<td>Face Brick &amp; Backers</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.710</td>
<td>6.71</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.10</td>
<td>0.29</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>1.22</td>
<td>1.40</td>
<td>1.70</td>
<td>1.45</td>
<td>0.19</td>
<td>0.038</td>
<td>0.179</td>
<td>0.39</td>
<td>ft/$ft^2$</td>
</tr>
<tr>
<td>Face Brick &amp; Re. Backers</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.85</td>
<td>5.85</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.20</td>
<td>0.31</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>1.39</td>
<td>1.60</td>
<td>1.92</td>
<td>1.45</td>
<td>0.21</td>
<td>0.330</td>
<td>0.148</td>
<td>0.24</td>
<td>ft/$ft^2$</td>
</tr>
<tr>
<td>Interior Block</td>
<td>Q</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>43,120</td>
<td>4312</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.90</td>
<td>0.20</td>
<td>-</td>
<td>-</td>
<td></td>
<td>ft$^2$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>0.49</td>
<td>0.53</td>
<td>0.63</td>
<td>0.55</td>
<td>0.06</td>
<td>-1.933</td>
<td>0.241</td>
<td>0.31</td>
<td>ft/$ft^2$</td>
</tr>
</tbody>
</table>

Table A1.1 (cont.)
<table>
<thead>
<tr>
<th>Component</th>
<th>Item</th>
<th>Type</th>
<th>Distribution</th>
<th>$X_{10}$</th>
<th>$X_p$</th>
<th>$X_{90}$</th>
<th>$m$</th>
<th>$S$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$t$</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blk. Backers Q RC Panels</td>
<td>Q</td>
<td>N</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10,440</td>
<td>1044</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Ft²</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>N</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.90</td>
<td>0.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>$$/Ft^2$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>0.50</td>
<td>0.53</td>
<td>0.60</td>
<td>0.55</td>
<td>0.04</td>
<td>-2.042</td>
<td>0.293</td>
<td>0.41</td>
<td>$$/Ft^2$</td>
</tr>
<tr>
<td>4 Ply Roof</td>
<td>Q</td>
<td>N</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>62,400</td>
<td>626</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Ft²</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>0.04</td>
<td>0.05</td>
<td>0.07</td>
<td>0.054</td>
<td>0.012</td>
<td>-3.042</td>
<td>0.241</td>
<td>0.005</td>
<td>$$/Ft^2$</td>
</tr>
<tr>
<td>Gravel Stops</td>
<td>Q</td>
<td>N</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1250</td>
<td>125</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Ft</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>1.88</td>
<td>2.00</td>
<td>2.42</td>
<td>2.20</td>
<td>0.33</td>
<td>-0.717</td>
<td>0.545</td>
<td>1.64</td>
<td>$$/Ft$</td>
</tr>
<tr>
<td>Cant Flshing</td>
<td>Q</td>
<td>N</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.70</td>
<td>34</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Ft</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td></td>
<td>1.50</td>
<td>1.60</td>
<td>2.12</td>
<td>1.77</td>
<td>0.28</td>
<td>-0.898</td>
<td>0.547</td>
<td>1.30</td>
<td>$$/Ft$</td>
</tr>
</tbody>
</table>

$x_{10}$ = value at 10 percentile

$x_p$ = mode

$x_{90}$ = value at 90 percentile

$m$ = mean

$s$ = standard deviation

$\mu$ = lognormal parameters

$\sigma$ = lognormal parameters

$t$ = lognormal parameters

$Q$ = quantity

$L$ = labor unit wage rate

$M$ = material unit price

$N$ = Normal distribution

$LN$ = lognormal distribution

Table A1.1 (cont.)
<table>
<thead>
<tr>
<th>Component</th>
<th>Item</th>
<th>Type</th>
<th>Distr.</th>
<th>(X_e)</th>
<th>(X_p)</th>
<th>(X_h)</th>
<th>(m)</th>
<th>(S)</th>
<th>(\mu)</th>
<th>(\sigma)</th>
<th>(t)</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excavation</td>
<td>Q</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>LN</td>
<td>400,000</td>
<td>440,000</td>
<td>495,000</td>
<td>446,175</td>
<td>37,271</td>
<td>5,800</td>
<td>0.142</td>
<td>113,952</td>
<td>-</td>
<td>$</td>
</tr>
<tr>
<td>Caissons</td>
<td>Q</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>N</td>
<td>2,200,000</td>
<td>2,600,000</td>
<td>3,000,000</td>
<td>2,400,000</td>
<td>312,012</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>$</td>
</tr>
<tr>
<td>Sheeting &amp;</td>
<td>Q</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Shoring</td>
<td>M</td>
<td>LN</td>
<td>705,000</td>
<td>755,000</td>
<td>830,000</td>
<td>765,287</td>
<td>49,222</td>
<td>5.832</td>
<td>0.142</td>
<td>420,828</td>
<td>-</td>
<td>$</td>
</tr>
<tr>
<td>Concrete</td>
<td>Q</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>N</td>
<td>9,020,000</td>
<td>10,020,000</td>
<td>11,020,000</td>
<td>10,020,000</td>
<td>780,031</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>$</td>
</tr>
<tr>
<td>Fireproof &amp;</td>
<td>Q</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Insulation</td>
<td>M</td>
<td>LN</td>
<td>1,800,000</td>
<td>2,000,000</td>
<td>2,400,000</td>
<td>2,082,148</td>
<td>240,011</td>
<td>6.861</td>
<td>0.241</td>
<td>1,099,379</td>
<td>-</td>
<td>$</td>
</tr>
<tr>
<td>Structural</td>
<td>Q</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Steel</td>
<td>M</td>
<td>LN</td>
<td>25,642,000</td>
<td>26,702,000</td>
<td>29,072,000</td>
<td>27,241,961</td>
<td>1,390,762</td>
<td>8.453</td>
<td>0.280</td>
<td>22,374,906</td>
<td>-</td>
<td>$</td>
</tr>
<tr>
<td>Metal Deck</td>
<td>Q</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>N</td>
<td>2,514,000</td>
<td>2,797,000</td>
<td>3,088,000</td>
<td>2,797,000</td>
<td>228,749</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table A1.2 Input Data for Example 2
<table>
<thead>
<tr>
<th>Component</th>
<th>X_p</th>
<th>X_s</th>
<th>m</th>
<th>S</th>
<th>μ</th>
<th>σ</th>
<th>ρ</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excavation</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Carriages</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sheetng</td>
<td>Q</td>
<td>Q</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Forming</td>
<td>M</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>M</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Fireproof</td>
<td>M</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Insulation</td>
<td>M</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Structural</td>
<td>M</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Metal Deck</td>
<td>M</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

**Table A1.3 Input Data for Example 3**