MANUAL OF

RISK ANALYSIS IN PROJECT APPRAISAL

Jonathan Haughton

Department of Economics
Suffolk University
Boston, MA 02108

June 2000; edited May 2020

E-mail: jhaughton@suffolk.edu. Tel: (617) 573 8127.
GOALS, ORGANIZATION

The purpose of this manual is to provide the essential tools needed in order to identify, quantify, analyze and present the effects of risk on the outcomes of projects. These skills are needed in order

a. to determine whether a project should be undertaken, and especially to cull bad projects, and
b. to assess whether a project should be redesigned, particularly so as to allocate risk more efficiently.

The manual is designed for readers who have some knowledge of the techniques of deterministic project appraisal (i.e. cost-benefit analysis), and who want to move on to the next step, which is to incorporate risk into the analysis. It is written for the practitioner who wants to start using the techniques straight away, and so provides a succinct and commonsense treatment. Readers wishing to explore the subject in more depth will want to turn to the growing technical literature, including some or all of the following:

Vose ... [insert more here]

In the first chapter we explain why risk analysis is needed, and the role of sensitivity analysis, scenario analysis and Monte Carlo simulation, illustrating the techniques with a simple example.

All the basic ingredients are set out in the first chapter. There are however plenty of refinements that need to be made to the basic techniques, and these are set out in the subsequent chapters. At a few points there are important disagreements about methodology, and we identify these as they arise.

Like project appraisal itself, risk analysis is part science, part art. Both improve with practice. This manual attempts to distill that experience and convey in in a comprehensible way. We would appreciate your comments and suggestions for improvements.
1. INTRODUCTION: FUNDAMENTALS OF RISK ANALYSIS

Should a project be undertaken in its present form? This is the fundamental question that is addressed by project appraisal. The question may be broken down further by asking:

1. is the project financially desirable for each of the stakeholders, and
2. is the project economically and socially desirable.

The economic appraisal is not of concern to private participants, but is of interest to government and to policy makers. But no matter how desirable a project may be from society’s point of view, it will fail unless each stakeholder finds the project to be beneficial. This typically includes equity holders/owners, bankers who lend to the project, and direct participants such as farmers and workers.

Traditional project appraisal is deterministic. It projects revenues and costs into the future, and arrives at a single best point estimate of the effects of the project.

This deterministic estimate is inadequate as a guide to the financial viability of a project, because it does not give a sense of how risky the project is. Yet risk is very much on the mind of those involved in most projects.

For instance, a farmer might be reluctant to plant high-yielding corn hybrids, even if they are likely to bring higher profits, because they may be associated with a higher risk of harvest failure. The deterministic project appraisal would show that this project is worthwhile; a risk-adjusted analysis might predict that it would fail.

Given the additional information on risk, we may have to recommend against the project. Or we may want to think creatively about redesigning the project – for instance, adding crop insurance in the above example – so that risk is both diversified away, and pushed onto those who can most afford to bear it.

There are three main approaches to measuring the risk associated with a project: sensitivity analysis, scenario analysis, and Monte Carlo simulation. It is easiest to discuss each in the context of a simple example of project appraisal. We will conclude that only intellectually and practically satisfactory approach is to use Monte Carlo simulation, and the rest of this manual is devoted to explaining how to apply the techniques effectively.

**Example 1: Planting high-yielding rice**

Consider the following project, which we use as an example for the rest of this chapter. A farmer is considering planting a new rice hybrid on a hectare of land. By purchasing seeds from a reputable source (at the cost of $20 per hectare), applying more water (cost of $20), working longer hours (an extra $16 worth of time) and applying an extra 0.35 tonnes of fertilizer, the farmer can expect to get an extra 1.5 tonnes of paddy rice. Fertilizer is expected to cost $220 per tonne, and the farmer believes that the paddy rice can be sold for $120 per tonne. Most of the inputs need to be purchased in January, when the crop is planted out, while the revenue will accrue in May when the crop is sold. The details are set out in Table 1, and show that if the cost of funds for the farmer is 20%, then the project has a positive net present value of $37.45.
Since the net present value of the project is positive, the normal recommendation would be to go ahead with the project.

