5 Variability, Uncertainty, and Probability

We wish truth, and we find only uncertainty.

Blaise Pascal

Risk is conventionally understood to be related to the concepts of uncertainty and probability (Chapter 1), because risk is derived from a sense that the future is imperfectly or incompletely known. Uncertainty concerning the future may be due to inherent randomness (e.g., quantum indeterminacy), effective limits on knowledge of future conditions even in deterministic systems (chaotic systems or simply the limits of data collection and analysis), or simple lack of knowledge (ignorance). The concept of probability and the related concepts of uncertainty, variability, likelihood, error, credibility, etc. are sources of confusion and controversy, which lead to linguistic uncertainty. Most environmental scientists are aware of two schools of statistics with different concepts of probability: Frequentist and Bayesian. Many are not aware that there are other approaches such as information-based statistics (Burnham and Anderson 1998, 2001) or evidence-based statistics (Taper and Lele 2004), or that many statisticians consider their field to be in need of a conceptual revolution (Gigerenzer et al. 1989; Salsburg 2001; Royall 2004). This chapter will avoid most of these issues by sticking to approaches that are in reasonably common use. Those who are uncertain about probability should consult Hacking's (2001) marvelously clear text. Those who want specifics about how quantitative methods are used in risk assessment might consult a text on quantitative methods in risk assessment such as Vose (2000), Burgman (2005), and Warren-Hicks and Moore (1998).

5.1 SOURCES OF UNPREDICTABILITY

We speak of risks and analyze probabilities because we wish to predict the future but realize that we cannot know the future. This is due to variability and uncertainty and to unquantified or unacknowledged factors.

5.1.1 VARIABILITY

Variability is the property of sets of entities or events that differ in some significant way. Examples include daily rainfall, weights of red foxes, or concentrations of a chemical in an effluent. Variability may be observed and estimated, but it cannot be reduced, because it is an inherent property of the system. Observations of a variable trait result in a distribution of frequencies of that trait. The response to variability is to form probabilities or equivalent expectations concerning that trait. Variability is an objective property of entities and events, so objectivist concepts and frequentist methods dominate its analysis. However, subjective

concepts and Bayesian methods are also used. In particular, not all variable traits are observed, so personal or expert judgment may be employed to estimate probabilities of variable traits.

5.1.2 UNCERTAINTY

Uncertainty is the lack of knowledge about a system. Unlike variability, it can be reduced by obtaining additional information. The responses to uncertainty are to form beliefs or suspend judgment until data can be generated. In probabilistic analyses, beliefs, like variability, should be expressed as distribution functions. However, they are commonly expressed qualitatively using terms such as likely, credible, and reasonably certain.

When mathematical models are considered to simulate a real-world phenomenon, and not simply describe a data set, issues of model uncertainty arise that go beyond simple lack of fit. Uncertainty concerning the parameters and forms of mathematical models is a difficult issue that is usually acknowledged without being quantified (Risk Assessment Forum 1996). Model uncertainty includes issues such as linearity, the inclusion of thresholds, the aggregation of species into trophic levels, and the inclusion of feedback processes. Like parameter uncertainty, model uncertainty can be reduced by research, but it is often more difficult than simply measuring a parameter. Research to resolve model uncertainties typically must address the mechanisms underlying the modeled processes. When there are multiple credible models, model uncertainty can be estimated from the variance among model results; otherwise it can be estimated from the variance due to changes in individual assumptions of a model (Gardner et al. 1980; Rose et al. 1991).

5.1.3 VARIABILITY UNCERTAINTY DICHOTOMY

The distinction between variability and uncertainty as aspects of unpredictability dates to the earliest writing on probability (Hacking 1975). The pair is more properly referred to as Aleatory (having to do with accounting) and Epistemic (having to do with how and what we know), but that terminology has not caught on (Hacking 2001). It is also tied to the two sources of probability: variability produces frequency-type probability and uncertainty produces belief-type probability, although the connection is not necessary or consistent. In particular, subjectivists treat all probabilities as beliefs, even those that arise from variability among entities or events.

The variability/uncertainty dichotomy has become conventional in environmental risk assessment (MacIntosh et al. 1994; McKone 1994; Price et al. 1996; Risk Assessment Forum 1997; Science Policy Council 2000; Linkov et al. 2001). It has practical as well as conceptual importance.

Variability is a property of the system while uncertainty is a property of the observer: This distinction is important to decision makers and stakeholders. In health risk assessment, this is reflected in the desire to know the range of responses of humans and the sources of that variability. Uncertainty concerning that variability is a secondary concern. Similarly, in ecological risk assessments we may be interested in the range of responses of populations or ecosystems due to variability in their exposure and sensitivity. We may then be concerned about our uncertainty concerning those estimates.

Variability is irreducible while uncertainty may be reduced by research: We cannot change the inherent variability in sensitivity to cadmium within a set of organisms or of species, but we can reduce our uncertainty concerning that variability by performing toxicity tests. Uncertainty about the magnitude of a constant such as the solubility of a chemical can be reduced as well. This distinction is clearly important when performing sensitivity analyses to decide how to allocate funds for measurement and research.

Estimates of variability can be directly verified but estimates of uncertainty cannot: Hence, if we estimate the distribution of a variable parameter, such as stream flow or invertebrate density, we can verify that estimate by taking measurements of flow or density over an appropriate temporal or spatial sampling frame. We can do the same for model-generated estimates. Consider, for example, a case in which we predict the proportion of fish species that will be affected if an effluent is released. We may estimate variability in exposure by toxicokinetic modeling and variability in sensitivity using a species sensitivity distribution. The resulting estimate of the proportion of species affected could be verified by monitoring if the effluent is permitted. In contrast, the uncertainty about a prediction cannot be verified, but a series of predictions may allow assessors to determine whether they have in fact over- or underestimated their uncertainties (Section 5.2.1).

These three aspects of the variability/uncertainty dichotomy correspond to different uses of assessments. These different uses could lead to treating different sets of parameters as variable or uncertain. For example, when performing a sensitivity analysis to prioritize a sampling or testing program, all parameters in the risk model that could be measured should be treated as uncertain, and uncertainty should be distinguished from irreducible variability. In contrast, to estimate a variable endpoint attribute, only those parameters that vary with respect to that attribute should be treated as variable. For example, flow varies over time, but, if the endpoint variable is the number of species responding, it would not be included as a source of variability, because it is the same for all species in a receiving stream.

5.1.4 Combined Variability and Uncertainty

The future is unpredictable because of the joint influence of variability and uncertainty. For example, we cannot predict the minimum flows in a stream because of variability in meteorology and hydrology and because of our uncertainty concerning the applicability of available data to the future (e.g., the climate is changing and the watershed is being developed). This combination has been termed total uncertainty.

The results of analyzing total uncertainty for an unreplicated event can be termed credibilities, following Russell (1948), to indicate that they are not probabilities in a conventional sense. An example is weather reports that present the credibility of rain tomorrow, the unreplicated event, based on uncertainty in measurements and models and the stochasticity of the atmosphere. If we are interested only in the credibility of an outcome, we need not distinguish variability and uncertainty. For example, if we wish to estimate the credibility that a drought will extirpate a reintroduced population of fish in a stream in the next 50 years, we might estimate and propagate all sources of variability and uncertainty together. For Bayesians, this concept of credibility is simply a case of degree of belief (Section 5.3.2).

5.1.5 Error

Error can be introduced at any stage in the generation and analysis of data including sampling, measurement, data transcription and analysis, model selection, and presentation of results. Generally, assessors deal with error by attempting to minimize it (Chapter 9) and hoping that it is inconsequential. Error, particularly measurement error, may, however, be large enough to significantly contribute to unpredictability (Sarda and Burton 1995). If measurement error is a particular concern or if it is important to estimate the true variability underlying a data set, statistical techniques are available to estimate the error and variability components of a data set (Zheng and Frey 2005). Alternatively, error may be estimated independently by quality assurance audits or equivalent studies. For example, from studies of interlaboratory variability in aquatic toxicity tests, error is generally in the range of

$3\times$ to $5\times$ (i.e., the highest result in a set of replicate tests is three to five times as great as the lowest) but may be $>10\times$ depending on the test, toxicant, and set of laboratories (Johnson et al. 2005). To the extent that we are aware of, and estimate, error, it is a component of uncertainty.

The importance of error in science is illustrated by the experience of physicists, who often find that data that differ from the null model by three to six standard deviations are illusory, apparently because of errors that bias the results (Seife 2000). It seems likely that such erroneous results are common in environmental studies, given the routine use of a two sigma standard (95% confidence), the many opportunities for error, and the absence of independent replication of environmental studies.

5.1.6 Ignorance and Confusion

Ignorance and confusion make the results of an assessment uncertain in ways that the assessors do not recognize. Ignorance is the source of unknown unknowns. That is, in some cases one is uncertain about some aspect of the problem without knowing it, because of ignorance that the issue even exists. For example, due to ignorance, some assessments of aquatic risks from low dissolved oxygen (DO) have been based on measurements of daytime DO, which is elevated by photosynthesis, rather than nighttime DO, which is depressed by respiration.

Confusion is the result of excess complexity of a problem. Often it results from attempts to mentally model a problem that requires a formal conceptual, mathematical, or logical model.

5.1.7 Summary of Sources

We can categorize input or output parameters in terms of the types of unpredictability that characterize them:

None—we know the answer and it is constant.
Variability—we know the answer but it varies.
Uncertainty—we are unsure of the answer.
Total uncertainty—we are unsure of the answer, but we know it varies.
Error, confusion, or ignorance—we do not know the answer at all.

