APPENDICES

A. BASICS OF PROBABILITY

A.1 Events

Any repeatable process for which the result is uncertain can be considered an experiment, such as counting failures over time or measuring time to failure of a specific item of interest. The result of one execution of the experiment is referred to as an **outcome**. Due to uncertainty associated with the process, repetitions or trials of a defined experiment would not be expected to produce the same outcomes. The set of all possible outcomes of an experiment is defined as the **sample space**.

Sample spaces can contain discrete points (such as pass, fail) or points in a continuum (such as measurement of time to failure). An **event** E is a specified set of possible outcomes in a sample space S (denoted $E \subset S$, where $\subset$ denotes subset).

Most events of interest in practical situations are compound events, formed by some composition of two or more events. Composition of events can occur through the union, intersection, or complement of events, or through some combination of these.

For two events, E_1 and E_2, in a sample space S, the **union** of E_1 and E_2 is defined to be the event containing all sample points in E_1 or E_2 or both, and is denoted by the symbol $(E_1 \cup E_2)$. Thus, a union is simply the event that either E_1 or E_2 occurs or both E_1 and E_2 occur.

For two events, E_1 and E_2, in a sample space S, the **intersection** of E_1 and E_2 is defined to be the event containing all sample points that are in both E_1 and E_2, denoted by the symbol $(E_1 \cap E_2)$. The intersection is the event that both E_1 and E_2 occur.

Figure A.1 shows a symbolic picture, called a **Venn diagram**, of some outcomes and events. In this example, the event E_1 contains three outcomes, event E_2 contains five outcomes, the union contains seven outcomes, and the intersection contains one outcome.

The **complement** of an event E is the collection of all sample points in S and not in E. The complement of E is denoted by the symbol $\overline{E}$ and is the outcomes in S that are not in E occur. In Figure A.1, the complement of E_1 is an event containing seven outcomes.

It is sometimes useful to speak of the **empty** or **null** set, a set containing no outcomes. In Figure A.1, the event E_3 is empty. It cannot occur.

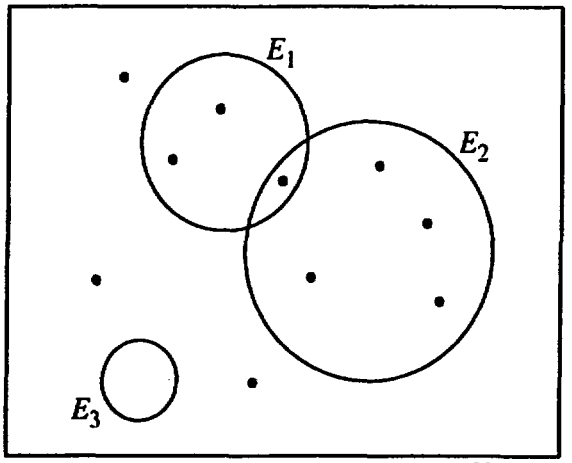

Figure A.1 Venn diagram, showing ten outcomes and three events.

Two events, E_1 and E_2, are said to be **mutually exclusive** if the event $(E_1 \cap E_2)$ contains no outcomes in the sample space S. That is, the intersection of the two events is the null set. Mutually exclusive events are also referred to as **disjoint** events. Three or more events are called mutually exclusive, or disjoint, if each pair of events is mutually exclusive. In other words, no two events can happen together.

A.2 Basic Probability Concepts

Each of the outcomes in a sample space has a probability associated with it. **Probabilities of outcomes** are seldom known; they are usually estimated from relative frequencies with which the outcomes occur when the experiment is repeated many times. Once determined, the probabilities must satisfy two requirements:

1. The probability of each outcome must be a number ≥ 0 and ≤ 1.

2. The probabilities of all outcomes in a given sample space must sum to 1.

Associated with any event E of a sample space S is the **probability of the event**, $\Pr(E)$. Since an event represents a particular set of outcomes of an experiment, the values of $\Pr(E)$ are built from the probabilities of the outcomes in E.

Probabilities are associated with each outcome in the sample space through a **probability model**. Probability

models are often developed on the basis of information derived from outcomes obtained from an experiment. Probability models are also formulated in the context of mathematical functions.

The values of Pr(E) estimated from the experimental outcomes are often defined as being representative of the *long-run relative frequency* for event E. That is, the relative frequency of an outcome will tend toward some number between 0 and 1 (inclusive) as the number of repetitions of the experiment increases. Thus, the probability of the outcome is the number about which the long-term relative frequency tends to stabilize.

This interpretation forms the basis of the **relative frequency definition of probability**, also referred to as the **frequentist view** of probability. In the frequentist view, a mathematical theory of probability is developed by deriving theorems based on the axioms of probability given in the next subsection. The probability of an event is considered to be a fixed quantity, either known or unknown, that is a property of the physical object involved and that can be estimated from data. A theorem derived from the three axioms describes the frequentist view:

If an experiment is repeated a large number of times, n, the observed relative frequency of occurrence, n_E/n, of the event E (where n_E = the number of repetitions when event E occurred) will tend to stabilize at a constant, Pr(E), referred to as the probability of E.

Another interpretation of probability leads to the so-called **classical definition of probability**, which can be stated as follows:

If an experiment can result in n equally likely and mutually exclusive outcomes and if n_E of these outcomes contain attribute E, then the probability of E is the ratio n_E / n.

For example, if each of the outcomes in Figure A.1 had equal probability, 0.1, then Pr(E_1) = 0.3, Pr(E_2) = 0.5, Pr($E_1 \cap E_2$) = 0.1, Pr($E_1 \cup E_2$) = 0.7, and Pr(E_3) = 0.

The classical definition is limited, because it assumes equally likely outcomes. However, it helps motivate the frequentist axioms mentioned above. These axioms provide a mathematical framework for probability, an overview of which is addressed in Section A.3. Some texts, including parts of this handbook, use the terms *classical* and *frequentist* interchangeably.

Another interpretation of probability is as a **subjective probability**. Probabilities obtained from the opinions of people are examples of subjective probabilities. In this concept, probability can be thought of as a rational measure of belief. Any past information about the problem being considered can be used to help assign the various probabilities. In particular, information about the relative frequency of occurrence of an event could influence the assignment of probabilities.

The notion of subjective probability is the basis for Bayesian inference. In contrast to the relative frequency definition of probability that is based on properties of events, subjective probability can be extended to situations that cannot be repeated under identical conditions. However, the assignment of subjective probabilities can be done according to certain principles so that the frequency definition requirements of probability are satisfied. All the mathematical axioms and theorems developed for frequentist probability apply to subjective probability, but their interpretation is different.

Martz and Waller (1991) present subjective probability as dealing not only with events but with propositions. A proposition is considered to be a collection of events that cannot be conceived as a series of repetitions, for example, a nuclear power plant meltdown. The degree of belief in proposition A, Pr(A), represents how strongly A is believed to be true. Thus, subjective probability refers to the degree of belief in a proposition. At the extremes, if A is believed to be true, Pr(A) = 1; if A is believed to be false, Pr(A) = 0. Points between 0 and 1 represent intermediate beliefs between false and true.

A.3 Basic Rules and Principles of Probability

The relative frequency, classical, and subjective probability definitions of probability satisfy the following axiomatic requirements of probability:

If Pr(E) is defined for a type of subset of the sample space S, and if

1. Pr(E) ≥ 0, for every event E,
2. Pr($E_1 \cup E_2 \cup \cdots$) = Pr(E_1) + Pr(E_2) + $\cdots$, where the events $E_1, E_2, \ldots$, are such that no two have a point in common, and
3. Pr(S) = 1,

then Pr(E) is called a **probability function.**

A probability function specifies how the probability is distributed over various subsets E of a sample space S.

From this definition, several rules of probability follow that provide additional properties of a probability function.

The probability of an impossible event (the empty or null set) is zero, written as:

$$\Pr(\varnothing) = 0,$$

where $\varnothing$ is the null set. The probability of the complement of E is given by:

$$\Pr(\overline{E}) = 1 - \Pr(E).$$

In general, the probability of the union of any two events is given by:

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2).$$

If E_1 and E_2 are mutually exclusive, then $\Pr(E_1 \cap E_2) = \Pr(\varnothing) = 0$, and

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2),$$

which is a special case of the second axiom of probability stated above and is sometimes referred to as the addition rule for probabilities.

For three events,

$$\begin{aligned}\Pr(E_1 \cup E_2 \cup E_3) = {} & \Pr(E_1) + \Pr(E_2) + \Pr(E_3) \\ & - \Pr(E_1 \cap E_2) - \Pr(E_1 \cap E_3) \\ & - \Pr(E_2 \cap E_3) + \Pr(E_1 \cap E_2 \cap E_3).\end{aligned}$$

This rule is also referred to as the **inclusion-exclusion principle** and can be generalized to n events. It is widely used in PRA to calculate the probability of an "or" gate (a union of events) in a fault tree (NRC 1994).

The inclusion-exclusion principle also provides useful upper and lower bounds on the probability of the union of n events that are *not* mutually exclusive. One such upper bound, referred to as the **rare event approximation**, is:

$$\Pr(E_1 \cup E_2 \cup \ldots \cup E_n) \le \Pr(E_1) + \Pr(E_2) + \ldots + \Pr(E_n).$$

The rare event approximation should only be used when the probabilities of the n events are all very small (NRC 1994). If the n events are mutually exclusive, the error is zero. A bound on the error is

$$\binom{n}{2} \max\,[\Pr(E_i)],$$

which is valid regardless of the independence of events (NRC 1994, though printed with a misprint there). The error in the rare-event approximation arises from the remaining terms in the full expansion of the left-hand side of the inequality. This approximation is frequently used in accident sequence quantification.

Many experimental situations arise in which outcomes are classified by two or more events occurring simultaneously. The simultaneous occurrence of two or more events (the intersection of events) is called a joint event, and its probability is called a **joint probability**. Thus, the joint probability of both events E_1 and E_2 occurring simultaneously is denoted by $\Pr(E_1 \cap E_2)$.

The probability associated with one event, irrespective of the outcomes for the other events, can be obtained by summing all the joint probabilities associated with all the outcomes for the other events, and is referred to as a **marginal probability**. A marginal probability is therefore the *unconditional* probability of an event, unconditioned on the occurrence of any other event.

Two events E_1 and E_2 are often related in such a way that the probability of occurrence of one depends on whether the other has or has not occurred. The **conditional probability** of one event, given that the other has occurred, is equal to the joint probability of the two events divided by the marginal probability of the given event. Thus, the conditional probability of event E_2, given event E_1 has occurred, denoted $\Pr(E_2|E_1)$, is defined as:

$$\Pr(E_2|E_1) = \Pr(E_1 \cap E_2) / \Pr(E_1), \qquad \text{(A.1)}$$

for $\Pr(E_1) > 0$. If $\Pr(E_1) = 0$, $\Pr(E_2|E_1)$ is undefined.

Rearranging this equation yields:

$$\begin{aligned}\Pr(E_1 \cap E_2) &= \Pr(E_1)\Pr(E_2|E_1) \\ &= \Pr(E_2)\Pr(E_1|E_2).\end{aligned} \qquad \text{(A.2)}$$

Calculation of joint probability requires the concept of **statistical independence**. An event E_2 is statistically independent of E_1 if the probability of E_2 does not change whenever E_1 occurs or does not occur. Thus, E_2 is **independent** of E_1 if

$$\Pr(E_2|E_1) = \Pr(E_2).$$

It follows from Equation A.1 that E_2 is independent of E_1 if their joint probability is equal to the product of the *unconditional*, or *marginal*, probabilities of the events:

$$\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2).$$

This is sometimes referred to as the multiplication rule for probabilities. In this formulation, it is clear that E_2 is independent of E_1 if E_1 is independent of E_2, and we say simply that E_1 and E_2 are statistically independent. If $\Pr(E_2)$ varies depending on whether or not event E_1 has occurred, then events E_1 and E_2 are said to be **statistically dependent**.

If $E_1, E_2, \ldots$ are mutually exclusive, and if the union of $E_1, E_2, \ldots$ equals the entire sample space, then the events $E_1, E_2, \ldots$ are said to form a **partition** of the sample space. Exactly one of the events must occur, not more than one but exactly one. In this case, the **law of total probability** says

$$\Pr(A) = \sum \Pr(A \mid E_i) \Pr(E_i) .$$

A special case can be written when there are only two sets. In this case, write E_1 simply as E and E_2 as $\overline{E}$.

Then the law of total probability simplifies to

$$\Pr(A) = \Pr(A \mid E)\Pr(E) + \Pr(A \mid \overline{E})\Pr(\overline{E})$$

for any event A. This formula is the basis for event trees, which are frequently used to diagram the possibilities in an accident sequence.

The concepts of mutually exclusive events and statistically independent events are often confused. If E_1 and E_2 are mutually exclusive events and $\Pr(E_1)$ and $\Pr(E_2)$ are nonzero, $\Pr(E_1 \cap E_2) = \Pr(\varnothing) = 0$. From Equation A.1, $\Pr(E_2 \mid E_1) = 0$, which does not equal $\Pr(E_2)$. Thus, the two events are not independent. Mutually exclusive events cannot be independent and independent events cannot be mutually exclusive.

Equation A.2 can be used to calculate the probability of the intersection of a set of events (the probability that all the events occur simultaneously). For two events E_1 and E_2, the probability of simultaneous occurrence of the events is equal to the probability of E_1 times the probability of E_2 given that E_1 has already occurred. In general, the probability of the simultaneous occurrence of n events can be written as:

$$\Pr(E_1 \cap E_2 \cap \ldots \cap E_n) =$$

$$\Pr(E_1) \Pr(E_2 \mid E_1) \Pr(E_3 \mid E_2 \cap E_1) \ldots \Pr(E_n \mid E_{n-1} \cap \ldots \cap E_1),$$

which is referred to as the **chain rule**. This rule can be used to calculate the probability that a given accident sequence occurs, with E_1 denoting the initiating event and the remaining events corresponding to the failure or success of the systems that must function in order to mitigate such an accident.

The probability of occurrence of at least one of a set of statistically independent events yields a result that is important to PRA and fault tree applications. If $E_1, E_2, \ldots, E_n$ are statistically independent events, the probability that at least one of the n events occurs is:

$$\Pr(E_1 \cup E_2 \cup \ldots \cup E_n) = \qquad \text{(A.3)}$$

$$1 - [1 - (\Pr(E_1)][1 - (\Pr(E_2)] \ldots [1 - (\Pr(E_n)],$$

which is equivalent (with expansion) to using the inclusion-exclusion rule. For the simple case where $\Pr(E_1) = \Pr(E_2) = \ldots = \Pr(E_n) = p$, the right-hand side of this expression reduces to $1 - (1 - p)^n$.

The general result in Equation A.3 has application in PRA and fault tree analysis. For example, for a system in which system failure occurs if any one of n independent events occurs, the probability of system failure is given by Equation A.3. These events could be failures of critical system components. In general, the events represent the modes by which system failure (the top event of the fault tree) can occur. These modes are referred to as the minimal cut sets of the fault tree and, if independent of each other (no minimal cut sets have common component failures), Equation A.3 applies. [See Vesely et al. (1981) for further discussion of fault trees and minimal cut sets.]

If the n events are <u>not</u> independent, the right side of Equation A.3 may be greater than or less than the left side. However, for an important situation that frequently arises in PRA, the right side of Equation A.3 forms an upper bound for the left side.

If the n events are cut sets that are positively associated [see Esary and Proschan (1970, 1963)], then the right side is an upper bound for $\Pr(E_1 \cup E_2 \cup \ldots \cup E_n)$ and is known as the **min cut upper bound** (NRC 1994). This name arises from common PRA applications where E_i is the i^{th} minimal cut set of a system or accident sequence of interest. In this case, the min cut upper bound is superior to the rare event approximation and can never exceed unity (as can happen with the rare event approximation). If the n events satisfy conditions similar to those of the rare event approximation, the min cut set upper bound is a useful approximation to the left side of Equation A.3. Note that the min cut upper bound is <u>not</u> applicable for mutually exclusive events.

A.4 Random Variables and Probability Distributions

A.4.1 Random Variables

A **random variable** is any rule that associates real numbers with the outcomes of an experiment. For example, the number of initiating events in one year, the number of failures to start in 12 demands, and the time to complete a repair of a pump are all random variables.

If the numbers associated with the outcomes of an experiment are all distinct and countable, the corresponding random variable is called a **discrete** random variable. The number of initiating events and the number of failures to start are examples of discrete random variables.

If the sample space contains an infinite number of outcomes (like those contained in any interval), the random variable is **continuous**. Time T is a common continuous random variable, for example, time to failure, time between failures, or time to repair, where the random variable T can assume any value over the range 0 to ∞.

A.4.2 Probability Distributions

A **probability function** (introduced at the beginning of Section A.3) associates a probability with each possible value of a random variable and, thus, describes the distribution of probability for the random variable. For a discrete random variable, this function is referred to as a **discrete probability distribution function** (p.d.f.). A discrete p.d.f., commonly denoted by f, is also referred to as a **discrete distribution**, or **discrete probability mass function**.

If x denotes a value that the discrete random variable X can assume, the probability distribution function is often denoted $\Pr(x)$. The notation used here is that a random variable is denoted by an *upper-case letter* and an observed or observable value of the random variable (a number) is denoted by a *lower-case letter*. The sum of the probabilities over all the possible values of x must be 1. Thus, we write $f(x) = \Pr(X = x)$, and require $\Sigma_i f(x_i) = 1$.

Certain discrete random variables have wide application and have therefore been defined and given specific names. The two most commonly used discrete random variables in PRA applications are the **binomial** and **Poisson** random variables, which are presented in Section A.6.

A continuously distributed random variable has a **density function**, a nonnegative integrable function, with the area between the graph of the function and the horizontal axis equal to 1. This density function is also referred to as the **continuous probability distribution function** (p.d.f.). If x denotes a value that the continuous random variable X can assume, the p.d.f. is often denoted as $f(x)$. The probability that X takes a value in a region A is the integral of $f(x)$ over A. In particular,

$$\Pr(a \le X \le b) = \int_a^b f(x)dx$$

and

$$\Pr(x \le X \le x + \Delta x) \approx f(x)\Delta x \qquad \text{(A.4)}$$

for small Δx. Also,

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

The most commonly used continuous distributions in PRA are the **lognormal**, **exponential**, **gamma**, and **beta** distributions. Section A.7 summarizes the essential facts about these distributions, and also about less common but occasionally required distributions: **uniform**, **normal**, **Weibull**, **chi-squared**, **inverted gamma**, **logistic-normal**, **Student's** t, F, and **Dirichlet**.

A.4.3 Cumulative Distribution Functions

Discrete probability distributions provide point probabilities for discrete random variables and continuous p.d.f.s provide point densities for continuous random variables. A related function useful in probability and PRA is the **cumulative distribution function** (c.d.f.). This function is defined as the probability that the random variable assumes values less than or equal to the specific value x, and is denoted $F(x)$.