**Sensitivity analysis**

But wait. This project is risky. What would happen if the additional yield of paddy does not live up to expectations? Or if the price of rice is lower than expected? Or fertilizer costs more than anticipated? Would changes in these variables – of a realistic size – be enough to turn this project into a loss-making proposition?

A straightforward way to answer this question is with sensitivity analysis, which measures the effects of a change in each project variable on the net present value. The technique is straightforward: change each variable in turn, and measure the effect on the overall NPV. From this we may learn which variables are most crucial to the success of the project, and whether there are realistic cases where the project may fail.

The results of a sensitivity analysis for the high-yielding rice example are shown in Table 2. For each variable we pick a plausible set of values, and for each of these we recompute the NPV. From this analysis we learn:

a. some variables are not very important to the analysis of risk. For instance, if the cost of water were $28 instead of $20, the NPV would fall from $37.45 to $29.68, but the project would still have a positive NPV and so be accepted. These unimportant variables can generally be left out of consideration in the more elaborate Monte Carlo simulation exercises.

b. Other variables are crucial. In this example, if the rice yield rises by 1 tonne rather than 1.5 tonnes per hectare then the NPV will fall from $37.45 to -$19.01, and we would reject the project. Clearly any further analysis of risk in this project will need to concentrate on this variable (and, to a lesser extent, on the prices of paddy rice and of fertilizer).

c. The sensitivity of NPV to a change in a variable depends on
   - The importance of the variable in computing NPV, and
   - The range of plausible values.

For instance, the price of paddy and the yield of paddy have comparable effects on NPV, but the results are more sensitive to the yield because it is likely to vary much more widely than is the price.

It is sometimes useful to compute an elasticity, which measures how much NPV changes (proportionately) when a variable changes (proportionately). More formally,

\[ \eta = \frac{\% \text{ change in NPV}}{\% \text{ change in variable}}. \]

For instance, suppose that for padi rice the elasticity is 3.20, which means that a 10% increase in the price of rice would raise the NPV of the project by about 32% (i.e. 3.2 times 10%). In Table 2, note the small and negative elasticities for the costs of seeds, labor and water; the negative sign indicates that higher values of the variable lead to a lower NPV.

Although often helpful, sensitivity analysis has some serious weaknesses. These include:
1. It changes the variables one at a time. This is not how the real world works, where prices and quantities and costs all tend to change together.

2. The results are usually fairly obvious. It comes as no surprise to learn that the yield of paddy is crucial to the success of the project.

3. The analysis does not give a sense of the probability of each outcome. For instance, how likely is a yield of 1 tonne (rather than 1.5 tonnes)? Of 2 tonnes? Unless we know the probability of each outcome, the sensitivity analysis is of no help as a guide to action; it gives us possibilities, not probabilities.

**Scenario Analysis**

Some analysts have responded to the problems of sensitivity analysis by computing NPVs for a range number of different plausible scenarios. It is quite common to see forecasts and project appraisals that report an optimistic case, a most likely case, and a pessimistic case. In Table 3 we report the results of a scenario analysis for the high-yielding rice example. The most likely outcome is a positive NPV of $37.45. But if everything turns out worse than hoped for – lower yields, a lower price of rice, a higher price of fertilizer, etc. – then the NPV would be -$49.77. And if all goes exceptionally well, then the NPV would be $135.96.

There is one case when the scenario analysis is useful. If even the worst-case NPV is positive, then the project should be accepted. However, in practice this is rare, in which case the exercise is unhelpful, because we do not know how probable each scenario is.

**Monte Carlo simulation**

The solution to all these problems is to undertake a simulation analysis. The technique is powerful, which explains its increasing popularity. The essential idea is to proceed as follows:

a. Build a spreadsheet that appraises the project, as one would do for the traditional deterministic analysis;

b. Using a software such as CrystalBall or @Risk, allocate a probability distribution all of the most important variables in the analysis. Thus, for instance, instead of assuming that the price of paddy will be $120 per tonne, we might instead assume that the price of paddy is normally distributed, with mean $120 and standard deviation $12. This would mean that 68% of the time, the price of paddy will be between $108 and $132 (using the empirical rule that for a normal distribution, 68% of the values are within one standard deviation of the mean).

c. Include any correlations that are considered to be important. For example, the price of rice might be positively correlated with the price of fertilizer, because when there is a currency devaluation, the price of both rises (in local prices).

d. Simulate the outcome of the project. In other words, get values for each variable, from their distributions, and compute the NPV. Then do this again, and again, and again, perhaps a thousand times. The result will be a distribution of NPV, which can be analyzed.
We now apply these steps to the high-yielding rice example. Try to follow the gist of the analysis; in subsequent chapters we will back up a bit and fill in the details.