5.2 WHAT IS PROBABILITY?

Since the founding of the probability theory in around 1660, it has been applied to multiple concepts (Hacking 1975). It is best to think of probability as a unit that scales 0–1 and follows certain logical rules. For example, if A and B are independent random events, the probability of observing both A and B (the joint probability) is $p(A \ \& \ B) = p(A) \times p(B)$. That relationship holds no matter how the constituent probabilities are defined or derived.

Note that probability is not the only possible unit of risk. Alternatives include qualitative scales such as acceptable/unacceptable or highly likely/likely/as likely as not/unlikely/highly unlikely. They also include frequencies and relative frequencies, which are equivalent to some probabilities, but more easily understood (Gigerenzer 2002). However, probability is the standard unit of risk.

Like any unit, probability can be applied to diverse concepts and situations and can be estimated in various ways. While there are other concepts of probability—Good (1983) describes six—the most common are probabilities as expressions of frequencies and as expressions of beliefs. Since probability is used as the measure of uncertainty in predictions, the sources of probability correspond to the sources of unpredictability (Section 5.1). The correspondence is, however, rough and imperfect.

5.2.1 Types of Probability: Frequency vs. Belief

5.2.1.1 Frequency

Probability is intuitively understood as an expression of frequency. That is, we can summarize the results of 100 coin flips as the frequency of heads (54 out of 100) and infer that the probability of heads with this coin is 0.54. The frequency becomes a probability when it is normalized as 0–1 and when it is assumed that the frequency is an estimate of an underlying property, the probability of heads. This classical view of probability and statistics is easily accepted, because it reflects the human propensity to extrapolate from experience (a series of coin flips) to a general rule (the probability of heads), which may be applied to individual instances (the next coin flip). However, this approach runs into problems when we try to apply this experience to other coins or other coin flippers. Experimental science can be thought of as a means of making real-world variability analogous to flipping a coin. Experimental design involves creating replicate systems and randomly assigning treatments to them, so that the frequency of responses can be extrapolated to probabilities of response of such systems in the real world. Problems occur when attempting to assign experimentally derived probabilities to individual events in the complex and uncontrolled real world. However, even without experimental replication, some probabilistic parameters are routinely derived from frequencies. Examples include frequencies of failures of treatment plants and of parameters resulting from meteorology such as frequencies of wind direction or of low stream flows.

Probabilities are derived from frequencies, because of the associated variability in the system that results in uncertain outcomes. If there were only one possible outcome, we would not bother to express its probability as 1. To the extent that particular types of errors can be identified and expressed as discrete variables, they can be used to estimate probabilities. If, for example, quality assurance audits provide the frequency of an error, that frequency can be used to estimate the probability that a value in a data set is erroneous. In that case, we are treating variability in the data generation process as equivalent to variation in nature. Ignorance and confusion are effectively immune to the frequentist version of probability.

5.2.1.2 Belief

Clearly, probability is applied to situations in which a frequency is not its source. For example, I might say that drought probably reduced the abundance of San Joaquin kit foxes on the Elk Hills Petroleum Reserve in the 1980s, and I might even state that the probability (or more properly the credibility) is 0.8. That is not a repeated event, so I am expressing my degree of belief rather than appealing to a frequency. In addition, even when we can appeal to frequencies our statements of probability may be better thought of as degrees of belief rather than as instances from a set of repeated observations. For example, if we speak of the probability of an effect in a stream due to a proposed domestic wastewater effluent, what frequency is relevant? If we require similar streams in the same region and very similar effluents, we are unlikely to find enough cases to estimate a frequency. However, if we accept all streams and domestic wastewater effluents, the relevance to our particular case is doubtful because of the range of stream and effluent flow rates, stream ecologies, effluent contents, technological reliability, etc. None of the frequencies provides exactly the probability that we want. This is a serious conceptual challenge for frequentist statisticians, but not for subjectivists who equate all probabilities with degrees of belief. However, it is a serious practical problem for assessment of risks for a single event, no matter what statistical framework is used.

Beliefs should correspond to frequencies in the long term. If a risk assessor says that the risk that a particular effluent will cause local extinction of brook trout is 0.7, that prediction

cannot be verified, because the extinction will or will not occur. There is no condition of 70% extinct. However, over a career, predicted effects should occur 70% of the time that a risk estimate of 0.7 was derived. That verification occurs for weather reports. Over the millions of weather predictions that have been made by the US National Weather Service, it rains on 70% of days at locations for which a 70% chance of rain is forecast. However, that sort of verification is not possible for ecological risk assessors, because there are too few predictions, methods and circumstances of prediction are too diverse, and predictions of significant risk result in management actions that negate the prediction.

Finally, subjectivists argue that even when observed frequencies are converted to probabilities, the conversion is based on belief rather than logic. We have no reason for certainty that the next ten coin tosses will be like the previous ones. However, that is the way to bet.

The concept that probabilities are expressions of belief is clearly more flexible than the frequentist concept. That is, beliefs concerning variability, uncertainty, error, ignorance, and confusion can all be expressed in units of probability.

5.2.2 TYPES OF PROBABILITY: CATEGORICAL VS. CONDITIONAL

Traditionally, statisticians have been concerned with estimating the probability of some event or hypothesis. An example is the probability of a fish kill below a treatment plant. This is a categorical probability $p(y)$. However, more commonly in recent years, the probability of some event given a prior event, or of some hypothesis given data, is of concern. An example is the probability of a fish kill given that a treatment plant has failed. This is a conditional probability $p(y|x)$. The basic formula of conditional probability is

$$p(y|x) = p(x \,\&\, y)/p(x) \tag{5.1}$$

where $p(x \,\&\, y)$ is the joint probability of x and y.

This distinction may not be obvious. The reader may say that the first example is conditioned on the existence of the treatment plant. However, if we treated that example as conditional, it would refer to the probability of a kill in a stream reach, given the probability that it has a treatment plant at its head. We would not be interested in that conditional probability, because we defined the assessment problem and delineated the reach based on the known existence of the plant. That is, $x = 1$, so $p(y|x)$ reduces to $p(y)$. The second example is more appropriately conditional, because we are interested in the probability of treatment failure, which is not constrained, and its influence on the probability of a fish kill.

Conditional probability is associated with Bayesian statistics, because Bayes' rule is a formula for estimating probabilities given prior beliefs and evidence. However, Bayesian statistics also involve concepts of subjectivism and updating that are not inherent in the concepts and calculus of conditional probabilities (Section 5.3.2).

5.3 WAYS TO ANALYZE PROBABILITIES

Some probabilities are simply calculated from the logical rules of probability. For example, if the probability that an out-migrating salmon will be killed by hydroelectric dams is 0.2 and by irrigation withdrawals is 0.08, the probability of being killed by these anthropogenic hazards is $(0.2 + 0.08) - (0.2 \times 0.08) = 0.26$, i.e., the probability of being killed by either minus the probability of being killed by both, because a fish cannot be killed twice. It is equivalent to the response addition model in mixtures toxicology (Section 8.1.2). More commonly, probabilities are estimated using one of the various forms of statistics.

5.3.1 FREQUENTIST STATISTICS

The dominant school of statistics termed frequentist is concerned with estimating the frequency of defined errors given an assumed model and an experimental or sampling design. This school of statistics has been primarily concerned with testing statistical hypotheses. In this context, "probability arises not to assign degrees of belief or confirmation to hypotheses but rather to characterize the experimental testing process itself: to express how *frequently* it is capable of discriminating between alternative hypotheses and how *reliably* it facilitates the detection of error" (Mayo 2004). Hence, it tells us about the probability of errors in our procedure and not, as many users believe, about the support for hypotheses provided by evidence. The latter is supplied by "evidential relation logics (whether Bayesian, likelihoodist, hypothetico-deductive or other)" (Mayo 2004). This aspect of frequentist statistics is the reason that your consulting statistician wants to be involved in designing your experiments. The hypothesis test is using the data to test the experimental design under the assumption that the null hypothesis is true.

Frequentist statistics actually belong to two schools: Fisher's testing of a null hypothesis using p values and Neyman–Pearson statistics, which estimates relative error rates, usually for a null hypothesis and an alternative. The standard approach to hypothesis testing in the natural sciences is a hybrid, which is interpreted in a way that is not strictly correct but which seems to work well enough for analysis of experiments (Gigerenzer et al. 1989; Hacking 2001; Salsburg 2001). In any case, statistical hypothesis testing is, in general, not appropriate to risk assessment (Box 5.1). Frequentist statistics are also applied to more useful analyses such as confidence intervals and statistical modeling techniques such as regression analysis (Section 5.5).

BOX 5.1
Hypothesis Testing Statistics

The statistical testing of null hypotheses is commonly perceived to be an objective and rigorous approach to scientific inference. Such tests are frequently used to inform environmental management, but these uses are nearly always inappropriate. Many publications have criticized the ascendancy of statistical hypothesis testing in applied science (Parkhurst 1985, 1990; Laskowski 1995; Stewart-Oaten 1995; Suter 1996a; Johnson 1999; Germano 1999; Anderson et al. 2000; Roosenburg 2000; Bailar 2005; Richter and Laster 2005). In fairness, it must be noted that some environmental statisticians still defend statistical hypothesis testing (Underwood 2000). The following are brief descriptions of a few of the problems with hypothesis testing.