For a discrete random variable X, with outcomes x_i, and the corresponding probabilities $\Pr(x_i)$, $F(x)$ is the sum of the probabilities of all $x_i \le x$. That is,

$$F(x) = \Pr(X \le x) = \sum_{x_i \le x} \Pr(x_i) .$$

For a continuous random variable X, $F(x)$ is the area beneath the p.d.f. $f(x)$ up to x. That is, $F(x)$ is the integral of $f(x)$:

$$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(y)dy .$$

If X takes on only positive values, the lower limit of integration can be set to 0. The upper limit is x, and $f(x)$ is the derivative of $F(x)$. Note that, because $F(x)$ is a probability, $0 \le F(x) \le 1$. If X ranges from $-\infty$ to $+\infty$, then

$F(-\infty) = 0$ and $F(+\infty) = 1$.

If X has a restricted range, with a and b being the lower and upper limits of X respectively, $a < X < b$, then

$F(a) = 0$ and $F(b) = 1$.

Also, $F(x)$ is a nondecreasing function of x, that is,

if $x_2 > x_1$, $F(x_2) \ge F(x_1)$.

Another important property of $F(x)$ is that

$\Pr(x_1 < X \le x_2) = F(x_2) - F(x_1)$

for discrete random variables and

$\Pr(x_1 \le X \le x_2) = F(x_2) - F(x_1)$

for continuous random variables.

An example of a p.d.f. and the associated c.d.f. for a continuous distribution is shown in Figure A.2.

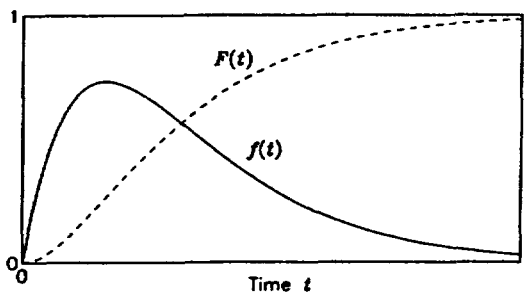

Figure A.2 Probability density function (p.d.f.) and cumulative distribution function (c.d.f.).

A.4.4 Reliability and Hazard Functions

A.4.4.1 Definitions

There are also characterizations that have special interpretations for time-to-failure distributions. Let T denote the random time to failure of a system. The **reliability function** of a system is defined as

$R(t) = \Pr(T > t)$.

Hence, $R(t)$, called the **reliability** at time t, is the probability that the system does not fail in the time interval $(0, t)$ or equivalently, the probability that the system is still operating at time t. (This discussion uses the notation (a, b) to mean the set of times $> a$ and $\le b$, but the distinction between $<$ and $\le$ is a mathematical fine point, not important in practice.) The reliability function is also sometimes called the **survival function**. It is equal to $1 - F(t)$.

When used as a reliability criterion, it is common to state a time, say t_0, called the **mission time**, and require for a system that the reliability at mission time t_0 be at least some prescribed level, say R_0. For example, a pump might be required to operate successfully for at least 12 hours with probability at least 0.95. The requirement in this case is $R_0 = 0.95$ and $t_0 = 12$. In terms of the reliability function, this would mean $R(12) \ge 0.95$. One interpretation would be that such a pump would perform for the required mission time for 95% of the situations when it is called on to do so. Another interpretation is that 95% of all such pumps would perform as required.

Consider a system that operates for a particular mission time, unless it fails. If it fails, no immediate repairs are attempted, so some authors call the system **nonrepairable**. A common way to characterize this system's reliability is in terms of the **hazard function**. Suppose that the system is still operating at time t, and consider the probability that it will fail in a small interval of time $(t, t + \Delta t)$. This is the conditional probability $\Pr(t < T \le t + \Delta t \mid T > t)$. The hazard function, h, is defined so that when Δt is small,

$$h(t)\Delta t \approx \Pr(t < T \le t + \Delta t \mid T > t) \ . \qquad \text{(A.5)}$$

This function is also encountered, under the name of λ, in some treatments of Poisson processes. Equation A.5 gives, approximately,

$$h(t)\Delta t \approx \frac{\Pr(t < T \le t + \Delta t)}{\Pr(T > t)} \approx \frac{f(t)\Delta t}{R(t)}$$

This is the basis for the formal definition of h:

$$h(t) = \frac{f(t)}{R(t)}$$

For details, see Bain and Engelhardt (1992, p. 541). Equation A.5 is analogous to Equation A.4, except that the probability in Equation A.5 is conditional on the system having survived until t, whereas Equation A.4 refers to all systems in the original population, either still surviving or not. Suppose a large number, say N, of identical systems are put into operation at time $t = 0$, and n is the number which fail in the interval $(t, t + \Delta t)$. It follows that $f(t)\Delta t \approx n/N$, the observed relative frequency of systems failed in the interval $(t, t + \Delta t)$. On the other hand, if N_t denotes the number of the original N systems which are still in operation at time t, then $h(t)\Delta t \approx n/N_t$, the observed relative frequency of *surviving* systems which fail in this same interval. Thus, $f(t)$ is a measure of the risk of failing at time t for any system in the original set, whereas $h(t)$ is a measure of the risk of failing at time t, but only for systems that have survived this long.

The hazard function is used as a measure of "aging" for systems in the population. If $h(t)$ is an increasing function, then systems are aging or wearing out with time. Of course, in general the hazard function can exhibit many types of behavior other than increasing with time. In actuarial science the hazard function is called the **force of mortality**, and it is used as a measure of aging for individuals in a population. More generally, the hazard function gives an indication of "proneness to failure" of a system after time t has elapsed. Other terms which are also used instead of hazard function are **hazard rate** and **failure rate**. The term *failure rate* is often used in other ways in the literature of reliability [see Ascher and Feingold (1984), p. 19].

A.4.4.2 Relations among p.d.f., Reliability, and Hazard

Any one of the functions F, f, R, and h completely characterizes the distribution, and uniquely determines the other three functions. The definition

$$h(t) = \frac{f(t)}{R(t)}$$

was given above. The right side can be written as the derivative of $-\ln[R(t)]$, leading to

$$R(t) = \exp\left(-\int_0^t h(u)du\right) = \exp\left(-H(t)\right)$$

where the function $H(t)$ is called the **cumulative hazard function**. The reliability function, $R(t)$, and the c.d.f., $F(t) = 1 - R(t)$, are therefore uniquely determined by the hazard function, $h(t)$, and the p.d.f. can be expressed as

$$f(t) = h(t)\exp\left(-\int_0^t h(u)du\right).$$

Figure A.3 shows the reliability, hazard and the cumulative hazard function for the example of Figure A.2.

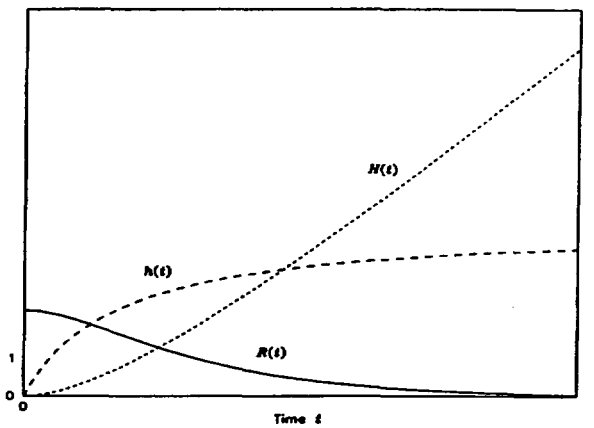

Figure A.3 The reliability function, hazard function and cumulative hazard function.

The hazard function in Figure A.3 is an increasing function of time. Therefore, it would be consistent with systems with a dominant wear-out effect for the entire life of the system. The lifetime of a system may be divided into three typical intervals: the **burn-in** or **infant** period, the **random** or **chance** failure period, and the **wear-out** period. During the useful period, the dominant cause of failures is "random" failures. For example, systems might fail due to external causes such as power surges or other environmental factors rather than problems attributable to the defects or wear-out in the systems. This example is somewhat idealized because for many types of systems the hazard function will tend to increase slowly during the later stages of the chance failure period. This is particularly true of mechanical systems. On the other hand, for many electrical components such as transistors and other solid-state devices, the hazard function remains fairly flat once the burn-in failure period is over.

A.4.5 Joint, Marginal, and Conditional Distributions

Many statistical methods are based on selecting a sample of size n from a probability distribution $f(x)$. Such a sample is denoted by

$$(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = (x_1, x_2, ..., x_n),$$

where $x_1, x_2, ..., x_n$ are the actual values of the random variable X which has the distribution $f(x)$.

The concepts of simultaneous events and joint, marginal, and conditional probability, discussed in Section A.3, also pertain to random variables and probability distributions. Two random variables X_1 and X_2 (both continuous, both discrete, or one of each) can have a **joint distribution**, or joint p.d.f., denoted $f(x_1, x_2)$. The point (x_1, x_2) can be thought of as a point in two-dimensional Euclidean space. Similarly, n random variables have joint distribution $f(x_1, x_2, ..., x_n)$. Also, the n random variables have joint cumulative distribution $F(x_1, x_2, ..., x_n)$.

The **marginal distribution** of X_i is defined as the joint p.d.f. integrated (for continuous random variables) or summed (for discrete random variables) over the $n-1$ other corresponding dimensions, resulting in a function of x_i alone. Thus, the marginal distribution of X_i is the unconditional p.d.f. of X_i, $f_i(x_i)$.

The **conditional distribution** of X_1 given X_2, denoted $f(x_1 \mid x_2)$, is defined by

$$f(x_1 | x_2) = \frac{f(x_1, x_2)}{f_2(x_2)},$$

where $f_2(x_2) \neq 0$. This conditional distribution can be shown to satisfy the requirements of a probability function. Sampling from a conditional p.d.f. would produce only those values of X_1 that could occur for a given value of $X_2 = x_2$. The concept of a conditional distribution also extends to n random variables.

Two random variables X_1 and X_2 are independent if their joint p.d.f. is equal to the product of the two individual p.d.f.s. That is,

$$f(x_1, x_2) = f(x_1) f(x_2).$$

In general, $X_1, X_2, ..., X_n$ are **independent random variables** if

$$f(x_1, x_2, ..., x_n) = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n).$$

A.4.6 Characterizing Random Variables and Their Distributions

A.4.6.1 Distribution Characteristics

Probability distributions have many characteristics of interest, some of which are described by **distribution parameters**. The term parameter is used to refer to a fixed characteristic. In contrast to a statistic, which changes from sample to sample, a parameter for a particular distribution is a constant and does not change. However, when a parameter's value is not known, sample statistics can be used to estimate the parameter value. Parameter estimation is discussed in Appendix B.

A very useful distribution characteristic is the parameter that serves as a measure of central tendency, which can be viewed as a measure of the middle of a distribution. When a change in the parameter slides the distribution sideways, as with the mean of a normal distribution, the parameter is referred to as the **location parameter**. It serves to locate the distribution along the horizontal axis. Sometimes, however, a change in the parameter squeezes or stretches the distribution toward or away from zero, as with the mean of the exponential distribution. In that case, the parameter is a **scale parameter**.

In any case, the most common measure of central tendency is the **mean**, μ, of the distribution, which is a weighted average of the outcomes, with the weights being probabilities of outcomes. For a discrete random variable X,

$$\mu_X = \sum_i x_i \Pr(x_i).$$

For a continuous random variable X,

$$\mu_X = \int_{-\infty}^{\infty} x f(x) dx.$$

(In Section A.4.6.2 below, the mean of X will be denoted $E(X)$, the "expected value" of X.)

Another distribution characteristic commonly used as a measure of central tendency, or location, is the **median**. For a continuous distribution, the median is the point along the horizontal axis for which 50% of the area under the p.d.f. lies to its left and the other 50% to its right. The median of a random variable, X, is commonly designated med(X) or $x_{.50}$ and, for a continuous distribution, is the value for which $\Pr(X \leq x_{.50}) = .50$ and $\Pr(X \geq x_{.50}) = .50$. In terms of the cumulative distribution, $F(x_{.50}) = .50$. The median is a specific case of the general 100αth **percentile**, x_α, for which $F(x_\alpha) = \alpha$. When the factor of 100 is dropped, x_α is called the α **quantile**. Along with the median as the 50th percentile (or equivalently, the 0.5 quantile), the 25th and 75th percentiles are referred to as **quartiles** of a distribution.

Figure A.4 shows the quartiles, $x_{0.25}$ and $x_{0.75}$, the median, $x_{0.50}$, and the mean. The quartiles and the median divide the area under the density curve into four pieces, each with the same area. Note that the mean is greater than the median in this example, which is the usual relation when the density has a long right tail, as this one does.

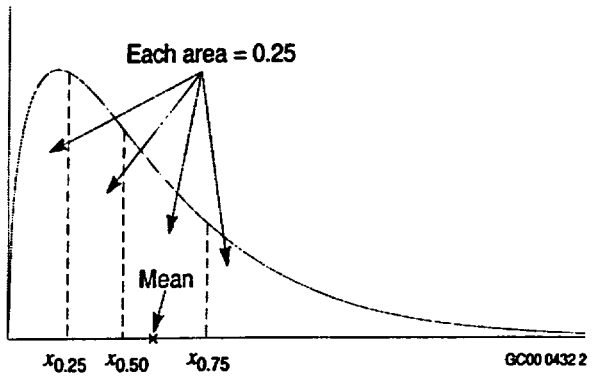

Figure A.4 Density, showing quartiles, median, and mean.

Figure A.5 shows the same quantities plotted with the c.d.f. By definition, the q quantile, x_q, satisfies $F(x_q) = q$.

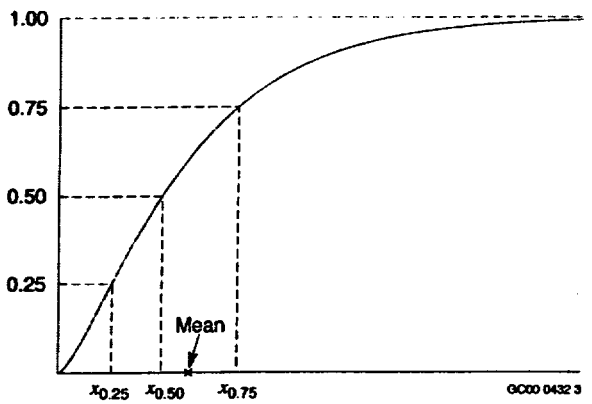

Figure A.5 Cumulative distribution function (c.d.f.) showing quartiles, median, and mean.

The mean and the median are used to measure the center or location of a distribution. Since the median is less affected by tail-area probabilities, it can be viewed as a better measure of location than the mean for highly-skewed distributions. For symmetric distributions, the mean and median are equivalent.

A different measure of center or location of a distribution is the **mode**, which indicates the most probable outcome of a distribution. The mode is the point along the horizontal axis where the "peak" or maximum of the p.d.f. is located. Note that the mode does not necessarily have to be near the middle of the distribution. It simply indicates the most likely value of a distribution. Note also that a peak does not have to exist and, in some cases, more than one peak can exist.

Another important characteristic of a distribution is its **variance**, denoted by σ^2. The variance is the average of the squared deviations from the mean. The **standard deviation**, σ, of a distribution is the square root of its variance. Both the variance and standard deviation are measures of a distribution's spread or dispersion. For a discrete random variable X,

$$\sigma_X^2 = \sum_i (x_i - \mu)^2 \Pr(x_i) .$$

For a continuous random variable X,

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx .$$

Though less used than the mean and variance, the **skewness** is defined as

$$E(X - \mu)^3 / \sigma^3 .$$

It measures asymmetry. It is usually positive if the density has a longer right tail than left tail, and negative if the density has a longer left tail than right tail. For example, the density in Figure A.4 has positive skewness.

A.4.6.2 Mathematical Expectation

The definitions of distribution means and variances arise from **mathematical expectation and moments of a distribution**, which form an important method for calculating the parameters of a known p.d.f. In general, the **expectation (expected value** or mathematical expectation) of a function $g(X)$, denoted $E[g(X)]$, is

$$E[g(X)] = \sum_i g(x_i) \Pr(x_i) ,$$

when X is discrete, and

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx ,$$

when X is continuous.

Because of their wide use, several expectations have special names. For $g(X) = X$, the expectation $E(X)$ becomes the mean of X. Thus, the mean is also commonly referred to as the expected value (or expectation) of the random variable X. In addition, for $g(X) = X$, the expectation $E(X)$ is known as the **first moment about the origin.**

The variance, σ_X^2, also denoted by $\mathrm{Var}(X)$, of a distribution is defined by mathematical expectation with $g(X) = (X - \mu_X)^2$. Thus,

$$\mathrm{Var}(X) = \sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - [E(X)]^2,$$

which is known as the **second moment about the mean.**

Ordinary moments (moments about the origin) of a random variable X are defined as

$$M_r = E(X^r),$$

for $r = 1, 2, \ldots$. Thus,

$$\mathrm{Var}(X) = \sigma_X^2 = E(X^2) - [E(X)]^2 = M_2 - M_1^2.$$

Central moments (moments about the mean) of a random variable X are defined as being equal to $E[(X - \mu)^r]$ for $r = 2, 3, \ldots$. The ordinary and central moments can be seen to define characteristics of distributions of random variables.

An important rule of expectation commonly used in PRA is that the expected value of a product of independent random variables is the product of their respective expected values. That is, $E(X_1 \cdot X_2 \cdot \ldots \cdot X_n) = E(X_1) \cdot E(X_2) \cdot \ldots \cdot E(X_n)$ when all X_i are independent. This rule also applies to conditionally independent random variables. If the random variables $X_2, X_3, \ldots, X_n$ are all conditionally independent given $X_1 = x_1$, then

$$f(x_2, x_3, \ldots, x_n | x_1) = f(x_2|x_1) \cdot f(x_3|x_1) \cdot \ldots \cdot f(x_n|x_1).$$

It follows that

$$E(X_2 \cdot X_3 \cdot \ldots \cdot X_n | x_1) = E(X_2|x_1) \cdot E(X_3|x_1) \cdot \ldots \cdot E(X_n|x_1).$$

Thus,

$$E(X_1 \cdot X_2 \cdot \ldots \cdot X_n) = E[X_1 \cdot E(X_2|x_1) \cdot E(X_3|x_1) \cdot \ldots \cdot E(X_n|x_1)].$$

The following facts are also often useful:

- $E(\Sigma_i X_i) = \Sigma_i E(X_i)$, whether or not the X_is are independent.
- $\mathrm{Var}(\Sigma_i X_i) = \Sigma_i \mathrm{Var}(X_i)$, if the X_is are independent.
- $E(aX + b) = aE(X) + b$.
- $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
- The last two give useful special cases when $a = 1$ or $b = 0$.

A.4.6.3 Moment-Generating Functions

Another special mathematical expectation is the **moment-generating function** of a random variable. For a random variable X with p.d.f. $f(x)$, the moment-generating function of X (or of the distribution) is defined by $M(t) = E(e^{tX})$, if M exists for some interval $-h < t < h$. Therefore, if X is a continuous random variable,

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

If X is a discrete random variable,

$$M(t) = \sum_i e^{tx_i} f(x_i).$$

Note that not every distribution has a moment-generating function.