**Example. A Monte Carlo simulation analysis of the high-yielding rice project.**

Let us assume that

- the price of paddy rice is normally distributed, with mean $120 and standard deviation $12. In other words, it comes from a distribution like the one shown in Figure 1. We will return to the question of whether this is an appropriate distribution, but for now it is enough to suppose it looks plausible.
- The price of fertilizer is normally distributed, with mean $220 and standard deviation $24. The general form of the distribution is like the one in Figure 1, but the mean and standard deviation will be different.
- Xxx output of padi
- The price of paddy is positively correlated with the price of fertilizer, and that the rank correlation coefficient is +0.5. In other words, when the price of paddy is higher, so is the price of fertilizer.

We then simulate the results a thousand times, using CrystalBall. This generates a distribution of NPV values. There are several ways of viewing the result: as a probability distribution as in Figure 2a, or a cumulative probability distribution as in Figure 2b, or from the statistics reported in Table 4.

From the results we learn that there is a 36% probability that the project will have a negative NPV. This is likely to create a problem for the risk-averse rice farmer, and raises doubts about the viability of the project, even though the expected NPV is positive.

To illustrate the power of Monte Carlo simulation, consider a modification to the project. Suppose that it includes an insurance provision. For a premium of $15, the National Cereal Bank will provide the farmer with 1.3-X tonnes of paddy whenever X<1.3, where X is the actual paddy output. If the harvest is good, the farmer receives nothing, but if the harvest is bad, the Bank will make up the difference (up to 1.3 tonnes).

From the point of view of the farmer, this lowers the expected NPV of the project by $15, to $22.45. However, it also reduces the probability of a negative NPV to xxx%, which might make an otherwise risky project attractive enough to undertake.

**What’s next?**

The discussion so far provides the bare bones of Monte Carlo simulation. But there are plenty of loose ends that need addressing, which we do in the chapters that follow. Among the key issues are the following:

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2</td>
<td>What distributions are appropriate for the variables in the risk analysis? What alternatives are available? How can one fit distributions? How may one generate distributions using expert opinion?</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>How are correlations estimated? What role does regression analysis play in specifying linkages between variables? What role does economic theory play in modeling these relationships?</td>
</tr>
</tbody>
</table>
Chapter 4: How may one forecast not only the values of a series, but also the distributions for these values?
Chapter 5: What is the best way to present the results of risk analyses?
Chapter 6: How may risk analysis be used to evaluate different forms of contracts?
Chapter 7: What is the appropriate discount rate to use in a risk analysis? Or put another way, how does one put a price on risk?

We now turn to these questions one by one.
3. FITTING DISTRIBUTIONS

One of the first steps in Monte Carlo simulation is to pick a probability distribution for each important variable. Some analysts believe this step to be very important; Vose (1996, p.51) writes that “in my experience, inappropriate use of probability distributions has proved to be a very common failure of risk analysis models.” On the other hand, Glenday (xxx) believes that the precise choice of probability distribution is less important than formulating the overall model appropriately, and coming up with a satisfactory measure of the variability of the variables.

At the outset it is worth emphasizing Vose’s golden rule: all distributions used must be plausible. Xxx p. xxx.

We begin by introducing some basic terminology. Then we introduce a number of important probability distributions. The most important part of the chapter deals with the problem of how to choose among the different possible distributions, and how to pick the appropriate parameters.

Terminology

Suppose we are interested in knowing how the price of rice is likely to vary in the future. A simple approach would be first to graph the price of rice over time, as is done in Figure xxx. If we assume that the price in one year is not influenced by the price in the previous year (i.e. the observations are independent), then we may gather together these observations into a frequency distribution, as shown in Figure xxx. In this case we show just four bars (“bins”), but we could vary the number of bars is desired.