We are not testing hypotheses: Hypothesis testing statistics were developed to test the reality of a hypothesized phenomenon by attempting to refute the null hypothesis that the phenomenon does not exist. "Most formal inference is not hypothesis testing but model construction, selection and checking (formal and informal), estimation of parameters and standard errors, or calculation of confidence regions or of Bayesian posterior distributions of parameters" (Stewart-Oaten 1995). This is particularly true in ecological risk assessment. We are not interested in testing the hypothesis that a chemical has no toxicity; we want to know what its effects will be at given levels of exposure. Similarly, we are not interested in testing the null hypothesis that two streams are identical; we know they are not. We want to know how they differ. Even when researching scientific hypotheses, it is better to compare genuine alternative hypotheses than to test a null hypothesis that nobody believes (Anderson et al. 2000; Taper and Lele 2004).

We are not interested in statistical significance: We are all taught, but often forget in practice, that statistical significance has no particular relationship to biological significance or societal significance. Real-world decisions should be based on real-world significance.

Continued

BOX 5.1 (Continued)
Hypothesis Testing Statistics

Statistical significance is subject to manipulation: If a responsible party wishes to avoid statistically significant effects in a test or field study, they can use few replicates, perform many comparisons, and use imprecise techniques so that the null will not be rejected. On the other hand, if environmentalists wish to find an effect, any difference can be statistically significant if the investigators are careful to keep method errors small and are willing to take enough samples. This is a result of the fact that statistical significance is about statistical designs, particularly the number of "replicates."

The relative degree of protection is biased by the statistics: If statistical significance is the criterion, more precisely quantified, more readily sampled, and more easily replicated life stages, species, and communities are more protected. For example, in biological surveys, a given proportional loss of species is more likely to be statistically significant in invertebrates than in fish, so biocriteria are less protective of fish (Suter 1996c). Some important ecological effects may never be detectible with statistical significance in practical or ethically acceptable studies (Roosenburg 2000).

Statistical hypothesis tests with field data are nearly always pseudoreplicated: The multiple samples of contaminated media and biota that are taken from an exposed environment and used as replicates in hypothesis tests are in fact pseudoreplicates (Hurlbert 1984). The message that environmental pseudoreplicates should not be treated as replicates has been widely heard but has not been appreciated. Few risk assessors would accept the results of a toxicity test in which one rat was treated with a chemical and another was not, even if replicate samples of blood from one rat had a "significantly" different hematocrit from the other. The reason that the test would not be accepted is that while the measurements were replicated the treatment was not. There are any number of reasons why the hematocrit of one animal might be lower than another's, which has nothing to do with the treatment. In studies of contaminated sites, the site is the "rat" and the waste, effluent, or spill is the unreplicated treatment. Ecological risk assessors should be at least as leery of testing hypotheses about differences between differently treated sites as they are about differently treated animals. There are more reasons why two ecosystems might differ than why two rats might differ. Note that pseudoreplication affects all estimates of variance, not just those in hypothesis tests. However, the issue is much more important for hypothesis tests because the variance is used to decide whether a treatment has an effect.

Statistical hypothesis tests with field data almost never randomly assign treatments: Even if we have true replication (e.g., by comparing sets of replicate streams with and without sewage treatment plants), field studies almost never randomly assign receptors to treatments. That is, we do not have a set of 20 streams and randomly pick 10 to have a sewage treatment plant each and 10 to have none. In fact, streams receiving treated sewage were picked by engineers and planners for nonrandom reasons and are likely to differ in ways that nullify any hypothesis test to determine the significance of differences.

Hypothesis tests invite misinterpretation: Even when testing a scientific hypothesis with an appropriate experimental design, the statistics are often misinterpreted. A common mistake is to accept a null hypothesis when one fails to reject it (Parkhurst 1985; Anderson et al. 2000). Null hypothesis tests assume that the null is correct and determine the probability of data at least as deviant from the null as the data obtained, given that assumption. Clearly a test cannot address its basic assumption. For example, in developing a model of the aquatic toxicity of polycyclic aromatic hydrocarbons (PAHs), it was assumed that the slopes of regressions of $\log LC_{50}$ vs. $\log K_{ow}$ are equal for all species (DiToro et al. 2000). The investigators tested the null hypothesis of equal slope for 33 species and, after correcting for multiple comparisons, accepted the null hypothesis and concluded that the slopes are all equal to a universal slope of -0.97. (The correction for $n = 33$ makes it very difficult to achieve a significant difference, but that is a less fundamental problem.) The conclusion that the slopes are effectively equal is defensible, but is not justified by the hypothesis test. To provide evidence for equality of slopes within a hypothesis testing approach, the investigators should have determined the deviation δ from the universal slope that would be biologically significant and then designed a study to determine whether that difference occurs. Those who recognize that they should not accept their null based on failure to reject are often tempted to perform a post hoc analysis of the power of the test. Such retrospective power analyses are, however, technically indefensible (Goodman and Berlin 1994; Gerard et al. 1998; Shaver 1999).

5.3.2 BAYESIAN STATISTICS

Although modern Bayesian statistical practices, like frequentist statistics, are concerned with making inductive inferences, Bayes himself was concerned with a problem in deductive inference. Classical deduction is represented by syllogisms such as A is greater than B and B is greater than C, therefore A is greater than C. This logic is beyond dispute until we apply it to the real world. Then we must admit that the magnitudes of A, B, and C are estimated with some degree of confidence. Hence, the inference is based on some prior degree of belief in the relative magnitudes, and even deduction must be admitted to be as subject to bias and error as induction. This position, that all inference is subject to uncertainty, is termed subjectivism.

Modern Bayesianism is characterized by three concepts, none of which is adopted by all Bayesians.

Subjectivism: Bayesian statisticians argue that we know nothing about underlying distributions, so we can say nothing about error rates in estimating true frequencies. All we know is the defensibility of beliefs or of changes in beliefs given evidence. Although there are personalistic Bayesians, subjective does not necessarily mean personal. Most Bayesians believe that their analyses provide a rational basis for interpersonal subjective beliefs.

Updating: Bayesian statistics provide an appropriate way to go about modifying our beliefs given new evidence. The presumption is that we have good reasons for believing what we believe (i.e., we are rational and our reasoning to date has been coherent). Bayes' theorem tells us how to go about that updating (we multiply by the likelihood of the new evidence given the hypothesis). Non-Bayesians respond that it often makes no sense in practice to update. For example, if we have an empirical model of plant uptake of soil contaminants and we have new data on plant uptake from the site that we are assessing, should we use the data directly or should we use it to update estimates from the model? Most assessors would favor using the data directly, because plant uptake is so dependent on site characteristics. As Dennis (2004) wrote concerning the conditioning of data analyses on prior beliefs: "Why dilute the Rothschild with Boone's Farm?" Bayesians can avoid this problem by using uninformative priors, and some Bayesians insist on uninformative priors. However, a nominally uninformative prior such as a uniform distribution is still an assumption that influences a Bayesian outcome. A more defensible but less common approach is to combine data from prior instances of the phenomenon being estimated as the prior (e.g., waste tanks of the same type or closely related species) (Myers et al. 2001; Goodman 2005).

Conditional probabilities: As discussed earlier, Bayesian statistics are associated with the calculation of conditional probabilities. Those who use Bayesian analysis of conditional probabilities, but do not subscribe to subjectivism, are called logical Bayesians. Bayes' rule appears in various versions of including versions for multiple conditions and for continuous variables. The basic version of Bayes' rule is the following formula:

$$p(B|A) = [p(A|B)p(B)]/p(A) \tag{5.2}$$

where $p(B|A)$ is the probability of B given A. B can be a parameter value, a state of a system or anything else that has a probability, but often it is described as a hypothesis. Hence, the formula may be recast as

$$p(\text{hypothesis}|\text{evidence}) = [p(\text{evidence}|\text{hypothesis})p(\text{hypothesis})]/p(\text{evidence}) \tag{5.3}$$

where $p(\text{evidence}|\text{hypothesis})$ is the likelihood function and $p(\text{hypothesis})$ is the prior.

Bayesian analysis of conditional probabilities has become a favorite tool for addressing the controversial question, does *Pfiesteria piscidae* cause fish kills (Stow 1999; Brownie et al. 2002; Newman and Evans 2002; Stow and Borsuk 2003). Newman and Evans's (2002) version is

$$p(\text{fish kill}|Pfiesteria) = [p(Pfiesteria|\text{fish kill})p(\text{fish kill})]/p(Pfiesteria) \tag{5.4}$$

As is so often the case with ecological assessments, problems come in parameterizing the formula. In this case,

> p(fish kill) is taken to be the daily rate of fish kills in the Neuse and Pamlico Rivers $= 0.081$.
> p(*Pfiesteria*) is from the frequency of *Pfiesteria* in samples from east coast sites $= 0.205$.
> p(*Pfiesteria*|fish kill) is from the frequency of *Pfiesteria*-like organisms in water samples at fish kills in the Neuse and Pamlico Rivers $= 0.52$.

Hence, by Bayes' rule,

$$p(\text{fish kill}|Pfiesteria) = 0.205 \tag{5.5}$$

Note that in this analysis, the prior, p(fish kill), is not really a prior belief. It is simply the number of fish kills divided by the number of observation days. A different answer would have been obtained if the authors had used their actual prior beliefs concerning the probability of a fish kill given an episode of *Pfiesteria* occurrence. Further, the use of a daily rate as the prior does not provide the desired probability per episode. These comments are not intended to discount the analysis, but rather to highlight the conceptual difficulties involved in just one simple term of the formula. Other controversies are associated with the other two terms.