The importance of the moment-generating function is that, when it does exist, it is unique and completely specifies the distribution of the random variable. If two random variables have the same moment-generating function, they have the same distribution.

It can be shown that the moments of a distribution can be found from the series expansion of $M(t)$. The moments of the distribution can also be determined from the moment-generating function by differentiating the moment-generating function with respect to t and setting $t = 0$. See Martz and Waller (1991) and any of several mathematical statistics texts, such as Hogg and Craig (1995), for further details on moment-generating functions.

A.4.6.4 Covariance and Correlation

For two random variables, X and Y, with means μ_x and μ_y, the expected value $E[(X - \mu_x)(Y - \mu_y)]$ is called the **covariance** of X and Y, denoted $\mathrm{Cov}(X, Y)$. The covariance of X and Y divided by the product of the standard deviations of X and Y is called the **correlation coefficient** (or correlation) between X and Y, denoted $\mathrm{Cor}(X, Y)$. That is,

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

$$= \frac{E(X-\mu_X)E(Y-\mu_Y)}{\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]}} .$$

The correlation coefficient measures the degree of association between X and Y, that is, the strength of a linear relationship between X and Y. It is always between -1 and 1. Positive correlation (correlation > 0) means that X and Y tend to be large together and small together in a linear way. Negative correlation means that X tends to be large when Y is small and vice versa, in a linear way. If X and Y are independent, then their correlation is zero. The converse is not true; examples can be constructed where X and Y are dependent (in a nonlinear way) yet have zero correlation.

A.4.7 Distribution of a Transformed Random Variable

This section considers the distribution of $Y = h(X)$, when X has a known distribution and h is a known function. The problem is straightforward when X has a discrete distribution. When X is continuous and h is monotone, either increasing or decreasing, the c.d.f.s are also related in the natural way, as follows. Let F be the c.d.f. of X and let G be the c.d.f. of Y. Then we have

$$G(y) = \Pr(Y \le y) = \Pr[h(X) \le y] .$$

If h is monotone increasing, this equals

$$\Pr[X \le h^{-1}(y)] = F(x),$$

where x and y are related by $y = h(x)$, $x = h^{-1}(y)$. In summary, $G(y) = F(x)$.

If, instead, h is monotone decreasing, then a similar argument gives

$$G(y) = 1 - F(x) .$$

The surprise comes with the densities. Differentiate both sides of either of the above equations with respect to y, to obtain the density of y. This involves using the chain rule for differentiation. The result is

$$g(y) = f(x)\left|\frac{dx}{dy}\right| .$$

That is, the density of Y is not simply equal to the density of X with a different argument. There is also a multiplier, the absolute value of the derivative.

Two important special cases are given here. If $Y = \exp(X)$, then

$$g(y) = f[\ln(y)](1/y) .$$

If $Y = 1/X$, then

$$g(y) = f(1/y)(1/y^2) .$$

These formulas form the basis for the densities of the lognormal distribution and the inverted gamma distribution.

A.5 Bayes' Theorem

It is frequently desired to calculate the probability of an event A given that another event B has occurred at some prior point in time. It can also be of interest to calculate the probability that a state of nature exists given that a certain sample is observed, for example, belonging to a certain population based on a sample measurement or observation. Conditional probability leads directly to **Bayes' Theorem**, which, along with subjective probability, forms the basis for Bayesian inference commonly used in PRA.

Recall the definition of a partition from Section A.3: $A_1, A_2, \ldots, A_n$ are a partition of the sample space if the are mutually exclusive and their union equals the entire sample space. Bayes' Theorem states that if $A_1, A_2, \ldots, A_n$ are a partition of the sample space and if B is any other event such that $\Pr(B) > 0$, then

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\Pr(B)}, \qquad \text{(A.6)}$$

where

$$\Pr(B) = \sum_{j=1}^{n} \Pr(B|A_j)\Pr(A_j) .$$

This last equation is the law of total probability (Section A.3). Equation A.6 follows from the definition of conditional probability in Equation A.1:

$$\Pr(A_i|B) = \frac{\Pr(B \cap A_i)}{\Pr(B)} = \frac{\Pr(B|A_i)\Pr(A_i)}{\Pr(B)} .$$

The $\Pr(A_i|B)$ is the **posterior** (or a posteriori) probability for the event A_i, meaning the probability of A_i once B is known. The $\Pr(A_i)$ is the **prior** (or a priori) probability of the event A_i before experimentation or observation. The event B is the observation. The $\Pr(B|A_i)$ is the probability of the observation given that A_i is true. The denominator serves as a normalizing constant.

Calculating the posterior probabilities $\Pr(A_i|B)$ requires knowledge of the probabilities $\Pr(A_i)$ and $\Pr(B|A_i)$, $i = 1, 2, ..., n$. The probability of an event can often be determined if the population is known, thus, the $\Pr(B|A_i)$ can be determined. However, the $\Pr(A_i)$, $i = 1, 2, ..., n$, are the probabilities that certain states of nature exist and are either unknown or difficult to ascertain. These probabilities, $\Pr(A_i)$, are called prior probabilities for the events A_i because they specify the distribution of the states of nature prior to conducting the experiment.

Application of Bayes' Theorem utilizes the fact that $\Pr(B|A_i)$ is easier to calculate than $\Pr(A_i|B)$. If probability is viewed as degree of belief, then the prior belief is changed, by the test evidence, to a posterior degree of belief. In many situations, some knowledge of the prior probabilities for the events $A_1, A_2, ..., A_n$ exists. Using this prior information, inferring which of the sets $A_1, A_2, ..., A_n$ is the true population can be achieved by calculating the $\Pr(A_i|B)$ and selecting the population that produces the highest probability.

Equation A.6 pertains to disjoint discrete events and discrete probability distributions. Bayes' Theorem has analogous results for continuous p.d.f.'s. The continuous version is given here. Suppose X is a discrete or continuous random variable, with p.d.f. depending on parameter θ, and with conditional p.d.f. of X, given θ, specified by $f(x|\theta)$. Suppose that θ has a continuous probability distribution with p.d.f. $g(\theta)$. This can happen in two ways: either θ is a possible value of the random variable Θ (using the convention of denoting random variables with uppercase letters), or else θ is an uncertain parameter with a subjective uncertainty distribution. The second case is the more common one. Call $g(\theta)$ the **prior p.d.f.** Then for every x such that $f(x) > 0$ exists, the **posterior p.d.f.** of θ, given $X = x$, is

$$g(\theta|x) = \frac{f(x|\theta)g(\theta)}{f(x)}, \qquad \text{(A.7)}$$

where

$$f(x) = \int f(x|\theta)g(\theta)d\theta$$

is the marginal p.d.f. of X. Again, the prior and posterior p.d.f.'s can be used to represent the probability of various values θ prior to and posterior to observing a value of another random variable X. This is valid whether "probability of θ" has a frequentist or subjective interpretation.

A.6 Discrete Random Variables

A.6.1 The Binomial Distribution

The **binomial** distribution describes the number of failures X in n independent trials. The random variable X has a binomial distribution if:

1. The number of random trials is one or more and is known in advance.
2. Each trial results in one of two outcomes, usually called success and failure (although they could be pass-fail, hit-miss, defective-nondefective, etc.).
3. The outcomes for different trials are statistically independent.
4. The probability of failure, p, is constant across trials.

Equal to the number of failures in the n trials, a binomial random variable X can take on any integer value from 0 to n. The probability associated with each of these possible outcomes, x, is defined by the binomial(n, p) p.d.f. as

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$x = 0, ..., n.$$

Here

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is the **binomial coefficient**. The symbol

$$n! = n(n-1)(n-2)...(2)(1)$$

denotes n factorial, with 0! defined to be equal to 1. This binomial coefficient provides the number of ways that exactly x failures can occur in n trials (number of combinations of n trials selected x at a time).

The binomial distribution has two parameters, n and p, of which n is known. (Although n may not always be known exactly, it is treated as known in this handbook.)

The mean and variance of a binomial(n, p) random variable X are

$E(X) = np$

and

$\text{Var}(X) = np(1 - p)$.

Figure A.6 shows three binomial probability distribution functions, with parameter $p = 0.25$, and $n = 4$, 12, and 40. In each case, the mean is np. The means have been aligned in the three plots.

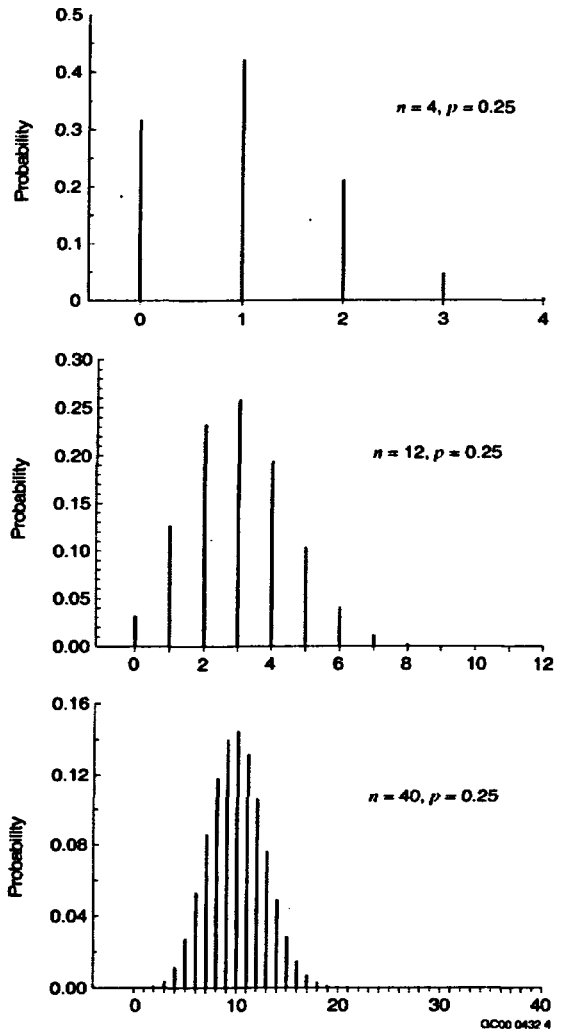

Figure A.6 Three binomial probability distribution functions.

A.6.2 The Poisson Distribution

The Poisson distribution provides a discrete probability model that is appropriate for many random phenomena that involve counts. Examples are counts per fixed time interval of the number of items that fail, the number of customers arriving for service, and the number of telephone calls occurring. A common use of the Poisson distribution is to describe the behavior of many rare event occurrences. The Poisson distribution is also frequently used in applications to describe the occurrence of system or component failures under steady-state conditions.

The count phenomena that occur as Poisson random variables are not necessarily restricted to occurring over a time interval. They could also be counts of things occurring in some region, such as defects on a surface or within a certain material. A process that leads to a Poisson random variable is said to be a **Poisson process.**

The Poisson distribution describes the total number of events occurring in some interval of time t (or space). The p.d.f. of a Poisson random variable X, with parameter $\mu = \lambda t$, is

$$\Pr(X = x) = \frac{e^{-\mu}\mu^x}{x!} = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \tag{A.8}$$

for $x = 0, 1, 2, \ldots$, and $x! = x(x - 1)(x - 2) \ldots (2)(1)$, as defined previously.

The Poisson distribution has a single parameter μ, denoted Poisson(μ). If X denotes the number of events that occur during some time period of length t, then X is often assumed to have a Poisson distribution with parameter $\mu = \lambda t$. In this case, X is considered to be a Poisson process with **intensity** $\lambda > 0$ (Martz and Waller 1991). The variable λ is also referred to as the **event rate** (or **failure rate** when the events are failures). Note that λ has units 1/time; thus, $\lambda t = \mu$ is dimensionless.

If only the total number of occurrences for a single time period t is of interest, the form of the p.d.f. in Equation A.8 using μ is simpler. If the event rate, λ, or various time periods, t, are of interest, the form of the p.d.f. in Equation A.8 using λt is more useful.

The expected number of events occurring in the interval 0 to t is $\mu = \lambda t$. Thus, the mean of the Poisson distribution is equal to the parameter of the distribution, which is why μ is often used to represent the parameter. The variance of the Poisson distribution is also equal to the parameter of the distribution. Therefore, for a Poisson(μ) random variable X,

$E(X) = \text{Var}(X) = \mu = \lambda t$.

Figure A.7 shows three Poisson probability distribution functions, with means $\mu = 1.0$, 3.0, and 10.0, respectively. The three means have been aligned in the graphs. Note the similarity between the Poisson distribution and the binomial distribution when $\mu = np$ and n is not too small.

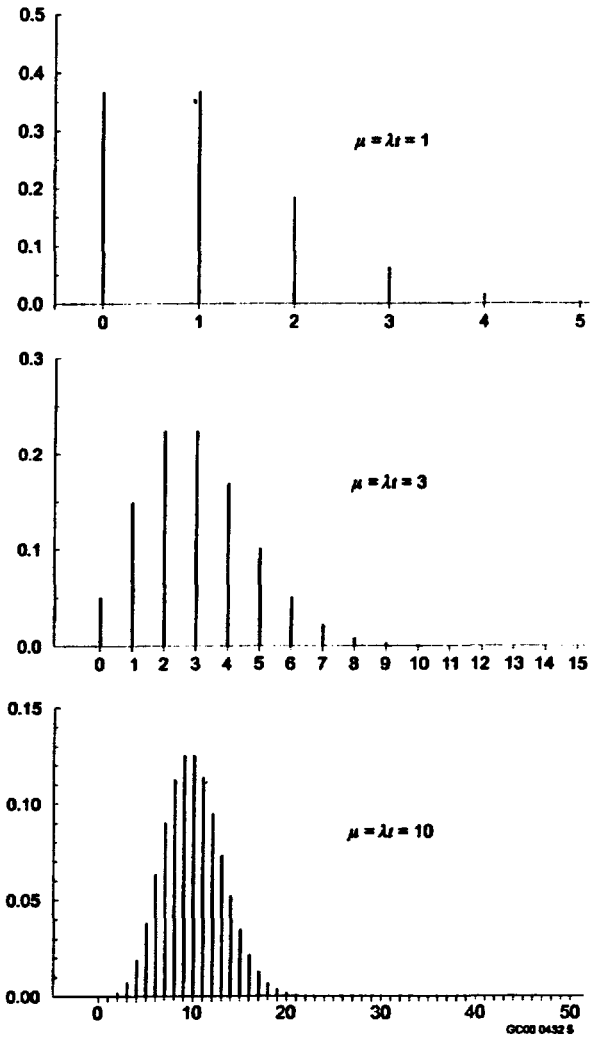

Figure A.7 Three Poisson probability distribution functions.

Several conditions are assumed to hold for a Poisson process that produces a Poisson random variable:

1. For small intervals, the probability of exactly one occurrence is approximately proportional to the length of the interval (where λ, the event rate or intensity, is the constant of proportionality).
2. For small intervals, the probability of more than one occurrence is essentially equal to zero (see below).
3. The numbers of occurrences in two non-overlapping intervals are statistically independent.

More precise versions of condition 2 are: (1) the probability of more than one event occurring in a very short time interval is negligible in comparison to the probability that only one event occurs (Meyer 1970), (2) the probability of more than one event occurring in a very short time interval goes to zero faster than the length of the interval (Pfeiffer and Schum 1973), and (3) simultaneous events occur only with probability zero (Çinlar 1975). All of these versions have the practical interpretation that common cause events do not occur. The phrase "do not occur" is used in this handbook, as it is in Thompson (1981).

The Poisson distribution also can serve as an **approximation to the binomial distribution**. Poisson random variables can be viewed as resulting from an experiment involving a large number of trials, n, that each have a small probability of occurrence, p, of an event. However, the rare occurrence is offset by the large number of trials. As stated above, the binomial distribution gives the probability that an occurrence will take place exactly x times in n trials. If $p = \mu/n$ (so that p is small for large n), and n is large, the binomial probability that the rare occurrence will take place exactly x times is closely approximated by the Poisson distribution with $\mu = np$. In general, the approximation is good for large n, small p, and moderate μ (say $\mu \leq 20$) [see Derman et al. (1973)].

The Poisson distribution is important because it describes the behavior of many rare event occurrences, regardless of their underlying physical process. It also has many applications to describing the occurrences of system and component failures under steady-state conditions. These applications utilize the relationship between the Poisson and **exponential** (continuous random variable, see Section A.7.4) distributions: the times between successive events follow an exponential distribution.

A.7 Continuous Random Variables

A.7.1 The Uniform Distribution

A **uniform** distribution, also referred to as a rectangular distribution, represents the situation where any value in a specified interval, say $[a, b]$, is equally likely. For a uniform random variable, X, because the outcomes are equally likely, $f(x)$ is equal to a constant. The p.d.f. of a uniform distribution with parameters a and b, denoted uniform(a, b), is

$$f(x) = \frac{1}{b-a}$$

for $a \le x \le b$.

Figure A.8 shows the density of the uniform(a, b) distribution.

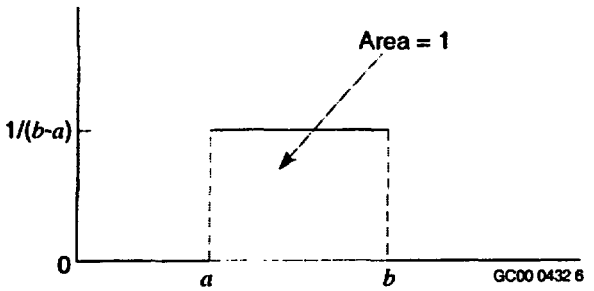

Figure A.8 Density of uniform(a, b) distribution.

The mean and variance of a uniform(a, b) distribution are

$$E(X) = \frac{b+a}{2}$$

and

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

A.7.2 The Normal Distribution

One of the most widely encountered continuous probability distributions is the **normal** distribution, which has the familiar bell shape and is symmetrical about its mean value. The importance of the normal distribution is due to: (1) its applicability in describing a very large number of random variables that occur in nature and (2) the fact that certain useful functions of nonnormal random variables are approximately normal. Details on the derivation of the normal distribution can be found in many basic mathematical statistics textbooks [e.g., Hogg and Craig (1995)].

The normal distribution is characterized by two parameters, μ and σ. For a random variable, X, that is normally distributed with parameters μ and σ, the p.d.f. of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \text{(A.9)}$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, and $\sigma > 0$. Increasing μ moves the density curve to the right and increasing σ spreads the density curve out to the right and left while lowering the peak of the curve. The units of μ and σ are the same as for X.

The mean and variance of a normal distribution with parameters μ and σ are

$$E(X) = \mu$$

and

$$\mathrm{Var}(X) = \sigma^2.$$

The normal distribution is denoted normal(μ, σ^2).

Figure A.9 shows two normal(μ, σ^2) densities. The distribution is largest at μ and is more concentrated around μ when σ is small than when σ is large.