By changing the scale, so that the area in the bars sums to 1, we construct a probability distribution. In our example, we see that the probability that the price of rice would be between xxx and xxx is 0.xxx. Note:

• the probability that the price falls into any range cannot be negative; and
• the sum of the probabilities of all possible events must equal 1.

From the probability distribution it is straightforward to construct the cumulative probability distribution. The calculations are shown in Table xxx, and the graph shown in Figure xxx.

The mean price of rice is given by \[ \mu = \frac{1}{n} \sum_{i=1}^{n} X_i \]. To measure the estimated dispersion in the price of rice, it is common to calculate the variance, which is calculated as \[ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 \]. This is often a large number, and to put it in the same units as the mean it is common to calculate the standard deviation, which is the square root of the variance; it essentially measures the average distance of values from the mean. Sometimes it is helpful to compute the coefficient of variation, defined as \( \frac{\sigma}{\mu} \). These measures are computed for the price of rice in Table xxx.

If we do not have the original data, but only a probability distribution, then we calculate the expected value of the distribution as \[ \sum_{i=1}^{n} pX_i \]. This is the analog of the mean for the underlying observations. Similarly, the variance
of a distribution may be calculated as $\sum_{i=1}^{n} p_i X_i^2 \mu^2$. The detailed calculations are shown in Table xxx for the frequency distribution of rice prices shown in Figure xxx.

### A Gallery of Probability Distributions

Probability distributions may be discrete or continuous. A discrete distribution allows only certain identifiable values. For instance, the number of visitors to a tourist attraction must be an integer, since one cannot have half a visitor. Continuous distributions allow any value within a range; examples include the yield of a crop, or the exchange rate, or the height of a person.

Probability distributions may be bounded or unbounded. An unbounded distribution may take on any value from minus to plus infinity. The normal distribution is unbounded, and in some cases it may be necessary to truncate it; for instance if we assume that crop yields are normally distributed, we may need to add a constraint that ensures that the crop is never negative. A number of useful distributions are partially bounded, and may only take on positive numbers (e.g. the lognormal distribution).

Vose makes a distinction between distributions that are theoretically derived and those that are more ad hoc ("general non-parametric distributions" in his terminology). For instance, both theory and experience suggest that people’s heights are normally distributed, so the normal distribution is appropriate here. On the other hand, we have no strong prior beliefs about how to model the distribution of, for instance, the probability of a drought, and so we may want to model it in a general, non-parametric way. Vose argues that general, non-parametric distributions are "intuitively easy to understand, extremely flexible and are therefore very useful," while theoretically-derived distributions are generally harder to justify.

Figure xxx shows the gallery of distributions that CrystalBall allows one to use in Monte Carlo simulations. Some additional comments are in order on the most useful of the distributions.

#### a. The Custom distribution.

Also known as the general, or custom, or step, or relative distribution, this is an entirely flexible distribution when the analyst specifies the probability of values falling into each bin. It is very useful in practice, particularly in reflecting the opinions of experts. It is sometimes written as General(min,max,{xi},{pi}); in this form it is a discrete, unbounded, non-parametric distribution.

An example of a custom distribution is shown in Figure xxx, and the associated cumulative custom probability distribution in Figure xxx. This example graphs General(0,5,{1,2,3,4},{6,12,6,1}).

A close relative is the Discrete distribution, which is a general function used for a variable that can take one of several discrete values, such as the number or doors on a car or kittens in a litter. It is written as Discrete({xi},{pi}), and an example is shown in Figure xxx.
The distribution is particularly useful in two important cases, which are discussed further below:

(i) when combining the differing opinions of experts; and

(ii) when modeling probabilistic branching. For instance, suppose a farmer believes that the yield of rice, in tonnes per hectare, is lognormal(1.6,2) provided that there is no saline intrusion. However, the farmer believes there is a 20% probability of saltwater entering the rice paddies, and seriously reducing the crop. In this case the yield is lognormal(0,.3). One could now model the rice yield as

\[
yield = \text{Discrete}\{\text{Lognormal}(1.6,2), \text{Lognormal}(0,.3)\},\{80\%,20\%\}).
\]

b. The Normal and Lognormal distributions

Sometimes referred to as the Gaussian distribution, or bell curve, the Normal distribution is useful in many applications. It applies to many human characteristics (e.g. height, weight), and is widely used in finance to capture the distribution of returns or prices.