Bayesian statistics provide an alternative to frequentist statistics. Bayesian analogs of mainstream frequentist techniques are available including Bayesian hypothesis testing, Bayesian confidence intervals, etc. Bayesians are common in academic statistics departments, and their methods are increasingly being applied in industry and government agencies for a variety of purposes (Box 5.2).

5.3.3   RESAMPLING STATISTICS

The availability of computers has made it possible to estimate probabilities by repeatedly sampling from a data set or distribution, and analyzing the results of those samples. If one wishes to estimate the distribution of a variable from a sample (e.g., the distribution of fish weights from a sample of five fish), one might estimate a distribution by assuming a distribution function (e.g., log normal) and using conventional frequentist statistics. However, one might doubt that we know the appropriate function and may doubt that five fish are

---

BOX 5.2
Bayesian Statistics and Sampling Design

Conventional approaches to sampling design begin with a pilot study that allows one to estimate the distribution of the variable of interest. Then, if acceptable Type I and Type II error rates are provided by the decision maker, an adequate number and distribution of samples can be estimated. This approach is the basis for the quantitative portion of the US EPA Data Quality Objectives process (Quality Assurance Management Staff 1994). In the author's experience, this is impractical, because decision makers will not admit to having acceptable error rates. This problem is avoided with Bayesian analysis, because calculations are based only on the data obtained and not on possible values given the statistical model. More importantly, Bayesian updating lends itself to efficient sampling. One simply keeps collecting data until probabilities are adequately defined or time or money run out. This approach has revolutionized clinical trials Berry et al. 2002. One keeps adding patients to the trial and updating the probability estimates until it is sufficiently clear that the drug works, causes unacceptable side effects, or does neither. This approach gets results more quickly and maximizes benefits to participants.

enough to define it. Remarkably, we can generate an estimate of the underlying distribution by repeatedly sampling from the set of five weights, with replacement, and then determining the distribution of the mean, standard deviation, or other sample statistics from that set of samples. This is the nonparametric Bootstrap method (Efron and Tibshirani 1993).

Resampling can also help us estimate probabilities from models. If we have a model with multiple parameters that are defined by distributions, we may be able to solve the model analytically by variance propagation (Morgan and Henrion 1990). However, such analytical solutions are often impractical because of nonlinearities and correlations among parameters. The solution is to use Monte Carlo simulation, which involves repeatedly sampling from each parameter distribution, solving the model, saving the results, and then reporting the distribution of those results as the distribution of the modeled variable (Section 5.5.5).

5.3.4 OTHER APPROACHES

The number of ways of analyzing and expressing variability and uncertainty is large and growing. In addition to frequentist, Bayesian, and resampling statistics, they include interval arithmetic, fuzzy arithmetic, p bounds, fault-tree analysis, and possibility theory. The lack of discussion of these techniques should not be taken to imply rejection of them. It reflects a judgment that current conventional methods are appropriate and much more likely to be accepted by reviewers and decision makers than exotic methods.

5.4 WHY USE PROBABILISTIC ANALYSES?

The first step in a probabilistic analysis must be to determine motivation. The form and content of the analysis depend on the desired output of the analysis. However, most guides to uncertainty analysis assume a particular motivation and desired output and proceed from that assumption. Reasons include the following.

5.4.1 DESIRE TO ENSURE SAFETY

Because of variability, a realized effect may be considerably larger than the most frequent effect. Because of uncertainty, true effects may be larger than estimated effects or may occur more frequently. Therefore, if the goal of an assessment is to ensure that all credible hazards are eliminated or at least accounted for in the decision, variability and uncertainty must be incorporated into the analysis. This may be done in at least four ways:

1. One may determine that the uncertainties are so large and poorly specified that no quantitative uncertainty analysis is possible. In such a case, a risk management decision may be made that all members of an allegedly hazardous class of chemicals or other hazardous entities should simply be banned. This is known as the precautionary principle. Once the risk management decision is framed in that way, the output of the risk analysis is a conclusion that a chemical or technology belongs or does not belong to a banned category.
2. One may make conservative assumptions. For example, in human health risk assessments it is assumed that an individual drinks 2 L of water a day from a contaminated source for a lifetime, consumes fish caught in contaminated waters, consumes vegetables grown on contaminated soil irrigated with contaminated water, etc. Following this example, ecological risk assessors may assume that an entire population of a wildlife species occupies the most contaminated portion of a site. By hypothesizing levels of exposure higher than are credible for any real human or wildlife population, these

conservative assumptions ensure that exposure is not underestimated, even though the exposure is uncertain. The product of stringing conservative assumptions together is a "worst case" or "reasonable worst case" estimate of risk.

3. One may apply safety factors to the components or results of the assessment. These are factors (usually 10, 100, or 1000) that are applied to ensure an adequate margin of safety (Section 5.5.1). They are based on expert judgment and simple analyses of past cases. The output of analysis using safety factors is a conservative risk estimate. However, because of the way factors are derived and the way in which they combine multiple sources of uncertainty, the degree of conservatism that results from safety factors is unclear.

4. One may perform a formal quantitative uncertainty analysis and choose as an endpoint a probability of effects that is very low. For example, one might declare that the probability must be less than 0.01 that the likelihood of extinction is as high as 0.0001 over the next 50 years.

5.4.2   Desire to Avoid Excessive Conservatism

As already discussed, the desire to ensure safety has led to the use of numerous conservative assumptions and safety factors in risk assessments. Some risk assessors, regulated parties, and stakeholders have objected that the resulting margins of safety are excessive (Kangas 1996). One response has been to argue for reduction of number and magnitude of factors and conservative assumptions or their elimination (i.e., use best estimates). An alternative is to develop anticonservative factors to correct, at least in part, the compounding of conservatism (Cogliano 1997). Another approach is to replace uncertainty factors and conservative assumptions with estimated distributions of parameters and to replace the compounding of factors with Monte Carlo simulation (Office of Environmental Policy and Assistance 1996). If low percentiles of the distributions of risk estimates are used to ensure safety, this approach is not necessarily less conservative than traditional regulatory approaches.

5.4.3   Desire to Acknowledge and Present Uncertainty

It is generally considered desirable to acknowledge and estimate the uncertainties associated with assessments. It is both safer to admit your uncertainties and more ethical than ignoring or hiding them. This is more the case with ecological risk assessments than with human health risk assessments, because estimated ecological effects are often detectable, and therefore a conservative deterministic estimate may be refuted by subsequent observations. A formal probabilistic analysis provides a clear and defensible method for estimating variability and uncertainty and justifying the estimates. However, many uncertainties are not estimated by conventional uncertainty analysis, such as the uncertainty associated with model selection or the uncertainty concerning assumptions about the future use of a site. Hence, presentations of uncertainty must include lists of issues, and qualitative judgments, as well as quantitative estimates.

5.4.4   Need to Estimate a Probabilistic Endpoint

Probably the least common reason for analysis of uncertainty in ecological risk assessment is the specification of a probabilistic endpoint by the risk manager. Probabilistic endpoints have been estimated by ecological risk assessors since the founding of the field (Barnthouse et al. 1982), but the impetus has come primarily from the assessors, not the risk managers. A conspicuous exception is population viability analysis, which estimates probabilities of extinction of species or populations given prescribed management practices (Marcot and

Holthausen 1987). Such analyses should be done in the event that a proposed action may pose a threat of extirpation. A more likely impetus for probabilistic endpoints are cases in which the ecological risks are driven by the probability of occurrence of an extreme event. Examples include the failure of a dam that holds a waste lagoon, an extremely wet period that brings contaminated groundwater to the surface, or the entry of a large flock of horned larks into a field recently treated with granular carbofuran. Finally, demands for probabilistic analysis may come from reviewers or from responsible parties. In any case, ecological assessment endpoints can be expressed as a probability, given variability or uncertainty in exposure or effects.

5.4.5 PLANNING SAMPLING AND TESTING

Ideally, the field and laboratory investigations that provide the data for risk assessments should be prioritized and planned on the basis of an analysis of uncertainty. The goal of the quantitative data quality objectives (DQO) process is to gather enough data to reduce uncertainty in the risk estimate to a prescribed, acceptable level (Chapter 9). This formalism is not directly applicable to ecological risk assessment, but one can still allocate resources on the basis of expected reduction in uncertainty. This use of uncertainty analysis requires a separate analysis of uncertainty that can be reduced by feasible sampling, analysis, or testing, rather than total uncertainty. For example, a model of mink and heron exposure to polychlorinated biphenyls (PCBs) and mercury was used in the Clinch River assessment to determine that the greatest source of reducible uncertainty in the exposure estimates was the distribution of PCB concentrations in water and sediment (MacIntosh et al. 1994). Analyses for this purpose are termed sensitivity analyses (Section 5.5.7).