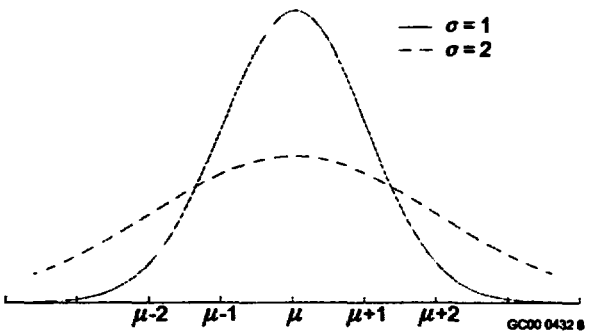

Figure A.9 Two normal densities.

Note the similarity of the normal density to a binomial p.d.f. with large np or a Poisson p.d.f. with large μ. This illustrates the fact that a normal distribution can sometimes be used to approximate those distributions.

The normal(0, 1) distribution is called the **standard normal** distribution, which, from Equation A.9, has p.d.f.

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad \text{(A.10)}$$

for $-\infty < x < \infty$. The cumulative distribution of the standard normal distribution is denoted by Φ. Tables for the standard normal distribution are presented in Appendix C and in almost all books on statistics.

It can be shown that the transformed random variable $Z = (X - \mu)/\sigma$ is normal(0, 1). Thus, to calculate probabilities for a normal(μ, σ^2) random variable, X, when $\mu \neq 0$ and/or $\sigma^2 \neq 1$, the tables for the standard normal can be used. Specifically, for any number a,

$$\Pr[X \le a] = \Pr[(X-\mu)/\sigma \le (a-\mu)/\sigma]$$

$$= \Pr[Z \le (a-\mu)/\sigma]$$

$$= \Phi[(a-\mu)/\sigma].$$

Part of the importance of the normal distribution is that it is the distribution that sample sums and sample means tend to possess as n becomes sufficiently large. This result is known as the **central limit theorem**, which states that, if $X_1, X_2, \ldots, X_n$, are independent random variables, each with mean μ and variance σ^2, the sum of these n random variables, $\Sigma_i X_i$, tends toward a normal($n\mu, n\sigma^2$) distribution for large enough n. Since the sample mean is a linear combination of this sum, the central limit theorem also applies. Thus, $\overline{X} = \Sigma_i X_i/n$ tends to a normal($\mu, \sigma^2/n$) distribution. The importance of the central limit theorem is it can be used to provide approximate probability information for the sample sums and sample means of random variables whose distributions are unknown. Further, because many natural phenomena consist of a sum of several random contributors, the normal distribution is used in many broad applications.

Because a binomial random variable is a sum, it tends to the normal distribution as n gets large. Thus, the normal distribution can be used as an **approximation to the binomial distribution**. One rule of thumb is that the approximation is adequate for $np \ge 5$.

A Poisson random variable also represents a sum and, as presented previously, can also be used as an approximation to the binomial distribution. It follows that the normal distribution can serve as an **approximation to the Poisson distribution** when $\mu = \lambda t$ is large. One rule of thumb is that the approximation is adequate for $\mu \ge 5$.

A.7.3 The Lognormal Distribution

Use of the **lognormal distribution** has become increasingly widespread. It is commonly used as a distribution for failure time and in maintainability analysis (Martz and Waller 1991). It has also been widely used as a prior distribution for unknown positive parameters.

The lognormal distribution arises from the *product* of many independent random variables. If $Y = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n = \Pi_i Y_i$ is the product of n independent positive random variables that are (nearly) identically distributed, then $\ln(Y) = \ln(\Pi_i Y_i) = \Sigma_i \ln(Y_i)$ is a sum that tends toward a normal distribution.

The distribution of Y is defined to be lognormal when the distribution of $\ln(Y)$ is normal. That is, when Y is lognormal, $\ln(Y)$ is normal(μ, σ^2). The parameters of the lognormal distribution are μ and σ, the parameters from the underlying normal distribution. For a random variable, Y, that is lognormally distributed with parameters μ and σ, denoted lognormal(μ, σ^2), the p.d.f. of Y is

$$f(y) = \frac{1}{\sigma y \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\ln y - \mu)^2\right]$$

for $0 < y < \infty$, $-\infty < \mu < \infty$, and $\sigma > 0$. Note the y in the denominator, for reasons explained in Section A.4.7. The mean and variance of a lognormal(μ, σ^2) distribution are

$$E(Y) = \exp(\mu + \sigma^2/2)$$

and

$$\mathrm{Var}(Y) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1].$$

In addition, the median of a lognormal distribution is $\exp(\mu)$ and the mode is $\exp(\mu - \sigma^2)$. See Martz and Waller (1991) for more information on the lognormal distribution.

Sometimes the median of $Y = \exp(\mu)$ is used as a parameter. In addition, a parameter commonly used in PRA is the **error factor** (EF), where $EF = \exp(1.645\sigma)$. This definition causes EF to satisfy

$$\Pr[\mathrm{med}(Y)/EF \le Y \le \mathrm{med}(Y)*EF] = 0.90.$$

Figure A.10 shows three lognormal densities. The value $\mu = -7$ corresponds to a median of about 1.E-3. [More exactly, it corresponds to $\exp(-7)$ = 9.E-4.] The value $\mu = -6.5$ corresponds to a median of about 1.5E-3. The value $\sigma = 0.67$ corresponds to an error factor EF = 3, and $\sigma = 1.4$ corresponds to an error factor EF = 10.

The two distributions with $\sigma = 0.67$ and different values of μ have essentially the same shape, but with different scales. The larger μ corresponds to spreading the distribution out more from zero. The distribution with $\sigma = 1.4$, and therefore EF = 10, has a very skewed distribution.

To calculate probabilities for a lognormal(μ, σ^2) random variable, Y, the tables for the standard normal can be used. Specifically, for any number b,

$$\Pr[\, Y \le b \,] = \Pr[\, \ln(Y) \le \ln(b) \,]$$

$$= \Pr[\, X \le \ln(b) \,]$$

$$= \Phi[\, (\ln(b) - \mu)/\sigma \,],$$

where $X = \ln(Y)$ is normal(μ, σ^2).

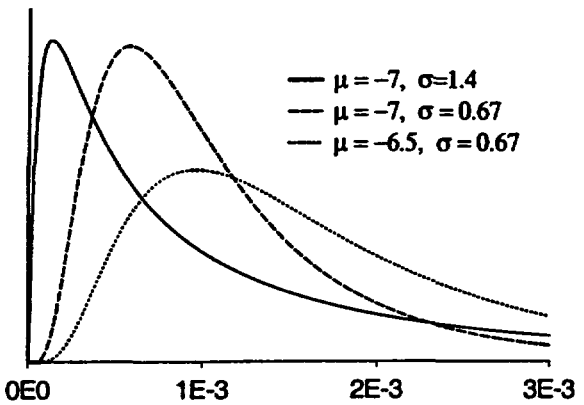

Figure A.10 Three lognormal densities.

A.7.4 The Exponential Distribution

The **exponential distribution** is widely used for modeling time to failure and is inherently associated with the Poisson process [see Martz and Waller (1991)]. For a Poisson random variable X defining the number of failures in a time interval t and for a random variable T defining the time to failure, it can be shown that T has the exponential p.d.f.

$$f(t) = \lambda e^{-\lambda t},$$

for $t > 0$. Thus, the time to first failure and the times between successive failures follow an exponential distribution and the number of failures in a fixed time interval follows a Poisson distribution.

Figure A.11 shows two exponential densities, for two values of λ. The intercept (height of the curve when $t = 0$) equals λ. Thus, the figure shows that the distribution is more concentrated near zero if λ is large. This agrees with the interpretation of λ as a frequency of failures and t as time to first failure.

The exponential distribution parameter, λ, corresponds to the λt parameterization of the Poisson p.d.f. in Equation A.8 and is referred to as the **failure rate** if the component or system is repaired and restarted immediately after each failure. It is called the **hazard rate** if the component or system can only fail once and cannot be repaired. Section A.4.4.2 discusses modeling duration times with different distributions and defines the hazard rate as $h(t) = f(t)/[1 - F(t)]$. For the exponential distribution, the hazard rate is constant, λ. It can be shown that the exponential distribution is the *only* distribution with a constant hazard rate.

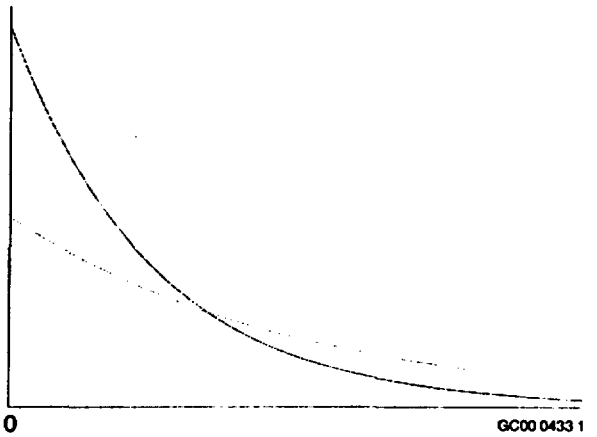

Figure A.11 Two exponential densities.

The c.d.f. of the exponential distribution is

$$F(t) = 1 - e^{-\lambda t}.$$

The exponential distribution with parameter λ is denoted exponential(λ). The mean and variance of an exponential(λ) distribution are

$$E(T) = 1/\lambda$$

and

$$\mathrm{Var}(T) = 1/\lambda^2.$$

The relationship of the exponential distribution to the Poisson process can be seen by observing that the probability of no failures before time t can be viewed in two ways. First, the number of failures, X, can be counted. The probability that the count is equal to 0 is given by Equation A.8 as

$$\Pr(X = 0) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Alternatively, the probability that first failure time, T, is greater than t is

$$\begin{aligned}\Pr(T > t) &= 1 - \Pr(T \le t)\\ &= 1 - F(t)\\ &= 1 - [1 - e^{-\lambda t}]\\ &= e^{-\lambda t}.\end{aligned}$$

Thus, the two approaches give the same expression for the probability of no failures before time t.

The assumptions of a Poisson process require a constant failure rate, λ, which can be interpreted to mean that the failure process has no memory (Martz and Waller 1991). Thus, if a device is still functioning at time t, it remains as good as new and its remaining life has the same exponential(λ) distribution. This constant failure rate corresponds to the flat part of the common **bathtub** curve (frequency of failures plotted against time) and does not pertain to initial (burn-in) failures and wear-out failures.

A different, sometimes useful, parameterization uses $\mu = 1/\lambda = E(T)$. For example, if T represents a time to failure, μ is called the mean time to failure. If T is the time to repair, or to fire suppression, or to some other event, the name for μ is the mean time to repair, or other appropriate name. The exponential(μ) distribution for T has density

$f(t) = (1/\mu)\exp(-t/\mu)$, for $t \geq 0$

and c.d.f.

$F(t) = 1 - \exp(-t/\mu)$, for $t \geq 0$.

The units of μ are the same as the units of t, minutes or hours or whatever the data have. The mean and variance are

$E(T) = \mu$
$\text{var}(T) = \mu^2$.

A.7.5 The Weibull Distribution

The **Weibull distribution** is widely used in reliability and PRA and generalizes the exponential distribution to include nonconstant failure or hazard rates (Martz and Waller 1991). Different Weibull distributions have been successfully used to describe initial failures and wear-out failures. The Weibull distribution is appropriate when a system is composed of a number of components, and system failure is due to any one of the components failing. It, therefore, is commonly referred to as a distribution corresponding to failure of the weakest link.

For a random variable, T, that has a Weibull distribution, the p.d.f. is

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t-\theta}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{t-\theta}{\alpha}\right)^{\beta}\right],$$

for $t \geq \theta \geq 0$ and parameters $\alpha > 0$ and $\beta > 0$. The parameter θ is a location parameter and corresponds to a period of guaranteed life that is not present in many applications (Martz and Waller 1991). Thus, θ is usually set to zero. The c.d.f. for T is

$$F(t) = 1 - \exp\left[-\left(\frac{t-\theta}{\alpha}\right)^{\beta}\right],$$

for $t \geq \theta$ and $\alpha > 0$ and $\beta > 0$.

The α parameter is a scale parameter that expands or contracts the density along the horizontal axis. The β parameter is a shape parameter that allows for a wide variety of distribution shapes. [See Martz and Waller (1991) for further discussion and examples.] When $\beta = 1$, the distribution reduces to the exponential distribution. Therefore, the Weibull family of distributions includes the exponential family of distributions as a special case.

A Weibull distribution with parameters α, β, and θ is referred to as Weibull(α, β, θ) and, when $\theta = 0$, Weibull(α, β). The mean and variance of the Weibull distribution are given by Martz and Waller (1991) as

$\theta + \alpha\Gamma(1 + 1/\beta)$

and

$\alpha^2[\Gamma(1 + 2/\beta) - \Gamma^2(1 + 1/\beta)$.

Here, Γ is the gamma function, defined in Section A.7.6.

Figure A.12 shows four Weibull densities, all with the same scale parameter, α, and all with location parameter $\theta = 0$. The shape parameter, β, varies. When $\beta < 1$, the density becomes infinite at the origin. When $\beta = 1$, the distribution is identical to the exponential distribution. Surprisingly, the distribution is *not* asymptotically normal as β becomes large, although it is approximately normal when β is near 3.

A.7.6 The Gamma and Chi-Squared Distributions

The **gamma distribution** is an extension of the exponential distribution and is sometimes used as a failure time model (Martz and Waller 1991). It is also often used as a prior distribution in Bayesian estimation (see Appendix B) of the failure rate parameter λ from Poisson(λt) or exponential(λ) data. The **chi-squared distribution** is a re-expression of a special case of the gamma distribution.

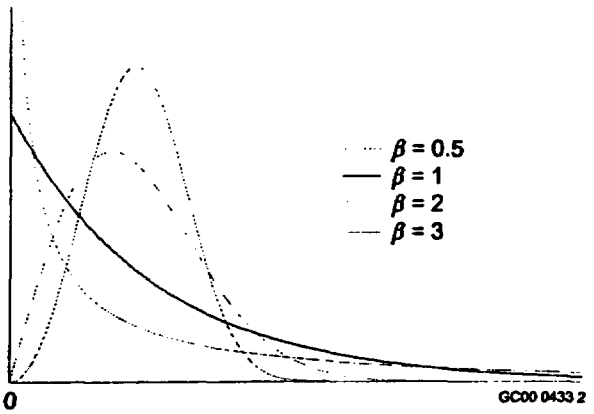

Figure A.12 Four Weibull densities, all having $\theta = 0$ and all having the same α.

The gamma distribution arises in many ways. The distribution of the sum of independent exponential(λ) random variables is gamma, which forms the basis for a confidence interval for λ from exponential(λ) data. Because the sum of n independent exponentially distributed random variables has a gamma distribution, the gamma distribution is often used as the distribution of the time, or waiting time, to the nth event in a Poisson process. In addition, the chi-squared distribution is the distribution for a sum of squares of independent, identically distributed normal random variables, which forms the basis for a confidence interval for the variance of a normal distribution. The gamma distribution is also often used as a distribution for a positive random variable, similar to the lognormal and Weibull distributions. In PRA work, it is often used as a Bayesian distribution for an uncertain positive parameter.

Two parameterizations of the gamma distribution are common, with various letters used for the parameters. The parameterization given here is most useful for Bayesian updates, the primary use of the gamma distribution in this handbook. For a random variable, T, that has a gamma distribution, the p.d.f. is

$$f(t) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \exp(-t\beta),$$

for t, α, and $\beta > 0$.

Here

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

is the **gamma function** evaluated at α. If α is a positive integer, $\Gamma(\alpha) = (\alpha - 1)!$.

A gamma distribution with parameters α and β is referred to as gamma(α, β). The mean and variance of the gamma(α, β) random variable, T, are:

$$E(T) = \alpha/\beta$$

and

$$\mathrm{Var}(T) = \alpha/\beta^2.$$

The parameters α and β are referred to as the shape and scale parameters. The shape parameter α allows the density to have many forms. If α is near zero, the distribution is highly skewed. For $\alpha = 1$, the gamma distribution reduces to an exponential(β^{-1}) distribution. Also, the gamma($\alpha = n/2$, $\beta = ½$) distribution is known as the **chi-squared distribution** with n degrees of freedom, denoted $\chi^2(n)$. The p.d.f. for the $\chi^2(n)$ distribution is found by substituting these values into the above formula for the gamma p.d.f. It also can be found in many statistics texts [e.g., Hogg and Craig (1995, Chapter 4)].

In addition, if T has a gamma(α, β) distribution, then $2\beta T$ has a $\chi^2(2\alpha)$ distribution, which forms the defining relationship between the two distributions. The gamma and chi-squared distributions can, therefore, be viewed as two ways of expressing one distribution. Since the chi-squared distribution usually is only allowed to have integer degrees of freedom, the gamma distribution can be thought of as an interpolation of the chi-squared distribution.

Percentiles of the chi-squared distribution are tabulated in Appendix C. These tables can be used as follows to find the percentiles of any gamma distribution. The $100{\times}p$ percentile of a gamma(α, β) distribution is $\chi^2_p(2\alpha)/(2\beta)$, where $\chi^2_p(2\alpha)$ denotes the $100{\times}p$ percentile of the chi-squared distribution with 2α degrees of freedom.

Figure A.13 shows gamma densities with four shape parameters, α. When $\alpha < 1$, the density becomes infinite at 0. When $\alpha = 1$, the density is identical to an exponential density. When α is large, the distribution is approximately a normal distribution.

As stated previously, the sum of exponential lifetimes or waiting times has a gamma distribution, with the shape parameter α equal to the number of exponential lifetimes. Also, it has been stated that in general the sum of independent, identically distributed random variables is approximately normal. This is the reason why the gamma distribution is approximately normal when α is large.

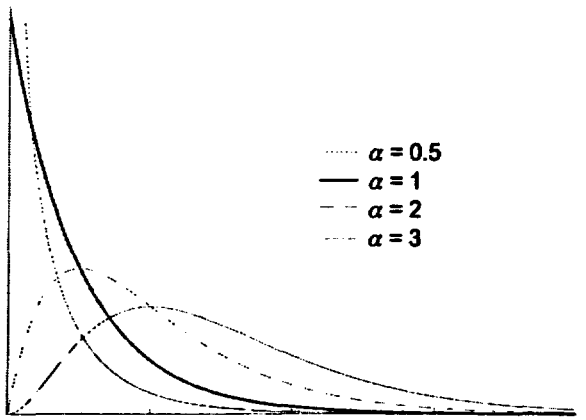

Figure A.13 Gamma densities with four shape parameters.

An alternative parameterization of the gamma distribution uses the scale parameter, say $\tau = \beta^{-1}$. If T has a gamma(α, τ) distribution, its p.d.f. is

$$f(t) = \frac{1}{\tau^{\alpha}\Gamma(\alpha)} t^{\alpha-1}\exp(-t/\tau)$$

for t, α, and $\tau > 0$. The mean and variance of the gamma(α, τ) random variable, T, are:

$$E(T) = \alpha\tau$$

and

$$\mathrm{Var}(T) = \alpha\tau^2.$$

This alternative parameterization is useful in a very small portion of this handbook.