The Normal distribution is a continuous, unbounded, “theoretical” distribution. It is entirely determined by just two parameters, the mean (\(\mu\)) and standard deviation (\(\sigma\)), and is typically written as \(N(\mu, \sigma)\). Figure xxx shows the probability density function (pdf) of the Normal distribution for \(N(3,2)\), while the associated cumulative density function (cdf) is given in Figure xxx. Note that in this case, xxx% of the values are negative.

There is a probability of approximately 68% that the values of the variable will like within one standard deviation of the mean; the probabilities are 95% and 99% respectively that the values will be within two, or three, standard deviations of the mean.

One reason that the Normal distribution is so common is because of the central limit theorem. Consider a variable \(X\), which may have any distribution at all. Now take a sample of values of \(X\) and compute the mean, giving \(x\bar{1}\). Repeat this process many times. Then one will see that the \(x\bar{1}\)s are distributed normally. This is a remarkable, and useful result; any variable that itself is the sum of other variables will be normally distributed.

If \(\ln(X)\) is distributed normally, then \(X\) is distributed lognormally. The Lognormal distribution is continuous and partially bounded – negative values are not possible. It is appropriate whenever a variable may be considered to be the product of other variables. [fn. From the central limit theory, the product of a large number of probability distributions is lognormally distributed.] For instance, the volume of gas in a reserve is often lognormally distributed because it depends on the product of the area of the formation, its thickness, formation pressure, porosity and the gas/liquid ratio (Vose, p.80).

The Lognormal distribution is also useful in representing variables that range from zero to +infinity, and are positively skewed. It has been used to model the length of words and sentences in a document, and the incubation period of infectious diseases. Two examples are graphed in Figure xxx.
c. **The Uniform distribution**

The Uniform distribution stipulates that, over a given range, every value is equally likely to be chosen. The commonest version is continuous, and the distribution is usually written $U(\text{min}, \text{max})$, and has a mean $(\text{min} + \text{max})/2$ and variance $((\text{max} – \text{min})^2)/12$. Although straightforward, the Uniform distribution is not very useful in practice, because it is rare that variables are equally lightly throughout a limited range but do not occur anywhere else; the transition is simply too abrupt. There is also a discrete version of the Uniform distribution, but other discrete distributions (e.g. Binomial, Geometric, Poisson) are almost always more appropriate.

Figure xxx gives two examples of Uniform distributions, and Figure xxx graphs the cumulative distribution function for $U(10,20)$.

d. **The Triangular distribution**

It is common to ask an expert for an assessment of the maximum, minimum and “most likely” values. An intuitively attractive way to map this into a distribution is with the Triangular (or Triang) distribution, which is written as $\text{Triang}(\text{min}, \text{most likely}, \text{max})$. By way of an example, the $\text{Triang}(2,5,7)$ distribution is graphed in Figure xxx, and its associated cumulative distribution function in Figure xxx.

The Triangular distribution is widely used, because it is quick and easy to use. However, Vose argues that it “usually overemphasizes the tails ... and underemphasizes the shoulders in comparison with other, more natural, distributions” (p.89).

A variant on this distribution is the Trigen distribution (see Vose p.169). It is determined by five parameters:

- $a$: the practical minimum
- $b$: the most likely value
- $c$: the practical maximum value (“most likely value”)
- $p$: the probability that the variable could be below $a$; and
- $q$: the probability that the variable could be above $c$.

An illustration is shown in Figure xxx, for a $\text{Trigen}(10,25,50,.1,.05)$. The distribution may be computed automatically using @RISK, but requires a little extra computation when used in CrystalBall. [Exercise: What is the Triang distribution that is generated by $\text{Trigen}(10,25,50,.1,.05)$?]

e. **The Beta and BetaPERT distributions**

One of the most flexible, and practically useful, of all distributions is the Beta distribution. It is determined by just two parameters, and may be written as $\text{Beta}(\alpha_1, \alpha_2)$. Some examples of the Beta distribution are given in Figures xxxx, xxxb and xxxc.
At the theoretical level, the Beta distribution measures the probability that an event will occur, given that we know that there were $r$ successes in $n$ trials. In this case $\alpha_1 = r + 1$ and $\alpha_2 = n - r + 1$. For example, suppose that in 10 attempts to grow carrots, the crop failed entirely twice. Then the Beta(3,9) would give the probability that a given carrot crop would fail.