5.4.6 COMPARING HYPOTHESES AND ASSOCIATED MODELS

Although conventional hypothesis testing has little place in ecological risk assessment (Box 5.1), statistics are needed to choose among alternative models, based on different hypotheses concerning the nature of the system, when mechanistic evidence is unclear. A simple common case is the selection of an exposure–response function to model a set of test data from among alternative functions that imply different characteristics of the response (linearity, thresholds, hormesis, etc.). Other cases involve selecting from among more complex alternative hypotheses such as compensatory, depensatory, and proportional responses of populations to mortality (Chapter 27) or bottom–up vs. top–down control of ecosystems (Chapter 28). Three approaches to model comparison are available:

1. One may choose the model that is best supported by the evidence. The most direct approach is to use likelihood ratios to compare the relative likelihoods of hypotheses given the available data (Royall 1997, 2004). This approach may be extended to include the goal of model parsimony (i.e., Occam's razor) by applying Akaike's information criterion. It is effectively the relative likelihood, normalized by the number of parameters in the model (Akaike 1973).
2. One may use knowledge of underlying mechanisms rather than the evidence provided by a data set to choose among hypotheses. This approach is particularly appealing when, as is often the case, data are not sufficient to confidently distinguish among models. For example, it is generally impossible to detect depensatory processes in population-monitoring data. However, one can estimate the influence of density-dependent predation and of difficulty in finding mates (the two mechanisms of depensatory responses) given the abundance and selectivity of predators and the abundance of mates in the ecosystem being assessed.

3. One may apply all plausible models. One can then present the range of results to the risk manager and stakeholders as an expression of model uncertainty (Section 5.1.2). Alternatively, one can combine them using Bayesian model averaging (Hoeting et al. 1999; Wasserman 2000). This is conceptually appealing, if information concerning underlying mechanisms is used to assign prior probabilities.

5.4.7 AIDING DECISION MAKING

Finally, the results of an uncertainty analysis may aid the risk manager in making a decision concerning the remedial or regulatory action. Decision analysis and some other decision-support tools require estimates of the probability of various outcomes, which must be derived by a probabilistic analysis of risks (Chapter 34). More generally, the additional information provided by an uncertainty analysis may lead to a better informed and more defensible decision, even without quantitative decision analysis.

5.4.8 SUMMARY OF REASONS

These reasons for evaluating variability and uncertainty are not mutually exclusive, so an assessor may have multiple motives. However, the chosen analytical method must be able to satisfy the most restrictive reason. For example, if one wishes to ensure safety, any analysis will do; but, if one wishes to ensure safety and present a full disclosure of uncertainties in the assessment, only a quantitative analysis will serve; and if one is using uncertainty analysis to help plan a program of sampling, testing, and analysis, only a quantitative analysis that distinguishes sources of uncertainty and variability will serve.

5.5 TECHNIQUES FOR ANALYSIS OF VARIABILITY AND UNCERTAINTY

This section presents six important classes of methods that are commonly used in ecological risk assessment.

5.5.1 UNCERTAINTY FACTORS

The most common technique for incorporation of uncertainty is uncertainty factors (also referred to as safety factors). These are numbers that are applied to either parameters of a risk model or the output of a model to ensure that risks are not underestimated. Most factors are based on expert judgment, informed by experience and simple analyses (Dourson and Stara 1983). For example, the NOAEL values used to calculate wildlife toxicity benchmarks are divided by a factor of 10 if they are based on subchronic studies, because of uncertainties concerning subchronic endpoints as estimators of chronic toxicity (EPA 1993e; Sample et al. 1996c). This factor is based on expert judgment that the threshold for chronic toxicity is unlikely to be more than a factor of 10 lower than a subchronic NOAEL. Most other uncertainty factors are also multiples of 10, reflecting their imprecision.

In addition to the informality of their derivation, the chief complaint against uncertainty factors is the way they propagate uncertainty through a model. If a model contains four parameters, which are multiplied, and each has an associated uncertainty factor of 10, the total uncertainty is a factor of 10,000. This implies that in the case being analyzed, all things are simultaneously, individually as bad as they can credibly be. The uptake factor is much higher than has been observed, the organisms have extremely small foraging ranges, the endpoint species is much more sensitive than the test species, etc. To avoid obtaining absurdly extreme estimates when using this method of uncertainty analysis, one should estimate a

maximum credible uncertainty (i.e., an overall uncertainty factor) in addition to factors for the individual parameters.

Uncertainty factors are operationally equivalent to the extrapolation factors discussed in Chapter 26. The distinction is simply that extrapolation factors account for identified systematic differences between measures of effect and assessment endpoints, while uncertainty factors account for uncertainties when systematic differences are not identifiable. For example, the subchronic tests discussed above are designed to be equivalent to chronic tests, so we do not expect them to be different, but we are uncertain of the truth of that assumption in any particular case. Therefore, an uncertainty factor is employed. If we knew that there was a predictable difference between subchronic and chronic test results, we might develop an extrapolation factor.

5.5.2 CONFIDENCE INTERVALS

Confidence intervals and their bounds are the most generally useful statistics for expressing variability or uncertainty. Although confidence intervals are often treated as equivalent to hypothesis tests, confidence and significance are distinct concepts within frequentist statistics (Hacking 2001). Unlike significance tests, confidence intervals allow us to estimate properties of the population from which a sample was drawn. Hence, their use reduces the sterility of performing hypothesis tests and reporting significance or lack thereof. Much more useful information is provided by presenting confidence intervals as interval estimates of a parameter or as a "best" (mean, median, etc.) estimate and confidence intervals on that estimate or on the data themselves. Confidence intervals on the data, given a model, are called prediction intervals (see, e.g., Figure 26.3). Information is increased by presenting multiple confidence intervals. That is, rather than presenting only the conventional 95% confidence interval, which has no particular relevance to environmental management, the 50%, 75%, 90%, and 95% intervals are presented (Figure 5.1).

Frequentist confidence intervals are, strictly speaking, statements about confidence in a method of sampling and its associated model (i.e., the method of estimating the interval is correct 95% of the time), not about the data (i.e., you cannot say that the true value falls
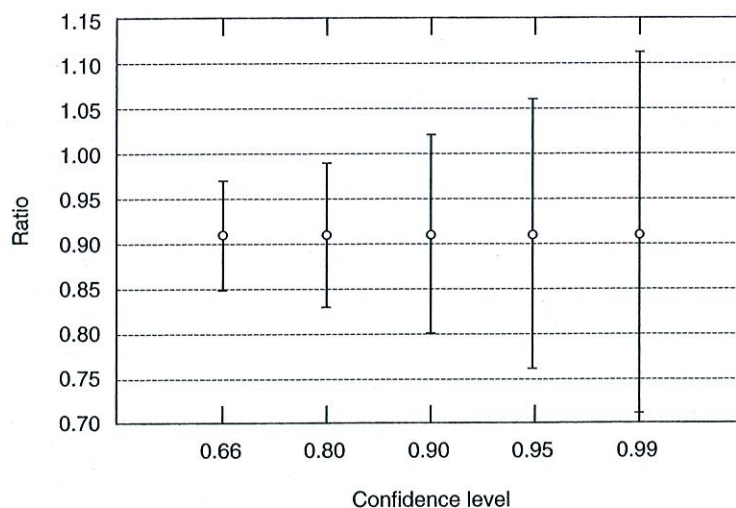

FIGURE 5.1 Sets of confidence intervals used to show how the uncertainty in a value (e.g., the ratio of exposure concentration to toxic concentration) increases as the required level of confidence increases.

in the interval with 95% confidence). To estimate confidence in sample-derived estimates themselves, you need Bayesian confidence intervals.

5.5.3 DATA DISTRIBUTIONS

Uncertainty often appears in the form of a distribution of realized values of a parameter. For example, concentrations of chemicals in repeated water samples or concentrations of chemicals at which individual organisms respond in a toxicity test have some distribution on the scalar, chemical concentration. These distributions can be used to estimate uncertainties concerning the mean water concentration during the period sampled, a future distribution of water concentrations, the concentration that will result in a fish kill, etc. As will be discussed, it is important to carefully consider the relationship between any data distribution and the distributions to be estimated in the assessment. In general, distributions can serve two functions. First, they can be used to represent the uncertainty or variability of a parameter in a mathematical model of exposure or effects. Second, when exposure or effect metrics are directly measured, the distribution of the measurements may directly represent the uncertainty or variability of exposure or effects.

An important decision to be made is whether to fit a function to a data distribution and, if so, which. Conventionally, one describes a distribution by fitting a mathematical function to it, such as the normal, lognormal, uniform, or logistic. The result is referred to as a parametric distribution function. However, if the distribution is not well fit by a parametric function or if the purpose is to display the actual distribution of the data, an empirical distribution function (EDF) may be used. In software for Monte Carlo analysis, EDFs are typically referred to as custom functions. One limitation of EDFs is that they do not describe the distribution beyond the data. This is a problem if the data set is not large and extreme values are of concern (e.g., estimating infrequent high exposure levels or responses of sensitive species). However, parametric functions may also poorly represent extreme values if they have infinite tails, if the extremes of the data are not symmetrical (e.g., the resistant organisms are more highly resistant than the sensitive organisms are sensitive), and if, as is typically the case, their fit is influenced primarily by the bulk of data near the centroid. Issues related to choosing and fitting functions are discussed in a book (Cullen and Frey 1999), a workshop report (Risk Assessment Forum 1999), and in two special issues of the journal *Risk Analysis* (vol. 14, no. 5; and vol. 19, no. 1).