A.7.7 The Inverted Gamma and Inverted Chi-Squared Distributions

The **inverted gamma distribution** is often used as a prior distribution for Bayesian estimation of the mean of an exponential distribution (Martz and Waller 1991). It is also used as a prior and posterior distribution for σ^2 when the data have a normal distribution with variance σ^2 (Box and Tiao 1973, Lee 1997).

For a gamma(α, β) random variable, T, $W = 1/T$ has an inverted gamma distribution with p.d.f.

$$f(w) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\left(\frac{1}{w}\right)^{\alpha+1}\exp\left(-\frac{\beta}{w}\right),$$

for w, α, and $\beta > 0$. The parameters here are the same as for the gamma distribution. For example, if T has units of time then w and β both have units 1/time. A comparison of this density with the gamma density shows that this density has an extra w^2 in the denominator, for reasons explained in Section A.4.7.

The parameters of the inverted gamma distribution are α and β and this distribution is denoted inverted gamma(α, β). Just as with the gamma(α, β) distribution, α is the shape parameter and β is the scale parameter. The distribution can also be parameterized in terms of $\tau = \beta^{-1}$.

The mean and variance of an inverted gamma(α, β) random variable, W, are

$$E(W) = \frac{\beta}{\alpha - 1}, \qquad \alpha > 1,$$

and

$$\mathrm{Var}(W) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \qquad \alpha > 2.$$

Note that for $\alpha \le 1$ the mean and higher moments do not exist. For $1 < \alpha \le 2$ the mean exists but the variance does not exist (Martz and Waller 1991).

Figure A.14 shows four inverted gamma distributions, all having the same scale parameter, β, and having various shape parameters, α.

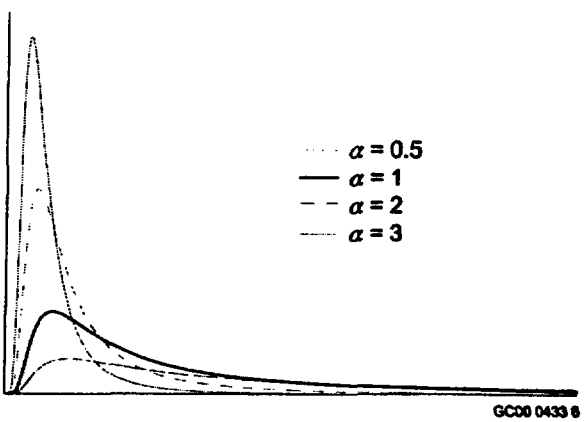

Figure A.14 Four inverted gamma densities, having the same scale parameter, β, and various shape parameters, α.

In the special case with $\alpha = n/2$ and $\beta = ½$, the distribution is called the **inverted chi-squared** distribution with n degrees of freedom. Values from this distribution are sometimes denoted $\chi^{-2}(n)$. This form of the distribution is often used in connection with a prior for σ^2 when the data are normally distributed.

A.7.8 The Beta Distribution

Many continuous quantitative phenomena take on values that are bounded by known numbers a and b. Examples are percentages, proportions, ratios, and distance to failure points on items under stress. The **beta distribution** is a versatile family of distributions that is useful for modeling phenomena that can range from 0 to 1 and, through a transformation, from a to b.

The beta distribution family includes the uniform distribution as well as density shapes that range from decreasing to uni-modal right-skewed to symmetric to U-shaped to uni-modal left-skewed to increasing (Martz and Waller 1991). It can serve as a model for a reliability variable that represents the probability that a system or component lasts at least t units of time. The beta distribution is also widely used in Bayesian estimation and reliability analysis as a prior distribution for the binomial distribution parameter p that represents a reliability or failure probability.

The p.d.f. of a beta random variable, Y, is

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

for $0 \le y \le 1$, with the parameters $\alpha, \beta > 0$. The distribution is denoted beta(α, β). The gamma functions at the front of the p.d.f. form a normalizing constant so that the density integrates to 1.

The mean and variance of the beta(α, β) random variable, Y, are

$$E(Y) = \frac{\alpha}{\alpha+\beta}$$

and

$$\mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Various beta distributions are shown in Figures A.15 and A.16. Figure A.15 shows beta densities with $\alpha = \beta$, and therefore with mean 0.5. When $\alpha < 1$, the density becomes infinite at 0.0, and when $\beta < 1$, the density becomes infinite at 1.0. When $\alpha = \beta = 1$, the density is uniform. When α and β are large, the density is approximately normal.

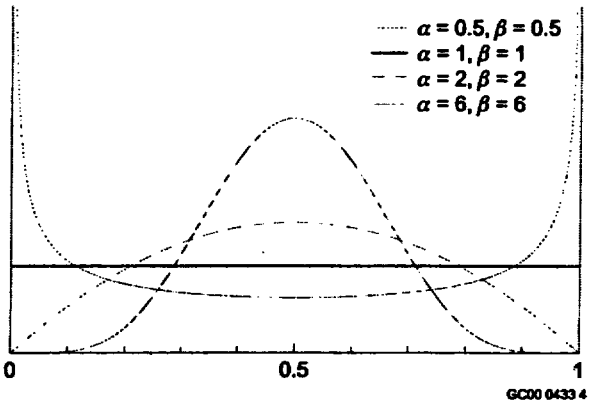

Figure A.15 Beta distributions with mean = 0.5.

Figure A.16 shows densities with mean 0.1. Again, when $\alpha < 1$, the density becomes infinite at 0.0, and when $\alpha > 1$, the density is zero at 0.0. As the parameters α and β become large, the density approaches a normal distribution.

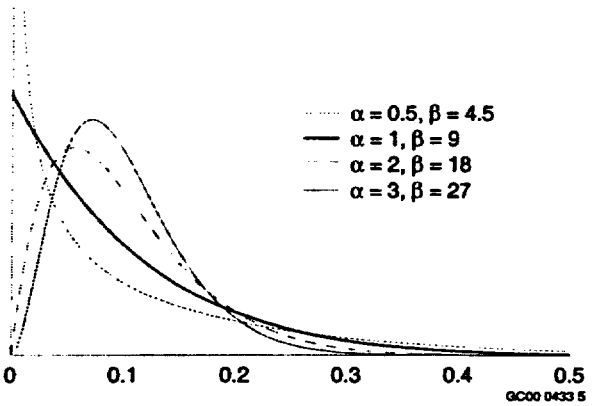

Figure A.16 Four beta distributions with mean 0.1.

Another parameterization of the beta distribution uses the parameters $x_0 = \alpha$ and $n_0 = \alpha + \beta$. This parameterization is used by Martz and Waller (1991) because it simplifies Bayes formulas and Bayesian estimation. The p.d.f. of a beta(x_0, n_0) is

$$f(y) = \frac{\Gamma(n_0)}{\Gamma(x_0)\Gamma(n_0 - x_0)} y^{x_0-1}(1-y)^{n_0-x_0-1},$$

for $0 \le y \le 1$, with the parameters x_0 and n_0 satisfying

$n_0 > x_0 > 0$.

The mean and variance of the beta(x_0, n_0) random variable, Y, are

$$E(Y) = \frac{x_0}{n_0}$$

and

$$\mathrm{Var}(Y) = \frac{x_0(n_0 - x_0)}{n_0^2(n_0 + 1)}.$$

Percentiles of the beta distribution occur in the formula for a confidence interval for p, and in the formula for a Bayes credible interval for p when a conjugate prior is used. Some percentiles are tabulated in Appendix C. In addition, many software packages, including some commonly used spreadsheets, can calculate these percentiles. If none of these work, Martz and Waller (1991) give a method for finding the beta percentiles from the corresponding percentiles of an F distribution, discussed in Section A.7.11. The F distribution is tabulated in most statistics books, and can be interpolated if necessary with good accuracy. The relation is

$$\mathrm{beta}_q(\alpha, \beta) = \alpha / [\alpha + \beta F_{1-q}(2\beta, 2\alpha)]$$

for small q, and

$$\mathrm{beta}_q(\alpha, \beta) = \alpha F_q(2\alpha, 2\beta) / [\beta + \alpha F_q(2\alpha, 2\beta)]$$

for large q. Here $\mathrm{beta}_q(\alpha, \beta)$ denotes the q quantile, or the $100{\times}q$ percentile, of the beta(α, β) distribution, and $F_q(d_1, d_2)$ denotes the q quantile of an F distribution with d_1 and d_2 degrees of freedom. So if all else fails, and a statistics book with F tables is nearby, the first formula can be used to find the lower percentile of the beta distribution and the second formula can be used to find the upper percentile. This method is not discussed further here, because it is not expected to be needed often.

A.7.9 The Logistic-Normal Distribution

While not widely used in PRA, this distribution is commonly used for Bayesian inference in other fields of application, especially as a prior for the binomial parameter p when p could plausibly be fairly large. X has a logistic-normal distribution if $\ln[X/(1 - X)]$ is normally distributed with some mean μ and variance σ^2. The function $\ln[X/(1 - X)]$ may appear strange, but it is common enough in some areas of application to have a name, the logit function. Therefore, the above statements could be rewritten to say that X has a logistic-normal distribution if logit(X) is normally distributed.

Properties of the logistic-normal distribution are summarized here.

- Let $y = \ln[x/(1 - x)]$. Then $x = e^y / (1 + e^y)$. This implies that x must be between 0 and 1.
- As x increases from 0 to 1, $y = \ln[x/(1 - x)]$ increases monotonically from $-\infty$ to $+\infty$. Thus, y can be generated from a normal distribution with no problem of forcing x outside its possible range.
- The monotonic relation between x and y means that the percentiles match. For example, the 95th percentile of Y is $\mu + 1.645\sigma$. Denote this by $y_{0.95}$. Therefore, the 95th percentile of X is $x_{0.95} = \exp(y_{0.95}) / [1 + \exp(y_{0.95})]$. Alternatively, this can be written as $y_{0.95} = \ln[x_{0.95} / (1 - x_{0.95})]$.
- If X is close to 0 with high probability, so that $X/(1 - X)$ is close to X with high probability, then the logistic-normal and lognormal distributions are nearly the same.

The third bullet shows how to find the percentiles of a logistic-normal distribution. Unfortunately there is no equally easy way to find the moments, such as the mean or variance. Moments must be found using numerical integration.

Figure A.17 shows several logistic normal distributions that all have median 0.5. These correspond to a normally distributed y with mean $\mu = 0$ and with various values of σ. Figure A.18 shows several logistic normal distributions that all have median 0.1. These correspond to a normally distributed y with mean $\mu = -2.2 = \ln[0.1/(1 - 0.1)]$.

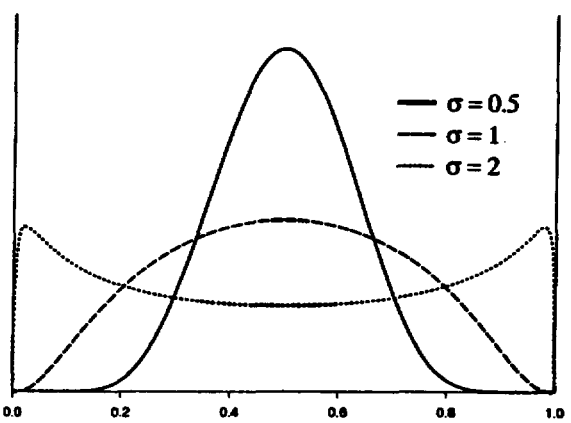

Figure A.17 Three logistic-normal densities with median = 0.5.

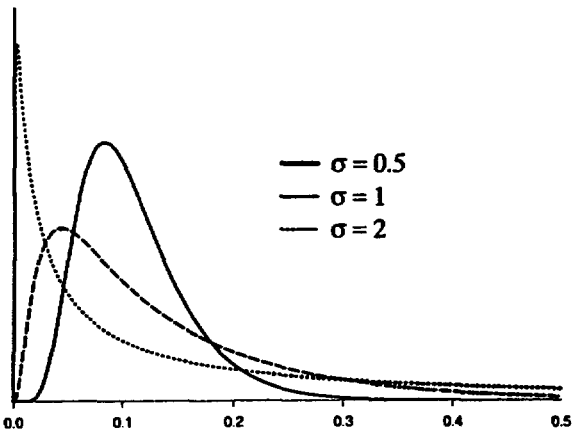

Figure A.18 Three logistic-normal densities with median = 0.1.

Note the general similarities to the beta distributions in Figures A.15 and A.16. Note also the differences: Logistic-normal distributions are characterized most easily by percentiles, whereas beta distributions are characterized most easily by moments. Also, the beta densities can be J-shaped or U-shaped, but the logistic-normal densities always drop to zero at the ends of the range.

A.7.10 Student's *t* Distribution

The Student's *t* distribution is not used in a central way in PRA. However, it appears in a peripheral way in places in this handbook, when dealing with the parameters of a normal or lognormal distribution, or in large-sample situations when a distribution is approximated as normal or lognormal. Therefore, the basic facts are summarized here.

If (1) Z has a standard normal distribution, (2) X has a chi-squared distribution with d degrees of freedom, and (3) Z and X are statistically independent, then

$$T = \frac{Z}{\sqrt{X/d}}$$

has a Student's *t* distribution with d degrees of freedom. Therefore, T has a distribution that is symmetrical about 0, and it can take values in the entire real line. If d is large, the denominator is close to 1 with high probability, and T has approximately a standard normal distribution. If d is smaller, the denominator adds extra variability, and the extreme percentiles of T are farther out than are the corresponding normal percentiles. Tables of the distribution are given in Appendix C.

Although not needed for ordinary work, the p.d.f. and first two moments of T are given here. [See many standard texts, such DeGroot (1975) or Bain and Engelhartd (1992).] The p.d.f. is

$$f(t) = \frac{\Gamma[(d+1)/2]}{(d\pi)^{1/2}\Gamma(d/2)}\left[1+(t^2/2)\right]^{-(d+1)/2} .$$

If $d > 1$ the mean is 0. If $d > 2$ the variance is $d/(d-2)$. If $d \le 2$ the variance does not exist. If $d = 1$, even the mean does not exist; in this case the distribution is called a **Cauchy distribution.**

A.7.11 *F* Distribution

The F distribution, also called Snedecor's F distribution, arises as follows. If Y and Z are independent chi-squared random variables with m and n degrees of freedom, respectively, then

$$X = \frac{Y/m}{Z/n}$$

has an F distribution with m and n degrees of freedom. This is sometimes written as an $F(m, n)$ distribution. This can be re-expressed in terms of a ratio of gamma-distributed variables, because the chi-squared distribution is a special case of a gamma distribution.

The density of an F distribution is almost never needed, although it is given in mathematical statistics books as

$$f(x) = \frac{\Gamma[(m+n)/2]m^{m/2}n^{n/2}}{\Gamma(m/2)\Gamma(n/2)}\frac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}}$$

for $x \ge 0$. Bain and Engelhardt (1992) give the moments:

$$E(X) = n/(n-2)$$

$$\mathrm{Var}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} .$$

The mean is defined only if $n > 2$, and the variance only if $n > 4$.

It follows from the definition in terms of a ratio of chi-squared variables that the percentiles are related to each other as follows. If $F_q(m, n)$ is the q quantile (that is, the $100q$ percentile) of an $F(m, n)$ distribution, then

$$F_q(m, n) = 1/F_{1-q}(n, m) . \qquad (A.11)$$

The F distribution is also related to the beta distribution, and Equation (A.11) forms the basis for the two different forms of the relation given near the end of Section A.7.8.

The distribution is not tabulated in Appendix C for two reasons: the distribution is used only minimally for the applications in this handbook, and the percentiles and probabilities are given by many commonly used software packages.

A.7.12 Dirichlet Distribution

The Dirichlet distribution is a multivariate generalization of the beta distribution. Let m variables $Y_1, ..., Y_m$ be such that $\Sigma_i Y_i = 1$. Their distribution can be described in terms of any $m - 1$ of them, such as $Y_1, ..., Y_{m-1}$ with

$$Y_m = 1 - \sum_{i=1}^{m-1} Y_i \ .$$

The m variables have a Dirichlet distribution with parameters $\alpha_1, ..., \alpha_m$ if the joint density of the first $m - 1$ variables is

$$f(y_1,\ldots,y_{m-1}) = \frac{\Gamma(\alpha_1+\ldots+\alpha_m)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_m)} \times y_1^{\alpha_1-1}\cdots y_{m-1}^{\alpha_{m-1}-1}\left(1-\sum_{i=1}^{m-1} y_i\right)^{\alpha_m-1}$$

Observe that when $m = 2$ this reduces to a beta distribution for Y_1 with parameters α_1 and α_2. (Some authors say that $Y_1, ..., Y_m$ have the Dirichlet distribution, while others say that $Y_1, ..., Y_{m-1}$ have this distribution. The distribution is the same whichever way it is described.)

Many of the properties of the distribution are described most easily in terms of an additional parameter α, defined as $\alpha = \alpha_1 + ... + \alpha_m$. Some of these properties are the following.

Individually, each Y_i has a beta(α_i, $\alpha - \alpha_i$) distribution. Therefore, we have

$E(Y_i) = \alpha_i / \alpha$, and

$\text{Var}(Y_i) = \alpha_i(\alpha - \alpha_i)/[\alpha^2(\alpha + 1)]$.

It can also be shown that the covariance terms are given by

$\text{Cov}(Y_i, Y_j) = -\alpha_i\alpha_j/[\alpha^2(\alpha + 1)]$.

Thus, the ratio of each α_i to α determines the corresponding mean. Once the means are fixed, the magnitude of α determines the variances and covariances, with large α corresponding to small variances. The covariances are negative, meaning that if one variable is larger than its mean, each other variable tends to be smaller than its mean; this is not surprising for variables that must sum to 1.

One application of the Dirichlet distribution in PRA is to multiple-branch nodes in event trees. If an event tree has a node with m branches, $m > 2$, the probability of the ith branch (also called the ith "split fraction") can be denoted p_i. The probabilities must satisfy $p_1 + ... + p_m = 1$. They are not known exactly, and therefore are assigned a joint distribution that describes their uncertainty in a Bayesian way. The Dirichlet distribution is a natural distribution to use.

For further information about this distribution, see the article in the *Encyclopedia of Statistical Sciences*, or Kotz, Balakrishnan, and Johnson (2000).

REFERENCES FOR APPENDIX A

Ascher, H., and H. Feingold, 1984, *Repairable Systems Reliability – Modeling, Inference, Misconceptions and Their Causes*, Marcel Dekker, Inc., New York.

Bain, L. J., and M. Engelhardt, 1992, *Introduction to Probability and Mathematical Statistics, 2nd Ed.*, PWS-Kent Pub., Boston.

Box, G. E. P., and G. C. Tiao, 1973, *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

Çinlar, E., 1975, *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs, NJ.

DeGroot, Morris H., 1975, *Probability and Statistics*, Addison-Wesley, Reading, MA.

Derman, C., L. J. Gleser, and I. Olkin, 1973, *A Guide to Probability Theory and Application*. Holt, Rinehart and Winston, NY.

Encyclopedia of Statistical Sciences, Vol. 2, 1982, S. Kotz and N. L. Johnson, eds., John Wiley and Sons, NY, pp. 386-387.