If there is a variable that can take on values between a minimum of $a$ and a maximum of $c$, then the Beta distribution can be rescaled using

$$a + \text{Beta}(\alpha_1, \alpha_2) \cdot (c-a).$$

Let $b$ be the “most likely” value. Then the BetaPERT(a,b,c) distribution is given by setting $\alpha_1 = \ldots \cdot [1.17]$. Some examples of BetaPERT distributions are shown in Figure xxx. These look a bit like smoothed versions of the Triangular distribution, with more weight on the shoulders and less on the extremes. This makes the BetaPERT distribution very useful when one has information on the minimum, most likely and maximum values of a variable – derived, for instance, from expert opinion.

### f. The Binomial distribution

This discrete distribution shows the number of successes from $n$ trials, where there is a probability $p$ of success for each trial. It is written as Binomial(n,p). For instance, suppose a firm is making computer components, and knows that there is a 1% probability that the component will not work; then if it ships 5000 components, how many will not work (and so have to be replaced, possibly at a high cost)? This may be modelled by the Binomial(5000,1%) distribution shown in Figure xxx. For $n$ greater than about 30, the Binomial can be approximated by the Normal(np,(npq)^.5), where $q = 1-p$.

The biggest problem with using the Binomial distribution in practice is that it assumes that the probability of success (i.e. $p$) does not change. This means that the firm does not improve its performance over time, or with experience, which is unrealistic.

There are numerous other distributions, but most of them are only applicable in specialized settings. Vose (1996) provides a clear and complete treatment.

### Choosing the Appropriate Distribution

There are two ways to pick a distribution for a variable: using available data, or referring to expert (including the analyst’s) opinion.

Programs such as CrystalBall and @Risk allow one to fit a distribution to a set of numbers. This is convenient, perhaps too convenient because of the danger that one will rush into fitting an inappropriate distribution. Before fitting a distribution, it is helpful to work though the following checklist (based loosely on Vose, pp.104-107):
a. Is the variable to be modelled discrete or continuous?

Prices, crop yields, exchange rates, the likelihood of a drought – these are all continuous variables. On the other hand, the number of ships built by a shipyard in a year might be better modelled as a discrete variable.

b. Is it better to use an empirical or a theoretical distribution?

Particularly when the number of data points is small, it is often better simply to use a Custom distribution, based on the observations directly.

[Add example.]

On the other hand, if there is a theoretical distribution that is known to fit the mathematics of the model, then there is a strong case for using such a distribution. This helps one decide when, for instance, to use the Binomial distribution. A standard discussion of the different distributions, and the circumstances under which they may be applied, is likely to be helpful here.

And if there is a theoretical distribution that is known to fit the distribution in practice, then this is likely to be suitable. The Normal distribution is often applied on the grounds that it is known, from experience, to fit the data well. For this the analyst is likely to need to have a solid knowledge of the relevant academic literature.

c. Does the fitted distribution cover the theoretical range of the variable?

Sometimes a distribution may extend beyond the range of a variable, and need to be truncated. For instance, the Normal distribution can take on negative values, and so may need to be constrained to be non-negative when it is used (for instance) to model the distribution of prices. Actual observations rarely lie at the extremes of a distribution, and so the fitted distribution should usually extend beyond what is measured.

D: Is the variable correlated with other variables?

In order to fit a stand-alone distribution to a variable X, we are assuming that X is not correlated with other variables of interest. If this assumption is not valid, then it is essential to model the links between variables, which is the subject of chapter 3.

Using data to fit a Custom distribution

Suppose that we have a small number of observations – perhaps 10 or so. In that case we may want to use, as our distribution, the actual distribution that we see for the observations.