Published distribution functions are available for some variables, either in prior risk assessments or open literature publications. For example, distributions of great blue heron exposure parameters are found in Henning et al. (1999), and such distributions may be expected in the future for other species. The following strategies may be employed to develop distribution functions:

- If the data set is large, one can employ statistical software packages to use statistical criteria to select the best function. In general, the best model provides the highest likelihood of the data, but other criteria such as least sum of squared deviations are also used and all should give the same ranking with good data. However, if there are few data points or the data are noisy, fitting algorithms may fail to give appropriate results. In addition, one must be aware of the fact that models with more parameters fit data better than those with fewer. Hence, one would not choose a three-parameter logistic model over a two-parameter logistic model unless the fit was significantly better, or the additional parameter had some mechanistic significance (e.g., to fit a variable maximum value). Akaike's information criterion, which is the log likelihood normalized by the number of parameters, provides an appropriate basis for comparing functions with

different numbers of parameters fit to the same data set (Burnham and Anderson 1998) as it does for comparing mechanistic hypotheses (Section 5.4.6).

- One may choose a function to fit based on experience or on knowledge of the underlying distribution from which the data are drawn. For example, one may know from experience that stream flow rates are nearly always lognormally distributed when sufficient data are available to confidently define the distribution for a site. Therefore, one would use that function at a new site even though the data set is too small for its form to be clearly defined.

- One may choose a function based on the processes that generate the distribution. The addition of a large number of random variables results in a normal distribution due to the central limit theorum. The multiplication of a large number of random variables results in a lognormal distribution. Counts of independent random events result in Poisson distributions. Time to failure or to death of organisms results in Weibull distributions.

- One may use parsimonious strategies, i.e., include in the distribution nothing beyond what is known with confidence. If one feels that the shape of the distribution cannot be specified but the bounds can, a uniform distribution may be defined. If only the bounds and centroid can be estimated, they can be used to define a triangular distribution.

- Finally, if the form of the distribution is unclear or clearly does not conform to any simple function (e.g., is polymodal), an empirical distribution may be used. Even if the form is clear and conforms reasonably well to a function, empirical distributions may be preferable because they reveal the actual form and variability of the data. The only technical difficulty is the proper choice of bins to avoid excessive smoothing (too few bins) or irregularity (too many bins with too few data per bin).

For mechanistic reasons (multiplicative variance) and because many environmental data sets are approximately lognormal in shape, the lognormal distribution is the most commonly used distribution in human health and ecological risk analysis (Koch 1966; Burmaster and Hull 1997).

The selection of a distribution should proceed by a logical process of determining what functions are likely, based on knowledge of the type of data, the mechanism by which the variance in the data was generated, and the goodness of fit statistics. The function should not be selected by inappropriately applying hypothesis testing. It is common practice to assume a function and then test the null hypothesis that the data have the assumed functional form. However, this practice is inappropriate for two reasons. First, it is logically inappropriate to accept the null hypothesis when one has failed to reject it, although this is the most common conclusion drawn from such analyses. Second, it is inappropriate to apply a tool developed to prove the occurrence of treatment effects, where there may be reason to favor the null, to a problem in estimation, where there is not. Rather the assessor should choose the best distribution based on prior knowledge and relative goodness of fit of those functions that are logically plausible.

5.5.4 STATISTICAL MODELING

Statistical modeling (also called empirical modeling) is the use of statistical techniques to generate a predictive and potentially explanatory mathematical model of the relationship between one or more independent variables and a dependent variable of interest. The most obvious example in ecological risk assessment is the generation of exposure–response models from test data (Chapter 23). Other examples include quantitative structure–activity relationships

to estimate fate and effects parameters (Chapter 22 and Chapter 26), models to extrapolate toxic responses between species, durations, etc. (Chapter 26), and stock-recruitment models (Chapter 27).

Statistical modeling, in its simplest form, resembles the fitting of distribution functions to data (Section 5.5.3). The function is the same, but in one case it is simply a description of the data and in the other it is a model of the process that generated the data. For example, one might interpret a log-normal function fit to the results of a fish toxicity test as simply the distribution of a parameter, such as the proportion dying, with respect to concentrations of the tested chemical. That is, the probability of death for fathead minnows has a log-normal cumulative distribution function with the median expressed as the LC_{50}. However, the fitted function is more usefully interpreted as a model of the responses of fish populations to the chemical that may be used to estimate responses in other circumstances. Clearly, we are more interested in toxicity test results as models. However, many other fitted functions are purely descriptive such as the distribution of concentrations of a chemical in soil at a contaminated site or the distribution of plant species richness in a forest.

These two interpretations have important implications for estimating uncertainty. The error statistics for a fitted function are, by definition, appropriate estimates of uncertainty concerning the function as a description of the data. However, those error statistics are not adequate estimates of the uncertainty associated with using the same function as a model of toxicity. In other words, departure of the data from the model is not an adequate estimate of the deviation of reality from the model. The uncertainty in the toxicity model should, depending on the use of the model, include interlaboratory variance, variance in water characteristics, variance among fish populations, or other variables.

At the other extreme of model complexity, statistical modeling blends into mathematical simulation modeling. That is, after variables have been independently estimated to the extent possible, the remaining variables in a simulation model may be estimated by fitting the model to data. This process is commonly termed model calibration. Increasingly, parameter estimation for complex models is performed using Bayesian techniques (Calder et al. 2003; Clark 2005).

5.5.5 MONTE CARLO ANALYSIS AND UNCERTAINTY PROPAGATION

When mathematical models are employed with multiple uncertain or variable parameters, appropriate error propagation techniques are required. Many risk models are simple enough to perform the propagation analytically (IAEA 1989; Morgan and Henrion 1990; Hammonds et al. 1994). However, the availability of powerful personal computers and user-friendly software packages for Monte Carlo analysis has resulted in the displacement of analytical solutions by that numerical technique. Monte Carlo analysis is a resampling technique that samples from the distributions assigned to each model parameter, solves the model, saves the solution, and repeats the process until a distribution of results is generated (Figure 5.2). Reviews and guidance for Monte Carlo analysis can be found in EPA documents (Risk Assessment Forum 1996, 1997), in relevant texts (Rubinstein 1981), and in a special issue of the journal *Human and Ecological Risk Assessment* celebrating the 50th anniversary of the technique (Callahan 1996).

5.5.6 NESTED MONTE CARLO ANALYSIS

As discussed above, situations involving risks can be thought of as involving both variability and uncertainty. Both contribute to the estimation of the probability that a specified effect will occur on a particular site, but in conceptually different ways. This distinction matters when one is estimating a variable endpoint (e.g., the probability of extinction of a species given variability in sensitivity among species) and wishes to estimate the associated uncertainties or
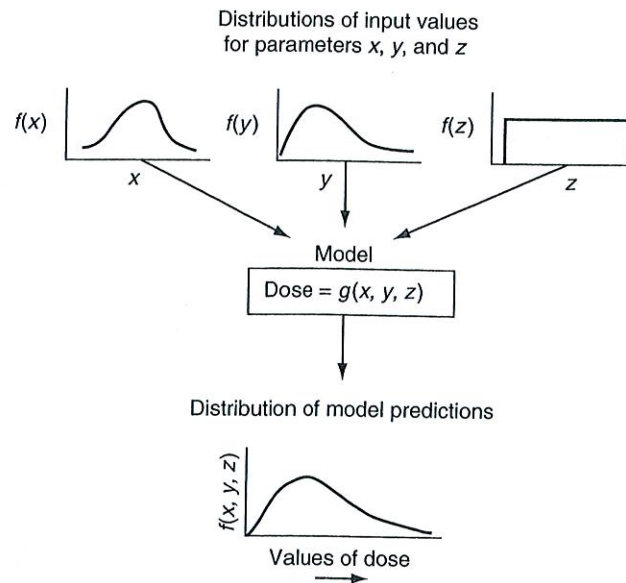
Distributions of input values
for parameters x, y, and z

FIGURE 5.2 A diagram depicting the process of Monte Carlo simulation.

when one is using models to plan a sampling and analysis program and needs to distinguish reducible uncertainties. In such cases, the parameters of models should be divided into those that are well-specified constants, those that are uncertain constants, those that are variable but well-specified, and those that are variable and uncertain. The nested Monte Carlo analysis (also called two-stage or two-dimensional Monte Carlo analysis) is begun by assigning distributions to the inherent variability of the variable parameters (e.g., dilution flow in a stream), uncertainty distributions to the uncertain parameters including the uncertain variable parameters (e.g., the uncertainty concerning the true distribution of flow), and constants to the well-specified parameters (e.g., molecular weight of the contaminant). Monte Carlo analysis is performed by first sampling from the variability distributions and then sampling from the uncertainty distributions for the uncertain variables and constants, and solving the model. By iterating the sampling, one generates a distribution of the model output based on variability and a distribution of the percentiles of that distribution based on uncertainty. An example of output from such an analysis is presented in Figure 5.3. Examples of the use of nested Monte Carlo analysis are presented in (MacIntosh et al. 1994; McKone 1994; Price et al. 1996).

Although this nested analysis is computationally complex, the greater difficulty is the conceptual problem of deciding how to classify the parameters. As discussed above, the assessor must determine how variability and uncertainty relate to the goals of the assessment and use that knowledge to consistently apply the analytical techniques. A nested analysis increases the conceptual complexity, but it may increase the likelihood of performing the analysis appropriately by compelling a more thorough review of the problem.