Esary, J. D., and F. Proschan, 1963, Coherent Structures of Non-Identical Components, in *Technometrics*, Vol. 6, pp. 191-209.

Esary, J. D., and F. Proschan, 1970, A Reliability Bound for Systems of Maintained, Interdependent Components in *Journal of the American Statistical Association*, Vol. 65, pp. 329-338.

Hogg, R. V., and A. T. Craig, 1995, *Introduction to Mathematical Statistics, 5th Edition*. Macmillan, NY.

Kotz, S., N. Balakrishnan, and N. L. Johnson, 2000, *Continuous Multivariate Distributions, Volume 1, Models and Applications, 2nd Edition*, Chapter 49. Wiley-Interscience, NY.

Lee, Peter M., 1997, *Bayesian Statistics: An Introduction, Second Edition*, Arnold, a member of the Hodder Headline Group, London.

Martz, H. F., and R. A. Waller, 1991, *Bayesian Reliability Analysis*. R. E. Krieger Publishing Co., Malabar, FL.

Meyer, P. L., 1970, *Introductory Probability and Statistical Applications*. Addison-Wesley Pub. Co., Reading, MA.

NRC, 1994, *A Review of NRC Staff Uses of Probabilistic Risk Assessment*, NUREG-1489, U. S. Nuclear Regulatory Commission, Washington, D.C.

Pfeiffer, P. E., and D. A. Schum, 1973, *Introduction to Applied Probability*. Academic Press, NY.

Snedecor, G. W., and W. G. Cochran, 1989, *Statistical Methods*. Iowa State University Press, Ames, IA.

Thompson, W. A., Jr., 1981, On the Foundations of Reliability, in *Technometrics*, Vol. 23, No. 1, pp. 1-13.

Vesely, W. E., F. F. Goldberg, N. H. Roberts, and D. F. Haasl, 1981, *Fault Tree Handbook*. NUREG-0492. U.S. Nuclear Regulatory Commission, Washington, D.C.

B. BASICS OF STATISTICS

B.1 Random Samples

When sampling from a distribution (or population), it is usually assumed that the n observations are taken at random, in the following sense. It is assumed that the n random variables $X_1, X_2, ..., X_n$ are independent. That is, the sample $X_1, X_2, ..., X_n$, taken from a distribution $f(x)$, has the joint p.d.f. h satisfying

$$h(x_1, x_2, ..., x_n) = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n).$$

This follows the definition of independent random variables given in Section A.4.5. A sample taken in this way is called a **random sample**. (As elsewhere in this handbook, upper case letters denote random variables and lower case letters denote particular values, number.)

The random variables $X_1, X_2, ..., X_n$ forming such a random sample are referred to as being independent and identically distributed. If n is large enough, the sampled values will represent the distribution well enough to permit inference about the true distribution.

B.2 Sample Moments

Mathematical expectation and moments provide characteristics of distributions of random variables. These ideas can also be used with observations from a random sample from a distribution to provide estimates of the parameters that characterize that distribution.

A **statistic** is a function of one or more random variables that does not depend on any unknown parameters. A function of random variables that can be computed from the collected data sample is thus a statistic. Note that a function of random variables is also a random variable that has its own probability distribution and associated characteristics.

If $X_1, X_2, ..., X_n$ denote a random sample of size n from a distribution $f(x)$, the statistic

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

is the **mean of the random sample**, or the **sample mean** and the statistic

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \qquad \text{(B.1)}$$

is the **variance of the random sample**. Note that $n - 1$ is used as the denominator in the S^2 statistic to make the statistic an *unbiased* estimator of the population variance, σ^2 (unbiased estimators are discussed in Section B.4.1). Some authors use n in the denominator instead of $n - 1$, with corresponding adjustment of formulas that involve S, but this handbook uses Equation B.1 consistently. In applications with computer packages, note which definition is used and make any necessary adjustments to formulas in this handbook.

Although not used as much as the sample mean and sample variance, the **sample skewness** is occasionally of interest. The definition can vary in detail, but one, used by SAS (1988) is

$$\frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (X_i - \overline{X})^3 \Big/ S^3 .$$

Similarly, the statistics defined by

$$m_r = \sum_{i=1}^{n} \frac{X_i^r}{n},$$

for $r = 1, 2, ...$, are called the **sample moments**.

One of the common uses of statistics is estimating the unknown parameters of the distribution from which the sample was generated. The sample mean, or average, $\overline{X}$, is used to estimate the distribution mean, or population mean, μ, the sample variance, S^2, is used to estimate the population variance, σ^2, and so forth.

B.3 Statistical Inference

Since values of the parameters of a distribution are rarely known, the distribution of a random variable is rarely completely known. However, with some assumptions and information based on a random sample of observations from the distribution or population, values of the unknown parameters can often be estimated. Probabilities can then be calculated from the corresponding distribution using these parameter estimates.

Statistical inference is the area of statistics concerned with using sample data to answer questions and make statements about the distribution of a random variable from which the sample data were obtained. Parameter **estimators** are functions of sample data that are used to estimate the distribution parameters. Statements about parameter values are inferred from the specific sample to the general distribution of the random variable or population. This inference cannot be perfect; all inference techniques involve uncertainty. Understanding the performance properties of various estimators has received much attention in the statistics field.

For the purposes of this handbook, statistical inference procedures can be classified as follows:

- parameter estimation
  - estimation by a point value
  - estimation by an interval
- hypothesis testing
  - tests concerning parameter values
  - goodness-of-fit tests and other model-validation tests.

Parametric statistical inference assumes that the sample data come from a particular, specified family of distributions, with only the parameter values unknown. However, not all statistical inference is based on parametric families. In many cases, in addition to not knowing the distribution parameter values, the form of the parametric family of distributions is unknown. **Distribution-free**, also called **nonparametric**, techniques are applicable no matter what form the distribution may have. **Goodness-of-fit tests** are an important type of nonparametric tests that can be used to test whether a data set follows a hypothesized distribution.

For statistical inference, two major approaches exist, the **frequentist** approach and the **Bayesian** approach. The two resulting sets of inference tools are summarized in Sections B.4 and B.5. In PRA work, Bayesian estimators are normally used for parameter estimation. See, for example, NUREG-1489 (NRC 1994). However, frequentist hypothesis tests are often used for model validation, especially when the hypothesis to be tested does not involve a simple parameter. This use of Bayesian techniques for estimation and frequentist techniques for model validation is also recommended by Box (1980).

NUREG-1489 (NRC 1994) lists a number of "advantages" and "disadvantages" for each of the Bayesian and frequentist approaches. An "advantage" is often in the eye of the beholder. For example, is it an advantage or disadvantage that frequentist methods use only the data at hand, not external or prior information? Therefore, the lists from that report are presented in modified and augmented form in Table B.1, where the points are not called advantages or disadvantages, but simply "features."

B.4 Frequentist Inference

Frequentist estimation of distribution parameters uses only the information contained in the data sample and assumptions about a model for the sample data. In contrast to **Bayesian estimation** (discussed in Section B.5), degree of belief is not incorporated into the estimation process of frequentist estimation.

In the **frequentist** approach to estimation, a distribution parameter is treated as an unknown constant and the data to be used for estimation are assumed to have resulted from a random sample. Information outside that contained in the sample data is used minimally. The random variability in the sample data is assumed to be due directly to the process under study. Thus, the frequentist approach addresses variation in parameter estimates and how far estimates are from the true parameter values.

Frequentist testing of a hypothesis follows the same spirit. The hypothesis is assumed, and the data are compared to what would have been expected or predicted by the hypothesis. The frequentist analyst asks whether the observed values come from the likely part of the distribution or from the extreme tails, and decides in this way whether the data are consistent with the hypothesis.

B.4.1 Point Estimation

Many situations arise in statistics where a random variable X has a p.d.f. that is of known functional form but depends on an unknown parameter θ that can take on any value in a set. The different values for θ produce a family of distributions. One member of the family corresponds to each possible value of θ. **Estimators** of the distribution parameter are functions of sample data that are used to estimate the distribution parameters. Thus, estimators are themselves random variables. The specific value of the estimator computed from a random sample provides an **estimate** of the distribution parameter. Note the distinction between *estimator*, a random variable, and *estimate*, a particular value. An estimate of a distribution parameter in the form of a single number is called a **point estimate** of that parameter. The sample mean is a point estimate of

Table B.1 Features of Bayesian and frequentist approaches.

Bayesian Approach	Frequentist Approach
Bayesian methods allow the formal introduction of prior information and knowledge into the analysis, which can be especially useful when sample data are scarce, such as for rare events. For the nuclear industry, this knowledge often exists in the form of **industry-wide generic data**. Thus, Bayesian estimation allows the use of various types of relevant generic data in PRA.	Results depend only on the data sample. Including relevant information about a parameter that is external to the random sample is complicated.
If the prior distribution accurately reflects the uncertainty about a parameter, Bayesian parameter estimates are better than classical estimates.	
Bayesian estimation can be sensitive to the choice of a prior distribution. Therefore: Identifying suitable prior distributions and justifying and gaining acceptance for their use can be difficult. The choice of a prior distribution is open to criticism that the choice is self-serving and may reflect inappropriate, biased, or incorrect views.	
Because Bayesian probability intervals can be interpreted as probability statements about a parameter, they are easily combined with other sources of uncertainty in a PRA using the laws of probability.	A confidence interval cannot be directly interpreted as a probability that the parameter lies in the interval.
Bayesian distributions can be propagated through fault trees, event trees, and other logic models.	It is difficult or impossible to propagate frequentist confidence intervals through fault and event tree models common in PRA to produce corresponding interval estimates on output quantities of interest.
Using Bayes' Theorem, Bayesian estimation provides a method to update the state of knowledge about a parameter as additional data become available.	Frequentist methods can update an earlier analysis if the original data are still available or can be reconstructed.
In complicated settings, Bayesian methods require software to produce samples from the distributions.	In complicated settings, frequentist methods must use approximations. In some cases they may be unable to analyze the data at all.
Bayesian hypothesis tests are commonly used only with hypotheses about a parameter value.	A well-developed body of hypothesis tests exists, useful for model validation. These are appropriate for investigating goodness of fit, poolability of data sources, and similar questions that do not involve a simple parameter.
Both Approaches	
When the quantity of data is large, both approaches produce good estimates.	
Both types of computation are straightforward when estimating a parameter in a simple setting.	

the mean of the distribution and the sample variance is a point estimate of the variance of the distribution. For another sample drawn from the same population, a different sample mean and variance would be calculated. In fact, these sample statistics are specific values of random variables. Thus, viewed as random variables the sample statistics have their own **sampling distributions**. For example, it can be shown that $\bar{X}$ has mean μ and variance σ^2/n, regardless of the distribution from which the samples are drawn.

Different techniques exist for obtaining point estimates for unknown distribution characteristics or parameters. Two of the most common methods are presented here [see Hogg and Craig (1995) for more information]: maximum likelihood estimation and the method of moments.

A distribution of a random variable X that depends on an unknown parameter θ will be denoted $f(x; \theta)$. If X_1, X_2, ..., X_n is a random sample from $f(x; \theta)$, the joint p.d.f. of X_1, X_2, ..., X_n is $f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$. This joint p.d.f. may be viewed as a function of the unknown parameter θ and, when so viewed, is called the **likelihood function**, L, of the random sample. Thus, the likelihood function is the joint p.d.f. of X_1, X_2, ..., X_n, denoted

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta) ,$$

viewed as a function of θ. The **maximum likelihood estimate** of θ is defined as the value $\hat{\theta}$ such that $L(\hat{\theta}; x_1, x_2, \ldots, x_n) \geq L(\theta, x_1, x_2, \ldots, x_n)$ for every value of θ. That is, the maximum likelihood estimate of θ is the value $\hat{\theta}$ that maximizes the likelihood function. In many cases, this maximum will be unique and can often be obtained through differentiation. Note that solving the derivative set to zero for θ may be easier using $\ln(L)$, which is equivalent since a function and its natural logarithm are maximized at the same value of θ.

The maximum likelihood estimate is a function of the observed random sample $x_1, x_2, \ldots, x_n$. When $\hat{\theta}$ is considered to be a function of the random sample X_1, $X_2, \ldots, X_n$, then $\hat{\theta}$ is a random variable and is called the **maximum likelihood estimator** of θ.

Another method of point estimation is the **method of moments**, which involves setting the distribution moments equal to the sample moments:

$$M_r = E(X^r) = m_r = \sum_i x_i^r / n,$$

for $r = 1, 2, \ldots, k$, if the p.d.f. $f(x; \theta_1, \theta_2, \ldots, \theta_k)$ has k parameters. The k equations can be solved for the k unknowns $\theta_1, \theta_2, \ldots, \theta_k$ and the solutions $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ are the method-of-moments estimators.

How "well" a point estimator estimates a parameter has received a large amount of attention. Numerous desirable properties of point estimators exist. One desirable property of estimators, alluded to previously in Section B.2, is **unbiasedness**. An **unbiased** estimator is one whose mean value is equal to the parameter being estimated. That is, an estimator $\hat{\theta}$ is unbiased for a parameter θ if $E(\hat{\theta}) = \theta$. For a random sample from a normal distribution, the sample mean, $\bar{X}$, and the sample variance, S^2, are unbiased estimators of μ and σ^2, respectively. (Recall that S^2 is defined by Equation B.1, with $n - 1$ in the denominator.) However, the method of moments estimator of the variance is biased. The **bias** of an estimator $\hat{\theta}$ is defined as $E(\hat{\theta}) - \theta$.

Minimum variance is another desirable property of an estimator. An unbiased estimator is said to have minimum variance if its variance is less than or equal to the variance of every other unbiased statistic for θ. Such an estimator is referred to as an unbiased, minimum variance estimator.

Another desirable property of estimators is **sufficiency**. For a random sample $X_1, X_2, \ldots, X_n$ from $f(x; \theta_1, \theta_2, \ldots, \theta_m)$, and $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ functions (statistics) of the X_is, the statistics $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ are **jointly sufficient statistics** if the conditional p.d.f. of the X_is given the statistics $\hat{\theta}_j$s, $g(x_1, x_2, \ldots, x_n | \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$, is independent of the parameters (Martz and Waller, 1991).

Sufficiency can be thought of as exhausting all the possible information about a parameter that is contained in the random sample. When a sufficient statistic exists, it may serve as the basis for a minimum variance or "best" estimator of the parameter. Sufficiency is also important because it simplifies Bayesian estimation methods.

Under certain commonly occurring conditions, as the sample size gets large, the maximum likelihood estimator is approximately normally distributed, approximately unbiased, and has approximately the minimum variance. It is, therefore, a very good estimator for large data sets. The maximum likelihood estimator is not necessarily good for small data sets.

Several other methods of estimation and desirable properties for estimators exist. Further information can be found in Hogg and Craig (1995) or Kendall and Stuart (1973).

B.4.2 Interval Estimation

Another way of estimating a parameter is to identify that it falls in some interval (*lcl*, *ucl*) with a specified degree of certainty, or confidence, where *lcl* denotes the lower confidence limit and *ucl* denotes the upper confidence limit. The interval (*lcl*, *ucl*) is referred to as an **interval estimate** of the parameter. The *lcl* and *ucl* values are calculated from the random sample from the given distribution. Associating a level of desired confidence with an interval estimate produces a **confidence interval**. The level of desired confidence is also referred to as the confidence coefficient.

Confidence intervals are based on estimators associated with a random sample (functions of the data), *LCL* for the lower confidence limit and *UCL* for the upper confidence limit, such that, prior to observing the random sample, the probability that the unknown parameter, θ, is contained in the interval [*LCL*, *UCL*] is known. That is,

$$\Pr[LCL \le \theta \le UCL] = 1 - \alpha$$

for $0 < \alpha < 1$.

Once the random sample has been generated, the functions *LCL* and *UCL* produce two values, *lcl* and *ucl*. The interval (*lcl*, *ucl*) is called a two-sided confidence interval with confidence level $1 - \alpha$, or equivalently, a $100(1 - \alpha)\%$ two-sided confidence interval. Similarly, upper one-sided confidence intervals or lower one-sided confidence intervals can be defined that produce only an upper or lower limit, respectively.

Since the true parameter value, although unknown, is some constant, the interval estimate either contains the true parameter value or it does not. A 95% confidence interval is interpreted to mean that, for a large number of random samples from the same distribution, 95% of the resulting intervals (one interval estimate of the same population parameter constructed the same way for each sample) would contain the true population parameter value, and 5% of the intervals would not. The $\alpha = .05$ risk of obtaining an interval that does not contain the parameter can be increased or decreased. Values for $1 - \alpha$ should be decided upon prior to obtaining the random sample, with .99, .95, and .90 being typical. Note that higher confidence levels result in wider interval estimates.

Confidence intervals cannot be interpreted as probability statements about the parameter being estimated, because the parameter is assumed to be an unknown constant and not a random variable. The level of confidence pertains to the percentage of intervals, each calculated from a different random sample from the same distribution, that are expected to contain the true parameter value. The confidence does not pertain to the specific calculated interval (it could be from the unlucky 5% of intervals that do not contain the true parameter value).

As an example, a confidence interval for the parameter μ can be produced from a random sample drawn from a normal(μ, σ^2) population by calculating the appropriate functions of the data. Recall that, if each sample value is drawn from a normal distribution, the sample mean $\bar{X}$ has a normal(μ, σ^2/n) distribution, where n is the sample size. Even if the sample values are drawn from a distribution that is not normal, by the central limit theorem, $\bar{X}$ will be approximately normal(μ, σ^2/n) for sufficiently large n. Assuming that σ^2 is known (from previous data and experience), the standardized normal random variable

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

is normal(0, 1), and tabulated in Appendix C. From these tables, values of w can be found for which

$$\Pr[-w \le Z \le w] = 1 - \alpha. \qquad \text{(B.2)}$$

For example, for $\alpha = .05$, $w = 1.96$. In this case, w is the 97.5th percentile of the standard normal distribution, commonly denoted $z_{0.975}$, or $z_{1-\alpha/2}$ for $\alpha = .05$.

Substituting for Z in Equation B.2 above, along with some algebraic manipulation, produces

$$\Pr[\bar{X} - w\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + w\frac{\sigma}{\sqrt{n}}] = 1 - \alpha,$$

which defines a $100(1 - \alpha)\%$ confidence interval for the population mean μ, where

$$LCL = \bar{X} - w\frac{\sigma}{\sqrt{n}} \qquad \text{(B.3)}$$

and

$$UCL = \bar{X} + w\frac{\sigma}{\sqrt{n}}, \qquad \text{(B.4)}$$

with $w = z_{1-\alpha/2}$.

A random sample will yield a specific sample mean. The numbers w and n are known, and σ was assumed to be known. Therefore, for a preassigned confidence level, values for *LCL* and *UCL* can be calculated to produce a specific $100(1 - \alpha)\%$ confidence interval for μ. Each of the random variables *LCL* and *UCL* is a statistic, and the interval (*LCL*, *UCL*) is a random interval formed from these statistics.