Example: We observe the following prices of paddy rice (per ton) over a ten-year period:

100 106 108 111 118 124 125 129 137 145
We could form a histogram:

<table>
<thead>
<tr>
<th>Range</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-109</td>
<td>3</td>
</tr>
<tr>
<td>110-119</td>
<td>2</td>
</tr>
<tr>
<td>120-129</td>
<td>3</td>
</tr>
<tr>
<td>130-139</td>
<td>1</td>
</tr>
<tr>
<td>140-149</td>
<td>1</td>
</tr>
</tbody>
</table>

This can now be used, in CrystalBall or @Risk, to specify a Custom distribution. Note that the price of paddy rice is a continuous variable, so we need to put the observations into bins. The choice of bins calls for some judgement. For instance, in the foregoing example, one could have had bins that started at 95-, 105-, etc., or smaller or larger bins. There is no uniquely correct way to do this. This procedure also works when we have a large number of observations.

Occasionally we have data that are already categorized. For instance, based on household survey data we might be told that when they applied an extra 80 kg of fertilizer, 30 farmers got an extra 50-99 kg of rice, 92 farmers got an additional 100-199 kg, and 71 farmers increased their yields by 200-299 kg. This makes it easy to create a Custom distribution.

Using data to fit a theoretical distribution

Consider again the problem of fitting a distribution to the price of rice. If we have enough data – at least 30 observations as a rule of thumb – then we may be in a good position to fit one of the theoretical distributions (e.g. Normal, Lognormal), and then use the fitted distribution in our risk analysis.

In CrystalBall, the procedure for fitting a distribution is straightforward. Put the actual observations in a column, ... [explain procedure, with an example.] It is tempting to ask CrystalBall to fit the best possible distribution, and we know that, like Alice in Wonderland, you will not resist this temptation. In the risk analysis itself however, it is important to use only a fitted distribution that makes good sense on theoretical as well as commonsense grounds.

Using Expert Opinion

Sooner or later every risk analyst needs to consult experts, especially when evaluating a project in a relatively unfamiliar area. Hard data may be unavailable, or too expensive to obtain, or obsolete, or incomplete, or simply irrelevant because the project is so innovative. Under these circumstances we are obliged to turn to experts.

For example, if the project involves developing community woodlots, we need to know what tree species grow well, and how fast, and how resistant they are to drought and disease; who would protect the trees as they grow; who would be able to harvest them, and when; how great is the risk of fire; what are the other possible uses of the land; and so on. For each of these variables we want to fit a distribution, and we will need to elicit this information from experts.
Be careful. Using expert opinion is harder than it looks, so before explaining how to turn the opinions into probability distributions, it is worth setting out a few cautions.

a. Do not assume that, as the analyst, you are best person to estimate all or even most of the variables in the model. Put another way, recognize your limitations, and listen genuinely to the opinions of experts.

b. Even expert opinion is subject to potentially serious biases. Among the most important are the following:

- **Availability.** An expert will base her estimates on her recollection of past events. If the event occurs regularly, then the recollection will be good. If the event was dramatic, such as a forest fire, then she may actually remember too well, and overestimate the probability of such an event. And if the event was widely reported, it may also get too much weight; many people believe, incorrectly, that it is more dangerous to travel a mile by plane than by car.

- **Anchoring.** Most individuals estimate uncertainty by first assessing the most likely value (the anchor), and then making adjustments for the minimum and maximum, using the anchor as a reference. These adjustments are typically insufficient, creating a range that is too narrow, and causing over-confidence in the estimates. Classroom exercises, where participants are asked to estimate the minimum, most likely, and maximum values for magnitudes (the length of the Nile, the area of New Zealand, etc.) routinely underestimate their ignorance!

- **Experts are human too.** This means that:
  - Some experts are not in fact expert, but may still try to come up with estimates “to be helpful.”
  - Organizational culture matters. Sales people tend to come up with optimistic estimates of sales; production managers may be pessimistic about what can be achieved, in order to leave a margin of error.
  - Experts may have a vested interest in the result. For instance, an expert might provide optimistic estimate of the cost of building a refinery, in the hope that her firm will land the contract to build it (and yet be able to bill for cost overruns). Of course, it is not good practice to ask for expert opinions from anyone who has a stake in the outcome of the analysis.
  - Second guessing. Many experts try to provide the answers that the analyst seems to want.
  - Some experts are too busy to do a serious job of providing estimates; here it is important to get managerial support for the risk analysis exercise.

c. When using expert opinion, be particularly careful to getting the units of measurement correct. For instance, a short ton (2,000 pounds) is not the same as a long ton (2,240 pounds), which differs from a metric tonne (2,204 pounds, or 1,000 kilos), so a ton is not a ton is not a tonne; and an English gallon is 25% larger than an American gallon.