Although the discrimination of variability and uncertainty is the most common use of nested Monte Carlo analysis, any category of probabilities may be nested. For example, one may be interested in separately analyzing variance among entities (organisms, populations, or communities) and variance among temporal intervals (days or years) or events (spills, pesticide applications, or species introductions). In particular, one might wish to distinguish variability among fields being treated with a pesticide from variability among applications due to factors such as time to first rainfall.
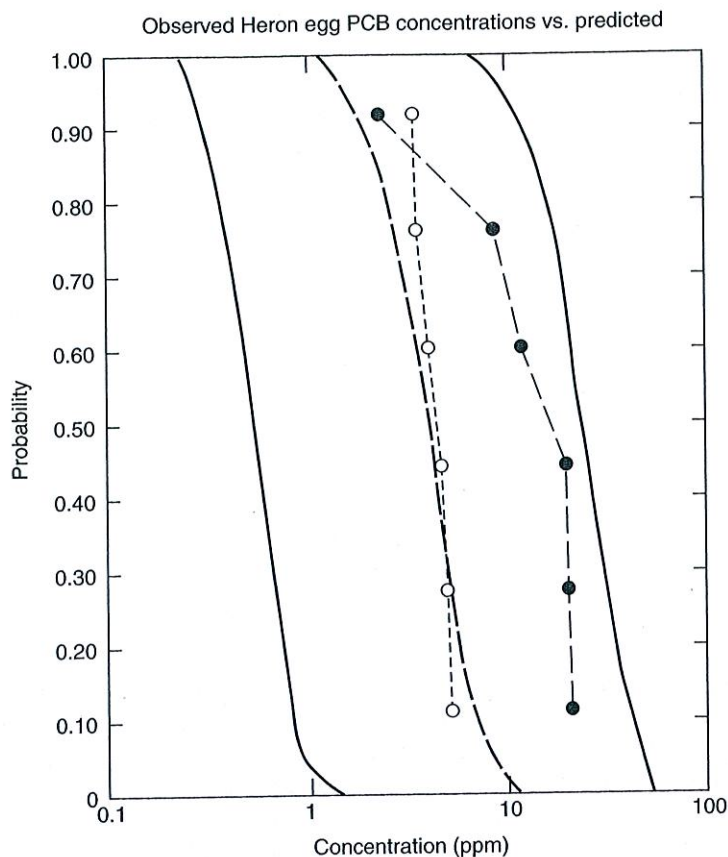
FIGURE 5.3 Complementary inverse cumulative distribution functions for concentrations of polychlor-inated biphenyls (PCBs) in great blue heron eggs. The dashed central curve represents the expected variance among eggs from different nests (i.e., due to variance among females). The solid outer curves represent the 5th and 95th percentiles on that distribution based on uncertainty in the parameters. The dots connected by dashed lines are measured concentrations in heron eggs from two rookeries. (Redrawn from MacIntosh, D.L., Suter, G.W., II, and Hoffman, F.O., *Risk Analysis*, 14, 405, 1994. With permission.)

5.5.7 SENSITIVITY ANALYSIS

Sensitivity analysis estimates the relative contribution of parameters to the outcome of an assessment. It may be performed a priori or a posteriori. A priori sensitivity analyses determine the inherent sensitivity of the model structure to variation in the parameter values. That is, if one knows the model structure but not the parameter values or their variance structures, one can still determine the rate of change in the model output with respect to an input parameter at any nominal value of the parameter by calculating the partial derivative. More commonly, the individual parameter values are displaced by a prescribed small percentage from a nominal value, and the magnitude of change in the output is noted. The ratio of the change in the output to the change in the input variable is termed sensitivity ratio or elasticity. The model is more sensitive to parameters with higher sensitivity ratios (at least in the vicinity of the assigned value). This technique is applicable even when no quantitative uncertainty analysis is performed as a means of identifying influential parameters, and it has

been recommended for determining which of the parameters should be treated probabilistically in the uncertainty analysis (Section 5.7). However, the sensitivity of the model to a parameter depends on the value assumed by the parameter (for models that are not fully linear) and its variance (for all models). Hence, the relative importance of parameters may differ from that predicted by an a priori sensitivity analysis (Gardner et al. 1981).

A posteriori sensitivity analyses determine the relative contribution of parameters to the model estimates. These analyses are typically done by performing a Monte Carlo analysis, recording the pairs of input and output values for each parameter, and regressing the input parameter values against the model output values. Greater slopes indicate greater sensitivity. Various specific techniques have been employed including single, multiple, and stepwise multiple regression (Bartell et al. 1986; Brenkert et al. 1988; IAEA 1989; Morgan and Henrion 1990). This sort of sensitivity analysis incorporates the expected values of the parameters and their assigned distributions due to variability or uncertainty. It can be used for communication of uncertainty in the risk estimates as well as to indicate which parameters of the model may be important to address in future iterations of the assessment. Good discussions of sensitivity analysis for risk models can be found in Iman and Helton (1988), Morgan and Henrion (1990), and Rose et al. (1991).

5.5.8 LISTING AND QUALITATIVE EVALUATION

Many uncertainties are not quantified because they are judged to be relatively trivial, either as a result of sensitivity analysis or based on judgment. In addition, some uncertainties are not quantified because they cannot be quantified in any reasonable manner or because the probabilities would not contribute to decision making. For example, uncertainty concerning future land uses cannot be quantified with any reasonable confidence and is more usefully treated as an assumption in the scenario than as a quantitative uncertainty in the risk models. Some uncertain parameters are not quantified, because they are set by policy. Finally, some uncertainties are not quantified because of limitations of time and resources. It is desirable to address unquantified uncertainties for the sake of openness (Section 34.2).

5.6 PROBABILITY IN THE RISK ASSESSMENT PROCESS

Unfortunately, uncertainty analysis techniques are often applied without sufficient thought to whether they make sense, given the goals of the assessment. In some cases, the results are misrepresented; more often they are presented ambiguously. As in other aspects of the risk assessment, it is appropriate to begin with a clearly defined assessment endpoint and determine the most appropriate way to estimate it (Chapter 16). For the analysis of uncertainty, one should begin by determining which of the cases in Table 5.1 is applicable. That is, must the assessment estimate a probability (e.g., probability of extinction) or a value (e.g., percent reduction in species richness), and are uncertainties of those probabilities or values to be quantitatively estimated (are they treated as determined or uncertain)? If probabilities derived from either variability or uncertainty are to be estimated, distributions must be derived.

The next step is to define those distributions, and that requires answering two questions:

- What is distributed?
- With respect to what is it distributed?

These questions need to be answered as clearly as possible. Risk distributions may result from variance or uncertainty in exposures, effects, or both.

TABLE 5.1
Types of Endpoints for Risk Assessment Categorized in Terms of Their Acknowledged Uncertainties

| | Endpoint | |
State of Knowledge	Single Value	Probability
Determined	Specified value	Probability from a specified distribution
Uncertain	Probability of an uncertain value	Probability from a probability of an uncertain distribution

Source: From Suter, G.W., II, *Guidance for Treatment of Variability and Uncertainty in Ecological Risk Assessment*, ES/ER/TM-228, Oak Ridge National Laboratory, Oak Ridge, TN, 1997.

5.6.1 DEFINING EXPOSURE DISTRIBUTIONS

In ecological risk assessments, exposure distributions are distributions of exposure metrics (e.g., concentration or dose) with respect to space, time, organisms, or belief. The specific type of space, time, or belief must be specified.

Space may be defined as arrays of points, linear units, or areas. Points are appropriate for immobile or near-immobile organisms such as plants or benthic invertebrates if the endpoint is defined in terms of individuals. For example, an assessor may be asked to determine whether any plants are exposed to toxic soils or to estimate the proportion of plants exposed to toxic soils. In those cases, one would estimate the distribution of point concentrations from the distribution of sample concentrations (assuming an appropriate sampling design). Streams are typically defined as linear units called reaches, and wildlife associated with streams such as kingfishers have territories defined in linear units. For example, an assessment of belted kingfishers would consider the distribution of exposures experienced in 0.4–2.2 km territories. Most wildlife are exposed within areas defined as territories or foraging areas. Other areas of exposure include distinct plant communities and distinct areas with a particular land use.

Time may be defined as a succession of instants, as intervals, or as incidents. Most samples are instantaneous, and distributions of such instants in time may be appropriate when temporal variance is purely random. However, few relevant exposures are instantaneous, so such distributions are most often of use when one is interested in estimating an average exposure over some period and its uncertainty. Most relevant exposures occur over some interval. For example, one may be interested in determining whether a chemical will cause an effect, which is known to require an exposure to relevant concentrations during a time x. One would then be interested in the distribution of concentrations over time intervals with duration x (i.e., moving averages). Another relevant interval is the seasonal exposure experienced by migratory species or sensitive life stages. That is, one would estimate the distribution of doses received during the seasonal interval when a life stage or species occupies the site. Finally, one may be interested in incidents that result in exposure or elevated exposure, such as storm events that flush contaminants into a stream or suspend contaminated sediments. These might be expressed as the distribution of concentrations over incidents of some specified duration or the joint distribution of concentration and duration of incidents.

Exposure may be distributed across individual organisms as in human health risk assessments, either because the endpoint is risks to individuals of an endangered species or other highly valued species, or because the endpoint is risks to populations expressed as the proportion of individuals experiencing some effect. Exposures of individuals may be

distributed due to variance in the areas they occupy, the food they consume, or inherent properties such as weight or food preferences.

Distributions of the credibility of exposure arise when the distributions are defined in terms of uncertainties or some mixture of uncertainties and variances. For example, a polyphagous and opportunistic species like mink may feed primarily on muskrats at one site, fish at another, and a relatively even mixture of prey at a third, depending on prey availability. Hence, the uncertainty concerning the mink diet at a site may be much greater than the variance among individuals at the site, in which case the fractiles of the distribution of individual dietary exposures are credibilities rather than proportions of individuals.