Usually the value of σ is not known. In this case, the unbiased estimator of the population variance, S^2, can be used to produce S, which can be used in the above equations in place of σ. Thus, the following standardized random variable, T, can be formed:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

(This formula requires the definition of S based on Equation B.1.) For sufficiently large n (say 25 or 30), T follows a normal(0, 1) distribution. If n is not sufficiently large, T follows a **Student's *t* distribution**, for which tabulated probabilities exist in many statistics books, and in Appendix C. The Student's t distribution depends on a parameter called the **degrees of freedom**. In the present example, this parameter equals $n - 1$. Confidence intervals for the population mean can then be calculated similarly to the case where σ is known, using either the Student's t distribution or, when n is large, the normal distribution.

Confidence intervals can also be constructed for differences of means and many other population parameters, such as variances, probabilities, quantiles, and distribution characteristics (see, for example, Hogg and Craig 1978).

B.4.3 Hypothesis Testing

Testing a statistical hypothesis is another major area of statistics. A **hypothesis** is a statement about the distribution of the observable random variable. Often this statement is expressed as a statement about one or more parameters of the distribution. As discussed previously, estimation uses information in the data from a random sample to infer something about the magnitude of a parameter value. Similar to estimation, **hypothesis testing** also uses information from the random sample. However, the objective of hypothesis testing is to determine whether the specific statement about the distribution is true.

The hypothesis to be tested is referred to as the **null hypothesis**, denoted by H_0. The alternative to the null hypothesis is referred to as the **alternative hypothesis**, denoted H_1 or H_a. A **test of a hypothesis** is a rule or procedure for deciding whether to reject or accept the null hypothesis. This rule or procedure is based upon information contained in the random sample and produces a single number, called a test statistic, which leads to a decision of whether the sample values do not support H_0. The entire set of values that the test statistic may assume is divided into two regions, one corresponding to the **rejection region** and the other to the **acceptance region**.

If the test statistic computed from a particular sample has a value in the rejection region, H_0 is rejected. If the test statistic falls in the acceptance region, H_0 is said to be accepted, due to lack of evidence to reject. For each of the two possible cases for H_0, true or false, the test either rejects or does not reject H_0, producing four distinct possibilities. These possibilities (using conditional probability notation), along with some concepts and terms associated with hypothesis testing, are summarized in Table B.2 (Martz and Waller, 1991).

Table B.2 Possible hypothesis test outcomes.

	H_0 True	H_0 False
Accept H_0	Pr(accept H_0 | H_0 is true) = $1-\alpha$ = Level of confidence	Pr(accept H_0 | H_0 is false) = β = Pr(Type II Error)
Reject H_0	Pr(reject H_0 | H_0 is true) = α = Level of significance = Pr(Type I Error)	Pr(reject H_0 | H_0 is false) = $1-\beta$ = Power

A stated null hypothesis is either true or false. One of two errors can occur in hypothesis testing:

1. rejection of the null hypothesis when it is true, referred to as the **Type I error**; and
2. acceptance of the null hypothesis when it is false, referred to as the **Type II error**.

The probability of making a Type I error, denoted by α, is referred to as the **significance level of the test.** Thus, $1 - \alpha$ is the probability of making a correct decision when H_0 is true. The probability of making a correct decision when H_0 is false, denoted $1 - \beta$, is referred to as the **power of the test.** The probability of making a Type II error is equal to one minus the power of the test, or β.

The goodness of a statistical hypothesis test is measured by the probabilities of making a Type I or a Type II error. Since α is the probability that the test statistic will fall in the rejection region, assuming H_0 to be true, increasing the size of the rejection region will increase α and simultaneously decrease β for a fixed sample size. Reducing the size of the rejection region will decrease α and increase β. If the sample size, n, is increased, more information will be available for use in making the decision, and both α and β will decrease.

The probability of making a Type II error, β, varies depending on the true value of the population parameter. If the true population parameter is very close to the hypothesized value, a very large sample would be needed to detect such a difference. That is, the probability of accepting H_0 when H_0 is false, β, varies depending on the difference between the true value and the hypothesized value. For hypothesis tests, α is specified prior to conducting the random sample. This fixed α specifies the rejection region. For a deviation from the hypothesized value that is considered practical and that is wished to be detectable by the hypothesis test, a sample size can be selected that will produce an acceptable value of β.

Different alternative hypotheses will result in different rejection regions for the same H_0. This is seen most easily for a hypothesis that is expressed in terms of a parameter, for example, H_0: $\mu = \mu_0$ for some given value μ_0. In this case, there is an exact correspondence between one-sided and two-sided confidence intervals and rejection regions for one-sided and two-sided alternative hypotheses. If the hypothesized value falls outside a $100(1 - \alpha)\%$ confidence interval for the corresponding population parameter, the null hypothesis would be rejected with level of confidence equal to $1 - \alpha$.

For the example presented in the previous section, Section B.4.2, the $100(1 - \alpha)\%$ two-sided confidence interval for a population mean is defined by the *LCL* and *UCL* in Equations B.3 and B.4. For the hypothesized value of the mean, say μ_0, if $\mu_0 < lcl$ or $\mu_0 > ucl$, H_0 would be rejected. Equivalently, the test statistic in Equation B.2 can be computed using $\mu = \mu_0$ and, for $\alpha = .05$, if it is greater than 1.96 or less than -1.96, H_0 would be rejected with 95% level of confidence.

To further illustrate these concepts, a more detailed example is presented. Atwood et al. (1998) assert that for non-momentary losses of offsite power with plant-centered causes, the recovery times are lognormally distributed with median 29.6 minutes and error factor 10.6. This is equivalent to X being normally distributed with $\mu = \ln(29.6) = 3.388$ and $\sigma = \ln(10.6)/1.645 = 1.435$, where $X = \ln$(recovery time in minutes). Suppose that a plant of interest has experienced five such losses of offsite power in recent history. It is desired to test whether the plant's recovery times follow the claimed distribution.

To simplify the situation, the question is formulated in terms of μ only, assuming that $\sigma = 1.435$. The null hypothesis is

H_0: $\mu = 3.388$.

Because only long recovery times are of concern from a risk standpoint, the alternative hypothesis is defined as

H_1: $\mu > 3.388$.

That is, values <3.388 are possible, but are not of concern. The test statistic, based on $n = 5$ recovery times, is to reject H_0 if

$$Z = \frac{\overline{X} - 3.388}{1.435/\sqrt{5}} > w \quad .$$

To make α, the probability of Type I error, equal to 0.05, w is chosen to be the 95th percentile of the standard normal distribution, 1.645. Then the test can be re-expressed as rejecting H_0 if

$\overline{X} > 4.44$.

The upper part of Figure B.1 shows the density of $\overline{X}$ when $\mu = 3.388$. The area to the right of 4.44 is

$\Pr(\overline{X} > 4.44 \mid H_0$ is true) ,

which equals 0.05.

What if H_0 is false? For example, a median 60-minute recovery time corresponds to $\mu = \ln(60) = 4.09$. The lower part of Figure B.1 shows the density of $\bar{X}$ when $\mu = 4.09$. The area to the right of 4.44 is

$$\Pr(\bar{X} > 4.44 \mid \mu = 4.09),$$

which is equal to 0.29. This value represents the power of the hypothesis test when $\mu = 4.09$ and is the probability of (correctly) rejecting H_0. The probability of a Type II error when $\mu = 4.09$ is 1 - 0.29 = 0.71.

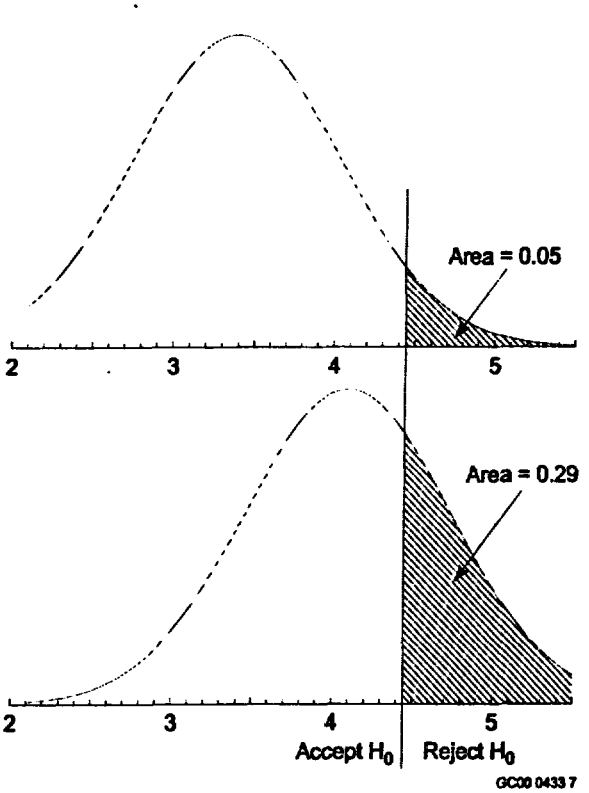

Figure B.1 Probability of rejecting H_0: $\mu = 3.388$, if in fact H_0 is true (upper distribution), and if H_0 is false with $\mu = 4.09$ (lower distribution).

It can be useful to plot the power as a function of μ. The plot is called a **power curve**. Figure B.2 shows two power curves, corresponding to $n = 5$ and $n = 10$. The probability of Type I error, that is, the probability of rejecting H_0 when H_0 is true, is shown as α. The probability of Type II error, that is, the probability of accepting H_0 when H_0 is false, is shown as β for one value of μ, and equals 1 minus the power. The two tests, with $n = 5$ and $n = 10$, have both been calibrated so that $\alpha = 0.05$. The power, for any value of μ in H_1, is larger when $n = 10$ than when $n = 5$; equivalently, the probability of Type II error is smaller.

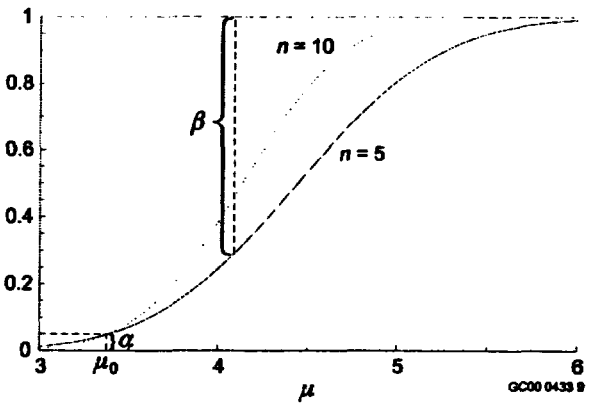

Figure B.2 Power curves when $n = 5$ and $n = 10$. The graph shows the probability of rejecting H_0, as a function of the true μ.

The interpretation of confidence in hypothesis testing is also the same as with confidence intervals. That is, the confidence is not in one specific test statistic. The confidence arises from the viewpoint that if the random sample was collected a number of times in the same way and if H_0 was true, $100(1 - \alpha)\%$ of the tests would result in not rejecting H_0.

As can be seen, interval estimation and hypothesis testing are closely related. Some experimenters prefer expressing inference as estimators. Others prefer to test a particular hypothesized value for the parameter of interest.

B.4.4 Goodness-of-Fit Tests

The methods presented above are concerned with estimating the parameters of a distribution, with the actual form of the distribution assumed to be known (or the central limit theorem applies with large n). Other hypothesis tests do not assume that only a parameter is unknown. In particular, **goodness-of-fit tests** are special hypothesis tests that can be used to check on the assumed distribution itself. Based on a random sample from some distribution, goodness-of-fit tests test the hypothesis that the data are distributed according to a specific distribution. In general, these tests are based on a comparison of how well the sample data agree with an expected set of data from the assumed distribution.

Perhaps the most familiar goodness-of-fit test is the **chi-square** test. The test statistic used for this statistical test has an approximate χ^2 distribution, leading to the name of the test. A random sample of n observations, $X_1, X_2, \ldots, X_n$, can be divided or binned into k groups or intervals, referred to as bins, producing an empirical

distribution. The assumed distribution under the null hypothesis, $f_0(x)$, is used to calculate the probability that an observation would fall in each bin, with the probabilities denoted by $p_1, p_2, ..., p_k$.

These probabilities are frequently referred to as cell probabilities. The k bins are also called cells. The k bin intervals do not overlap and they completely cover the range of values of $f_0(x)$. It follows that $\sum_{i=1}^{k} p_i = 1$. The *expected* frequency of the ith bin, denoted e_i, is $e_i = np_i$, for $i = 1, 2, ..., k$. The e_i are commonly referred to as the expected cell counts. The *observed* frequencies for each of the k bins, denoted O_i, are referred to as observed cell counts.

The chi-square goodness-of-fit test compares the observed frequencies to the corresponding expected frequencies for each of the k groups by calculating the test statistic:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} .$$

If the observations come from some distribution other than that specified in the null hypothesis, the observed frequencies tend to agree poorly with the expected frequencies, and the computed test statistic, X^2, becomes large.

The distribution of the quantity X^2 can be approximated by a chi-square distribution. The parameter that specifies the chi-square distribution is called the **degrees of freedom**. Its value depends on the number of unknown parameters and how they are estimated. When the null-hypothesis distribution is completely specified, such as normal with both μ and σ known, the degrees of freedom are $k - 1$. If, instead, H_0 specifies the form of the distribution but not the parameters, the degrees of freedom must be adjusted. In the example, if $\bar{X}$ and S^2 from the sample are used to estimate μ and σ^2 when testing the distribution, the degrees of freedom are between $k - 1$ and $k - 1 - m$, where m is the number of estimated parameters, 2. If the quantity X^2 is greater that the $1 - \alpha$ quantile of the $\chi^2(k - 1)$ distribution, the hypothesized probability distribution is rejected. If X^2 is less than the $1 - \alpha$ quantile of the $\chi^2(k - 1 - m)$ distribution, the data are concluded to be adequately modeled by $f_0(x)$.

When the sample size is small, the χ^2 distribution still applies as long as the expected frequencies are not too small. Larger expected cell counts make the chi-square distribution approximation better. The problem with small expected frequencies is that a single random observation falling in a group with a small expected frequency would result in that single value having a major contribution to the value of the test statistic, and thus, the test itself. In addition, small expected frequencies are likely to occur only in extreme cases. One rule of thumb is that no expected frequency should be less than 1 (see Snedecor and Cochran, 1989). Two expected frequencies can be near 1 if most of the other expected frequencies are greater than 5. Groups with expected frequencies below 1 should be combined or the groups should be redefined to comply with this rule. Note that k is the number of groups after such combination or redefinition.

Comparing how well sample data agree with an expected set of data leads to another common use of the chi-square test: testing whether two or more classification criteria, used to group subjects or objects, are independent of one another. Although not a goodness-of-fit test, the **chi-square test for independence** is similar to the chi-square goodness-of-fit test.

For two grouping criteria, the rows of a two-way **contingency table** can represent the classes of one of the criteria and the columns can represent the classes of the other criterion. To test the hypothesis that the rows and columns represent independent classifications, the expected number, e_{ij}, that would fall into each cell of the two-way table is calculated and used to compute the following chi-square test statistic:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} ,$$

where $i = 1, 2, ..., r$ (the number of rows); $j = 1, 2, ..., c$ (the number of columns); and O_{ij} is the number observed to belong to the ith row and jth column. The e_{ij} are calculated by

$$e_{ij} = \frac{R_i C_j}{n} ,$$

where R_i and C_j are the total observed in the ith row and jth column, respectively, and n is the total sample size ($n = \Sigma R_i = \Sigma C_j$).

For this test, the χ^2 test statistic follows a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom. If the calculated X^2 exceeds the $1 - \alpha$ quantile of the χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom, the null hypothesis of independence is rejected and the rows and columns are concluded to not represent independent classifications.

The **Kolmogorov goodness-of-fit test** tests the hypothesis that the observed random variable has c.d.f. $F_0(x)$, versus the alternative hypothesis that the observed random variable does not have c.d.f. $F_0(x)$. It does this by comparing the sample c.d.f. (the empirical distribution function) to the hypothesized c.d.f. For a random sample of n observations, $X_1, X_2, ..., X_n$, the test statistic is defined as the maximum vertical distance between the empirical c.d.f., $\hat{F}(x)$ and $F_0(x)$. The actual procedure for calculating the test statistic can be found in many statistics texts, including Martz and Waller (1991) and Conover (1999). The test statistic is then compared to the $1 - \alpha$ quantile of tabled values for the Kolmogorov test, e.g. in Table C. If the calculated test statistic exceeds the $1 - \alpha$ quantile, the hypothesized c.d.f. is rejected. Otherwise, $F_0(x)$ is concluded to describe the data. The Kolmogorov goodness-of-fit test is based on each individual data point and therefore is equally effective for small or large samples.

As an example, consider the previous example of loss-of-offsite-power recovery times. Suppose that five recovery times have been observed at the plant: 7, 22, 94, 185, and 220 minutes. The corresponding values of x = ln(recovery time in minutes) are 1.95, 3.09, 4.54, 5.22, and 5.39. The null hypothesis and alternative hypothesis are:

H_0: X is normal with $\mu = 3.388$, $\sigma = 1.435$
H_1: X has some other distribution .

Note, all possible alternative distributions are considered, not just normal distributions, or distributions with $\sigma = 1.435$.

Figure B.3 shows the distribution function specified by H_0 (the smooth curve) and the empirical distribution function specified by the data (the step function). The maximum distance between the two distributions is D, the Kolmogorov test statistic. If D is large, the test rejects H_0 in favor of H_1.

If the sample size is small, the Kolmogorov test may be preferred over the chi-square test. The Kolmogorov test is exact, even for small samples, while the chi-square test is an approximation that is better for larger sample sizes. The Kolmogorov statistic can also be used to construct a confidence region for the unknown distribution function.

The Kolmogorov goodness-of-fit test is sometimes called the Kolmogorov-Smirnov one-sample test. Statistics that are functions of the maximum vertical distance between $\hat{F}(x)$ and $F_0(x)$ are considered to be Kolmogorov-type statistics. Statistics that are functions of the maximum vertical distance between *two* empirical distribution functions are considered to be Smirnov-type statistics. A test of whether two samples have the same distribution function is the Smirnov test, which is a two-sample version of the Kolmogorov test presented above. This two-sample test is also called the **Kolmogorov-Smirnov two-sample test**. Conover (1999) presents additional information and tests.

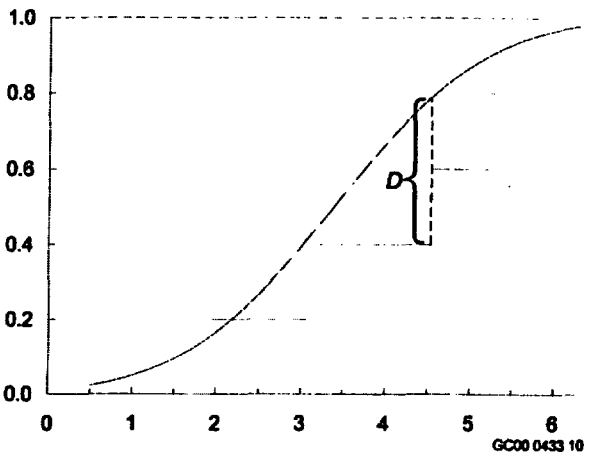

Figure B.3 The hypothesized distribution, the empirical distribution, and the Kolmogorov test statistic, D.