---

**Converting expert opinion into distributions**

It is rare that experts are able to specify a theoretical parametric distribution (e.g. Normal, Binomial) that represents the distribution of a variable.

As a practical matter, it is usually much easier to get an expert to specify a minimum value, most likely value, and maximum value. A simple way to model this information is by using a Triangular distribution, which is the main reason for the popularity of the Triangular distribution. However, there are two problems here.
• First, the Triangular distribution is too thick on the shoulders, and too fat on the tails. This problem may be solved by using a BetaPERT distribution, as explained above (see page xxx).

• The second problem is that it can be hard to get an accurate estimate, from an expert, of the maximum and minimum values of a variable; typically the expert will give a “practical minimum” rather than an absolute minimum. One way to deal with this is to apply a Trigen distribution.

Sometimes experts are able to set out the whole distribution, in which case it is often easiest to use a Custom distribution from the very beginning.

Example. The ShipR corporation sells goods using the Web. Customers submit orders, which are checked by the sales office, passed on to the warehouse for fulfilment, and sent out by package mail. The company wants to know how long is the interval, for the customer, between the moment when an order is submitted and the time when the package arrives on the doorstep. More specifically, the company wants to know what interval is long enough for 90% of the orders to be completely fulfilled; it would like to be able to advertise with a slogan like “from order to delivery in 27 hours!” ...

It is relatively straightforward to model the opinions of a single expert, but what if there are several opinions, and they differ? In this case one needs to either reconcile, or aggregate, these differences of opinion.

Aggregating expert opinion

In aggregating the views of experts, the best approach is to use a Discrete distribution, to combine the distributions (e.g. Triang, BetaPERT) that reflect the opinions of the individual experts; heavier weights may be put on the opinions of experts that are believed to be particularly knowledgeable or trustworthy.

There are some pitfalls to be avoided here. Do not simply pick the most pessimistic estimate; the purpose of risk analysis is to come up with an unbiased estimate of risk, not a worst-case scenario. And do not take an average of the expert opinions, as this will create too tight a distribution. Vose (p.180) gives a clear example: if two experts agree that a variable is distributed Normal(100,10), then this is the distribution to apply; however, the average of two normal distributions would give Normal(100,10/\sqrt{2}), which has too little variation, and so is incorrect.

Example. xxx

Reconciling differences of opinion

Very often, when experts hold different opinions, it is because they differ in their appreciation of the problem at hand. In this case it is frequently productive to follow a two-step procedure – similar to the Delphi technique pioneered by the RAND corporation – which begins with a brainstorming session and follows with individual interviews.
The purposes of the brainstorming session are to ensure that all of the experts have the same relevant information, and to provide an opportunity to discuss the uncertainties involved in the problem. The session will be more efficient if any relevant written information is circulated in advance; the meeting concentrates on discussing the uncertainty of each variable; the session is structured; and minutes are taken and circulated afterwards to participants.

As a practical matter, it is very difficult to get a group of people to agree collectively on the appropriate distribution to fit to a variable. The loudest voices tend to get too much weight, and the lack of agreement can cause the meeting to drag on interminably. The solution is to interview the experts individually after the meeting. By then each expert will have a good sense of the problem, will have had time to reflect on it, and will have formed her own opinion; the opinions of soft-spoken experts can then be given the weight they are due. Some experts are reluctant to come up with a distribution of values, but can usually be persuaded to specify minimum, maximum and “most likely” values; others may be coaxed into providing fuller information, and it is often helpful to begin to sketch a plausible distribution during the interview, even on a computer.
4. Modeling Dependencies

Variables are related to each other, and this is at the heart of making risk analysis both realistic and useful. For instance, the yield of rice will depend on the quantity of fertilizer that a farmer applies to the field. In analyzing a project that involved rice cultivation, it would not be appropriate to fit separate, and unrelated, distributions to rice yields, and to fertilizer applications.

xxx
5. Forecasting Variables: non-stochastic methods
6. Forecasting Variables: stochastic methods
7. Evaluating Contracts
8. The Cost of Capital debate
9. Presenting the Results
10. Conclusions