5.6.2 DEFINING EFFECTS DISTRIBUTIONS

In ecological risk assessments, effects distributions are distributions of responses of organisms, populations, or communities (e.g., death, abundance, or species richness) with respect to exposure. It is necessary to specify the interpretation of the effects metric and the sort of exposure with respect to which it is distributed (see above). Four general cases will be considered: effects thresholds, exposure–response relationships from toxicity tests, distributions of measures of effects, and outputs of effects simulation models.

Effects thresholds are often defined by thresholds for statistical significance such as NOAELs or LOAELs. These values do not have associated variances or other uncertainty metrics and are conventionally treated as fixed values. However, while their inherent variance is unspecified, the uncertainty associated with extrapolating them between taxa, life stages, durations, etc. can be estimated (Chapter 26). The most common approach is uncertainty factors.

In conventional laboratory or field testing, organisms are exposed to a series of chemical concentrations, doses, or some other exposure variable, and the number of organisms responding or the magnitude of responses at each exposure level is recorded. Models are fit to those data that permit the calculation of either the exposure causing a certain level of the effect or the level of effect at a given exposure (Chapter 23). If the response is dichotomous (e.g., dead or alive) or dichotomized (e.g., the continuous variable weight can be converted to normal or underweight), a frequency of response can be treated as a probability of effect in the endpoint organisms (e.g., probability of dying). Alternatively, the frequency can be treated deterministically as the proportion of an endpoint population experiencing the effect (e.g., proportion dying). If one is concerned about the uncertainties associated with these results, one might begin with the variance in the model estimate (e.g., confidence intervals) or the variance of the observations around the model (e.g., prediction intervals). However, these variances are generally smaller than the variance among tests, which is a more relevant measure of uncertainty. The minimum variance among well-performed acute tests using the same protocol and species results in test endpoint values within $\pm$ a factor of 2 or 3 (McKim 1985; Gersich et al. 1986). However, ranges of results in less uniform test sets may be more than a factor of 10 (Section 5.1.5). In addition to this variance, which is inherent in the test, extrapolations between test species, life stages, response parameters, etc. should be represented by subjective uncertainty factors, empirically derived factors, or extrapolation models (Chapter 26).

Exposure–response relationships may also be derived from observational field data. That is, the observations of co-occurring levels of a biological variable and one or more environmental characteristics may be used to generate a model (Chapter 23). For example, the number of insect taxa in streams might be related to the total suspended solids concentration. Although these models are based on real-world data, they are not necessarily better representations of causal relationships. Field data have high inherent variance due to biological, chemical, and physical variability among sites and over time, have high error due to lower quality of measurements in the field, and, most importantly, are confounded due to correlations among

the "independent" variables. In extreme cases, biological responses may be modeled as a function of a measured variable, when the true cause is an unmeasured variable. For example, episodic exposures to pesticides may be affecting streams in agricultural areas, but routine water quality monitoring seldom includes pesticides and is unlikely to detect those episodes. As a result, effects may be related to sediment rather than agricultural chemicals or to habitat rather than lawn chemicals. Hence, study design is a larger source of uncertainty in field-derived exposure–response models than the error and variance measures provided by statistics.

Finally, mathematical simulation models are used to estimate effects, particularly those on ecosystem properties or on populations mediated by population or ecosystem processes (Chapter 27 and Chapter 28). The uncertainties associated with these effects estimates are usually generated using Monte Carlo analysis.

5.6.3 ESTIMATING RISK DISTRIBUTIONS

Risk is a function of exposure and effects. If only one of these components of risk is treated as a distributed variate, the estimation of risk as a distributed variate is relatively straightforward. If the dose has been estimated to be distributed over organisms due to variance among individuals, and the effects are assumed to have a determinate threshold, the output of the risk characterization is the probability that an individual in the exposed population will receive an effective dose (Figure 5.4). However, if the exposure and effects are both estimated probabilistically, and the risk will be expressed as the joint probability, the concordance of the distributions must be ensured. If both distributions are derived as degrees of belief, concordance is not a problem; however, if they are based on variance, concordance should be assured. If one is estimating the probability that organisms of a particular species are affected, both the exposure and effects metrics must be distributed with respect to organisms. For example, if effects are distributed based on the variance among organisms of a species observed in a toxicity test, the variance in exposure should be limited to variation among organisms of a species due to weight, diet, water consumption, etc. Even though ecological risks have been
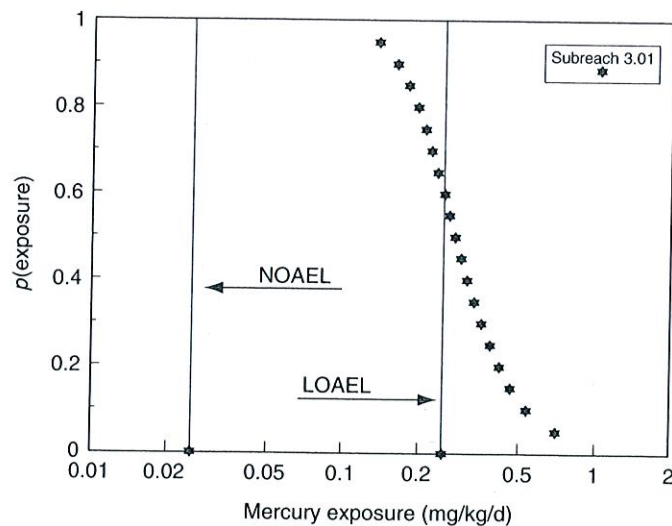

FIGURE 5.4 Inverse cumulative distribution function for exposure of rough-winged swallows to mercury, derived by Monte Carlo simulation of an oral exposure model. The vertical lines are the no observed adverse effect level (NOAEL) and the lowest observed adverse effect level (LOAEL). The probability that an individual would receive a dose greater than the LOAEL is 0.6.
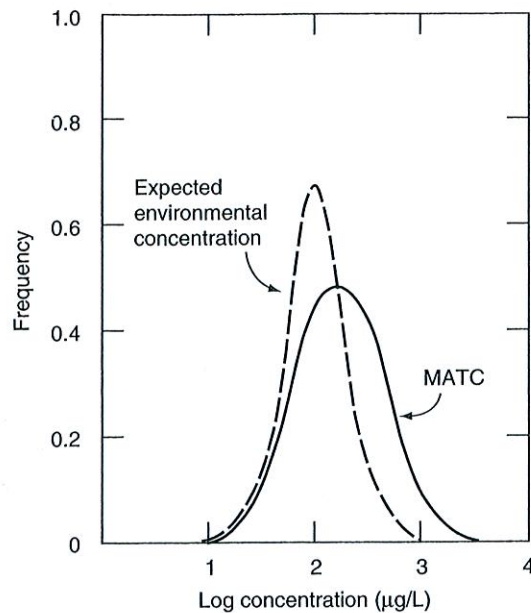
FIGURE 5.5 Probability density functions for a predicted *Salvelinus* maximum acceptable toxicant concentration (MATC) (solid line) and an expected environmental concentration (dashed line). (From Barnthouse, L.W. and Suter, G.W., II, *User's Manual for Ecological Risk Assessment*, ORNL-6251, Oak Ridge National Laboratory, Oak Ridge, TN, 1986. With permission.)

defined as the joint probability of an exposure distribution and an effects distribution since the earliest published methods (Figure 5.5), little attention has been devoted to determining and explaining what those probabilistic risks represent. Methods for addressing this issue are discussed in Chapter 30.

5.7 PARAMETERS TO TREAT AS UNCERTAIN

If the number of parameters in a risk model is large, or if research is required to define distributions, it is necessary to decide which parameters to treat probabilistically. The number of parameters is likely to be large in the early stages of a risk assessment when few data are available or in assessments that never perform much testing, measurement, sampling, or analysis. For example, once contaminant concentrations in plants have been measured, the plant uptake submodel of the wildlife exposure model can be eliminated. Hence, a multiparameter model is replaced with a single measured parameter. If a probabilistic analysis is performed that does not include all parameters, the following criteria should be used in the selection:

- If a probabilistic analysis is replacing an analysis that included uncertainty factors or conservative assumptions, the parameters to which those factors or assumptions were applied should be treated as uncertain.
- If the regulators, responsible parties, or other stakeholders have expressed concern that misspecification of a variable or uncertain parameter may affect the outcome of the assessment, that parameter should be treated as variable or uncertain.
- If the probabilistic analysis is performed in support of a decision, the parameters relevant to the decision must be treated as uncertain. For example, if the analysis is performed to

aid development of a sampling and analysis plan, the parameters that may be measured must be treated as uncertain.

- The EPA and others have specified that parameters determined to be influential by the sensitivity analysis should be treated as uncertain (Hansen 1997; Risk Assessment Forum 1997). This requirement should be applied when other more relevant criteria are not applicable. This requirement could cause the selection of parameters, such as the molecular weight of the chemical, that are not significantly uncertain or variable and are not directly relevant to the decision.

5.8   SUMMARY

This chapter provides an overview of issues related to probability that should allow the reader to understand the discussions and examples in the rest of the book. In particular, the reader should understand that there is no uniform practice of probabilistic ecological risk assessment. Rather, a variety of techniques may be applied depending on the goals of the risk assessment, the audiences for the results, the available data and models, and other considerations. These general issues will be revisited with respect to the development of the analysis plan (Chapter 18) and the characterization of variability and uncertainty (Chapter 34).