Another useful goodness-of-fit test is the **Anderson-Darling** goodness-of-fit test and test for normality. The Anderson-Darling test measures the squared difference between the empirical distribution function (EDF) of a sample and the theoretical distribution to be tested. It averages this squared difference over the entire range of the random variable, weighting the tails more heavily than the center. This statistic is recommended to guard against wayward observations in the tail and has generally good power.

Because many statistical methods require the assumption of normality, some assessment of whether data come from a normal population is helpful when considering appropriate analysis techniques. The Anderson-Darling statistic provides a measure of how much normal probability scores for the data (normal probability plot values) deviate from a straight line that would arise under normality. A computer package is often used to calculate this statistic and compare it to tabled values for the statistic. If the calculated statistic is too high, the deviations from the straight line are too large to be attributed to the variation due to sampling observations from a normal distribution. Thus, the hypothesis of normality is rejected. See Stephens (1982) for more information on the Anderson-Darling goodness-of-fit test.

Certain patterns of deviations from linearity in normal probability plots indicate common types of nonnormal characteristics, such as skewness or kurtosis (presence of long or short tails of the p.d.f.). Test for skewness or kurtosis are also available. See Snedecor and Cochran (1989) for more information on these tests.

B.5 Bayesian Estimation

B.5.1 Purpose and Use

Bayesian estimation is the other major class of statistical inference methods. Similar to frequentist estimation, both point and interval estimates can be obtained. However, Bayesian estimation is different from classical estimation in both practical and philosophical perspectives (NRC, 1994). Bayesian estimation incorporates degree of belief and information beyond that contained in the data sample, forming the practical difference from classical estimation. The subjective interpretation of probability forms the philosophical difference from frequentist methods.

The prior belief about a parameter's value is contained in what is referred to as the **prior distribution**, which describes the state of knowledge (or subjective probability) about the parameter, prior to obtaining the data sample. Therefore, in the Bayesian approach, the parameters of the sampling distribution have probability distributions. These probabilities do not model random variability of the parameters, but the analyst's degree of belief about the true values of the unknown parameters. The distributions are sometimes called "uncertainty distributions." A Bayesian uncertainty distribution satisfies all the rules of a probability distribution.

Bayesian estimation consists of two main areas, both of which use the notion of subjective probability. The first area involves using available data to assign a subjective, prior distribution to a parameter, such as a failure rate. The degree of belief about the uncertainty in a parameter value is expressed in the prior distribution. This assignment of a prior distribution does not involve the use of Bayes' Theorem. The second area of Bayesian estimation involves using additional or new data to update an existing prior distribution. This is called **Bayesian updating**, and directly uses Bayes' Theorem.

Bayes' Theorem, presented in Section A.5, can be seen to transform the prior distribution by the effect of the sample data distribution to produce a **posterior distribution**. The sample data distribution, $f(x|\theta)$, can be viewed as a function of the unknown parameter, instead of the observed data, x_i, producing a likelihood function, as discussed in Section B.4.1. Using the likelihood function, the fundamental relationship expressed by Bayes' Theorem is

$$\text{Posterior Distribution} = \frac{\text{Prior Distribution} \times \text{Likelihood}}{\text{Marginal Distribution}}.$$

The marginal distribution serves as a normalizing constant.

In Bayesian updating, the sampling distribution of the data provides new information about the parameter value. Bayes' Theorem provides a mathematical framework for processing new sample data as they become sequentially available over time. With the new information, the uncertainty of the parameter value has been reduced, but not eliminated. Bayes' Theorem is used to combine the prior and sampling distributions to form the posterior distribution, which then describes the updated state of knowledge (still in terms of subjective probability) about the parameter. **Point and interval estimates** of the parameter can then be obtained directly from the posterior distribution, which is viewed as containing the current knowledge about the parameter. This posterior distribution can then be used as the prior distribution when the next set of data becomes available. Thus, Bayesian updating is successively implemented using additional data in conjunction with Bayes' Theorem to obtain successively better posterior distributions that model plant-specific parameters.

Bayesian point and interval estimates are obtained from both the prior and posterior distributions. The interval estimates are **subjective probability intervals**, or **credible intervals**. The terminology is not yet universally standard. Berger (1985) and Bernardo and Smith (2000) both use the term **credible interval**, but Box and Tiao (1973) use **Bayes probability interval**, Lindley (1965) uses **Bayesian confidence interval**, and other authors have used other terms. A credible interval can be interpreted as a subjective probability statement about the parameter value, unlike classical interval estimates. That is, the interpretation of a 95% Bayesian posterior probability interval (a, b) is that, with 95% subjective probability, the parameter is contained in the interval (a, b), given the prior and sampling distributions.

B.5.2 Point and Interval Estimates

Bayesian parameter estimation involves four steps. The first step is identification of the parameter(s) to be estimated, which involves consideration of the assumed distribution of the data that will be collected. The second step is development of a prior distribution that

appropriately quantifies the state of knowledge concerning the unknown parameter(s). The third step is collection of the data sample. The fourth and final step is combining the prior distribution with the data sample using Bayes' Theorem, to produce the desired posterior distribution.

For PRA applications, determining the prior distribution is usually based on generic data and the data sample usually involves site-specific or plant-specific operating data. The resulting posterior distribution would then be the site-specific or plant-specific distribution of the parameter.

The plant-specific data collected are assumed to be a random sample from an assumed sampling distribution. The data are used to update the prior, producing the posterior distribution. Point estimates, such as the most likely value (the mode), the median, or (most commonly) the mean value, and probability interval estimates of the parameter can then be obtained. Other bounds and other point values can also be obtained with the Bayesian approach because the posterior parameter distribution is entirely known and represents the available knowledge about the parameter.

Bayesian interval estimation is more direct than classical interval estimation and is based solely on the posterior p.d.f.. A symmetric $100(1 - \alpha)\%$ **two-sided Bayes probability interval estimate** of the parameter is easily obtained from the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution. **Lower and upper one-sided Bayes probability interval estimates** can similarly be calculated. Again, note that the Bayes interval estimates are explicit probability statements of the true parameter being contained in the interval.

In some applications, such as a planned facility, plant-specific data do not exist. In these cases, Bayes' Theorem is not used. Only the generic data are used and parameter estimates are based solely on the assumed prior distribution. Investigation of the sensitivity of the results to the choice of the prior distribution is important for these cases.

B.5.3 Prior Distributions

The prior distribution is fundamental to any Bayesian analysis and represents all that is known or assumed about the parameter θ prior to collecting any data. The information summarized by the prior distribution can be objective, subjective, or both. Operational data and data from a previous but comparable experiment could be used as objective data. Subjective information could involve personal experience and opinions, expert judgement, assessments of degree of belief, and design information.

The selection of prior distributions can be seen to be somewhat subjective. A particular prior must be evaluated to determine the sensitivity of the choice of that prior on the parameter estimates. Consistency of the prior information and data with the prior distribution must be tested.

Choices for the initial prior distribution and techniques for handling various kinds of data are described in detail in several references, such as Martz and Waller (1991), Raiffa and Schlaifer (1961), and Siu and Kelly (1998).

B.5.3.1 Noninformative Prior Distributions

One class of prior distributions that is widely used is termed **noninformative priors**, also referred to as priors of ignorance, or **reference priors** (Bernardo and Smith 1994). These names refer to the situation where very little *a priori* information about a parameter is available in comparison to the information expected to be provided by the data sample, or there is indifference about the range of values the parameter could assume.

One might think that this indifference could be expressed by a prior distribution that is uniformly distributed over the interval of interest. Every value in the interval is equally likely and no knowledge about any specific value over another value is imposed.

However, uniform distributions do not necessarily best reflect true noninformativeness (Box and Tiao 1973), because models can be parameterized in various ways. For example, if the time to failure, T, is exponentially distributed, it is common to write the density of T as

$$f(t) = \lambda e^{-\lambda t}$$

or alternatively as

$$f(t) = \frac{1}{\mu} e^{-t/\mu}.$$

The two parameters are related by $\lambda = 1/\mu$.

A uniform distribution cannot be said to automatically reflect ignorance and be used as a standard noninformative prior distribution. For the exponential example here, ignorance of λ implies ignorance of μ, but λ and μ cannot both have a uniform distribution. In fact, suppose that λ has the uniform distribution in

some finite range, say from a to b. Then μ has a density proportional to $1/\mu^2$ in the range from $1/b$ to $1/a$, as stated in Appendix A.4.7. The distribution of μ is *not* uniform.

Jeffreys' rule (Jeffreys 1961) guides the choice of noninformative prior distributions and provides the Jeffreys prior distribution (Box and Tiao, 1973). The Jeffreys prior distribution is commonly used in PRA and involves using a specific parameterization of the model (distribution). Jeffreys' method is to transform the model into a parameterization that is in terms of a **location parameter**, a parameter that slides the distribution sideways without changing its shape. (See Box and Tiao 1978, Secs. 1.2.3 and 1.3.4). This method then uses the uniform distribution as the noninformative prior for the location parameter. It is reasonable to regard a uniform distribution as noninformative for a location parameter. The distribution for any other parameterization is then determined, and is called noninformative.

In the exponential example, working with log(time), let $\theta = \log(\mu)$, $S = \log(T)$, and $s = \log(t)$. Using algebraic formulas given in Section A.4.7 of Appendix A, it can be shown that the density in this parameterization is

$$f(s) = \exp(s - \theta)e^{-\exp(s-\theta)} .$$

Because θ only appears in the expression $s - \theta$, a change in θ simply slides the distribution sideways along the s axis. Therefore, θ is a location parameter. The Jeffreys noninformative prior is a uniform distribution for θ. This distribution translates to a density for λ which is proportional to $1/\lambda$, and a density for μ which is proportional to $1/\mu$. These are the Jeffreys noninformative prior distributions for λ and μ.

A further argument for Jeffreys prior distributions is that the resulting Bayes intervals are numerically equal to confidence intervals (Lindley 1958), and the confidence intervals are based on the data alone, not on prior belief. Unfortunately, the above approach cannot be followed exactly when the data come from a discrete distribution, such as binomial or Poisson. The original parameter can only approximately be converted to a location parameter. The resulting distribution is still called the Jeffreys prior, however, even though it only approximates the Jeffreys method.

To avoid the appearance of pulling prior distributions out of the air, the general formula for the Jeffreys prior is stated here, as explained by Box and Tiao (1973) and many others. All the particular cases given in this handbook can be found by working out the formula in those cases. Let θ denote the unknown parameter to be estimated. Let $L(\theta, x)$ denote the likelihood corresponding to a single observation. It is a function of θ, but it also depends on the data, x. For example, x is the number of Poisson events in a single unit of time, or the number of failures on a single demand, or the length of a single duration. Calculate

$$-\frac{d^2}{d\theta^2}\ln[L(\theta, x)] .$$

Now replace the number x by the random variable X, and evaluate the expectation of the calculated derivative:

$$E\left(-\frac{d^2}{d\theta^2}\ln[L(\theta, X)]\right) .$$

The Jeffreys noninformative prior is a function of θ proportional to the square root of this expectation.

B.5.3.2 Conjugate Prior Distributions

A **conjugate prior distribution** is a distribution that results in a posterior distribution that is a member of the same family of distributions as the prior. Therefore, conjugate prior distributions are computationally convenient. The methodology for obtaining conjugate priors is based on sufficient statistics and likelihood functions (see Martz and Waller, 1991).

The beta family of distributions is the conjugate family of prior distributions for the probability of failure of a component in a binomial sampling situation. The resulting posterior beta distribution can then be used to provide point and interval estimates of the failure probability.

A time-to-failure random variable is often assumed to follow an exponential distribution, with the failure events arising from a Poisson process. For this model, with either exponential or Poisson data, the gamma family of distributions is the conjugate family of prior distributions for use in Bayesian reliability and failure rate analyses.

Figure B.4 is a schematic diagram showing the relation of the kinds of priors that have been mentioned so far. There are many nonconjugate priors. A relatively small family of priors is conjugate, typically a single type such as the gamma distributions or beta distributions. Finally, the Jeffreys noninformative prior is a single

distribution, shown in the diagram by a dot. In many cases, the Jeffreys prior is also conjugate, as indicated in the figure.

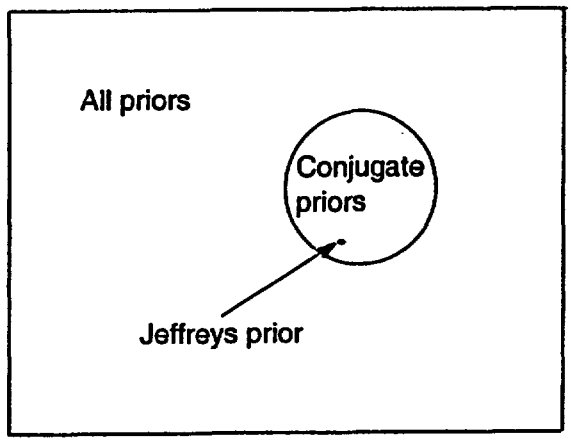

Figure B.4 Schematic diagram of types of priors.

A popular nonconjugate prior is the lognormal distribution. It can be used as a prior distribution for both the binomial sampling and Poisson process models above, although it is not conjugate.

Conjugate prior distributions provide convenience, but accurate modeling of prior degree of belief should not be sacrificed for mathematical convenience. However, when one expression of prior belief is viewed to be as correct as another, the more convenient expression is usually selected for use.

B.5.3.3 Other Prior Distribution Approaches

The prior distribution is the distribution of degree of belief before data that provide new information are obtained. Usually, the prior probabilities do not have a direct frequency interpretation and cannot be experimentally confirmed.

When the prior distribution does have a frequency interpretation, the observed data can be used to estimate the prior distribution. This situation represents another class of methods of statistical inference called **empirical Bayes methods**. The empirical Bayes prior distribution is empirically determined, for example, using observed plant-specific data for a given set of plants. Bayes' Theorem can then be applied to combine this prior with observed data from a specific plant to produce a posterior distribution. Thus, empirical Bayes methods are useful when data from similar, but not identical, sources exist. This situation also gives rise to the use of so-called **hierarchical Bayes** methods (see Gelman, et al., 1995, and Carlin and Louis, 1996).

Attempts have been made to remove some of the subjectivity present in selecting prior distributions, with the goal being to obtain *one* distribution for the same given information. That is, different analysts using the same information would decide upon the same prior distribution. The result has been development of the **method of maximum entropy**. If θ is a parameter with uncertainty distribution g, the entropy is defined as

$$-\int g(\theta)\ln[g(\theta)]d\theta \quad .$$

The distribution g that maximizes this expression is called the maximum entropy distribution. For finite ranges, the p.d.f. with the largest entropy is the uniform, or flat, distribution. Thus, entropy can be viewed as a measure of the variability in the height of a p.d.f., and a maximum entropy prior would be the one with the required mean that is as flat as possible. Siu and Kelly (1998, Table 2) give the maximum entropy distributions for a number of possible constraints.

Maximum entropy methods may see more use in the future, but still do not produce a systematic approach to selecting only one prior from a set of possible priors. In fact, the same problem that the Jeffreys' method attempts to address (Section B.5.3.1) is present with the maximum entropy approach: if a model is parameterized in two different ways, the maximum entropy priors for the two parameters are inconsistent with each other.

To address this lack of invariance, **constrained noninformative priors** are obtained. They are based on the maximum entropy approach in conjunction with Jeffreys' method. That parameterization is used for which the parameter is a location parameter. Giving maximum entropy to this parameter produces a distribution called the constrained noninformative prior distribution. Atwood (1996) presents constrained noninformative priors and their application to PRA. Constrained noninformative prior distributions are seeing use in PRA, although not as much as Jeffreys' priors.

Other ways of defining noninformative prior distributions exist. See Martz and Waller (1991) and Berger (1985) for more information.

REFERENCES FOR APPENDIX B

Atwood, C. L., 1996, Constrained Noninformative Priors in Risk Assessment, in *Reliability Engineering and System Safety*, vol. 53, pp. 37-46.

Atwood, C. L., D. L. Kelly, F. M. Marshall, D. A. Prawdzik and J. W. Stetkar, 1998, *Evaluation of Loss of Offsite Power Events at Nuclear Power Plants: 1980-1996*, NUREG/CR-5496, INEEL/EXT-97-00887, Idaho National Engineering and Environmental Laboratory, Idaho Falls, ID.

Berger, J. O., 1985, *Statistical Decision Theory and Bayesian Analysis, Second Edition*, Springer-Verlag, New York.

Bernardo, J. M., and A. F. M. Smith, 1994, *Bayesian Theory*, John Wiley & Sons, New York.

Box, G. E. P., 1980, Sampling and Bayes' Inference in Scientific Modelling and Robustness (with discussion) in *Journal of the Royal Statistical Society, Series A*, Vol. 143, Part 4, pp. 383-430.

Box, G. E. P., and G. C. Tiao, 1973, *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

Carlin, B. P., and T. A. Louis, 1996, *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

Conover, W. J., 1999, *Practical Nonparametric Statistics*, 3rd ed. John Wiley & Sons, NY.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 1995, *Bayesian Data Analysis*, Chapman & Hall, London.

Hogg, R. V., and A. T. Craig, 1995, *Introduction to Mathematical Statistics*, Macmillan, NY.

Jeffreys, H., 1961, *Theory of Probability, Third Ed.*, Clarendon Press, Oxford, England.

Kendall, M. G., and A. Stuart, 1973, *The Advanced Theory of Statistics*, Vol. 2, Hafner, NY.

Lindley, D. V., 1958, Fiducial Distributions and Bayes' Theorem, in *Journal of the Royal Statistical Society, Series B*, Vol. 20, pp. 102-107.

Lindley, D. V., 1965, *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Cambridge University Press, Cambridge, UK.

Martz, H. F., and R. A. Waller, 1991, *Bayesian Reliability Analysis*, R. E. Krieger Publishing Co., Malabar, FL.

NRC, 1994, *A Review of NRC Staff uses of Probabilistic Risk Assessment*, NUREG-1489, U. S. Nuclear Regulatory Commission, Washington, D.C.

Raiffa, H., and R Schlaifer, 1961, *Applied Statistical Decision Theory*, The M.I.T. Press, Cambridge, MA.

SAS, 1988, *SAS Procedures Guide, Release 6.03 Edition*, SAS Institute, Inc., Cary, NC.

Siu, N., and D. Kelly, 1998, Bayesian Parameter Estimation in Probabilistic Risk Assessment, in *Reliability Engineering and System Safety*, Vol. 62, pp. 89-116.

Snedecor, G. W., and W. G. Cochran, 1989, *Statistical Methods*, Iowa State University Press, Ames, IA.

Stephens, M. A., 1982, Anderson-Darling Test for Goodness of Fit, in *Encyclopedia of Statistical Sciences*, Vol. 1., pp. 81-85, S. Kotz and N. L. Johnson, eds., John Wiley & Sons, NY.