Assessment of the reliability of a hydrosystems infrastructural system or its components involves the use of probability and statistics. This chapter reviews and summarizes some fundamental principles and theories essential to reliability analysis.

2.1 Terminology

In probability theory, an experiment represents the process of making observations of random phenomena. The outcome of an observation from a random phenomenon cannot be predicted with absolute accuracy. The entirety of all possible outcomes of an experiment constitutes the sample space. An event is any subset of outcomes contained in the sample space, and hence an event could be an empty (or null) set, a subset of the sample space, or the sample space itself. Appropriate operators for events are union, intersection, and complement. The occurrence of events $A$ and $B$ is denoted as $A \cup B$ (the union of $A$ and $B$), whereas the joint occurrence of events $A$ and $B$ is denoted as $A \cap B$ or simply $(A, B)$ (the intersection of $A$ and $B$). Throughout the book, the complement of event $A$ is denoted as $A'$.

When two events $A$ and $B$ contain no common elements, then the two events are mutually exclusive or disjoint events, which is expressed as $(A, B) = \emptyset$, where $\emptyset$ denotes the null set. Venn diagrams illustrating the union and intersection of two events are shown in Fig. 2.1. When the occurrence of event $A$ depends on that of event $B$, then they are conditional events,

*Most of this chapter, except Secs. 2.5 and 2.7, is adopted from Tung and Yen (2005).
which is denoted by $A \mid B$. Some useful set operation rules are

1. Commutative rule: $A \cup B = B \cup A$; $A \cap B = B \cap A$.
2. Associative rule: $(A \cup B) \cup C = A \cup (B \cup C)$; $(A \cap B) \cap C = A \cap (B \cap C)$.
3. Distributive rule: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
4. de Morgan’s rule: $(A \cup B)' = A' \cap B'$; $(A \cap B)' = A' \cup B'$.

Probability is a numeric measure of the likelihood of the occurrence of an event. Therefore, probability is a real-valued number that can be manipulated by ordinary algebraic operators, such as $+, -, \times$, and $\div$. The probability of the occurrence of an event $A$ can be assessed in two ways. In the case where an experiment can be repeated, the probability of having event $A$ occurring can be estimated as the ratio of the number of replications in which event $A$ occurs $n_A$ versus the total number of replications $n$, that is, $n_A/n$. This ratio is called the relative frequency of occurrence of event $A$ in the sequence of $n$ replications. In principle, as the number of replications gets larger, the value of the relative
frequency becomes more stable, and the true probability of event $A$ occurring could be obtained as

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n} \quad (2.1)$$

The probabilities so obtained are called *objective or posterior probabilities* because they depend completely on observations of the occurrence of the event.

In some situations, the physical performance of an experiment is prohibited or impractical. The probability of the occurrence of an event can only be estimated subjectively on the basis of experience and judgment. Such probabilities are called *subjective or prior probabilities*.

### 2.2 Fundamental Rules of Probability Computations

#### 2.2.1 Basic axioms of probability

The three basic axioms of probability computation are (1) *nonnegativity*: $P(A) \geq 0$, (2) *totality*: $P(S) = 1$, with $S$ being the sample space, and (3) *additivity*: For two mutually exclusive events $A$ and $B$, $P(A \cup B) = P(A) + P(B)$.

As indicated from axioms (1) and (2), the value of probability of an event occurring must lie between 0 and 1. Axiom (3) can be generalized to consider $K$ mutually exclusive events as

$$P(A_1 \cup A_2 \cup \cdots \cup A_K) = P\left( \bigcup_{k=1}^{K} A_k \right) = \sum_{k=1}^{K} P(A_k) \quad (2.2)$$

An *impossible event* is an empty set, and the corresponding probability is zero, that is, $P(\emptyset) = 0$. Therefore, two mutually exclusive events $A$ and $B$ have zero probability of joint occurrence, that is, $P(A, B) = P(\emptyset) = 0$. Although the probability of an impossible event is zero, the reverse may not necessarily be true. For example, the probability of observing a flow rate of exactly 2000 m$^3$/s is zero, yet having a discharge of 2000 m$^3$/s is not an impossible event.

Relaxing the requirement of mutual exclusiveness in axiom (3), the probability of the union of two events can be evaluated as

$$P(A \cup B) = P(A) + P(B) - P(A, B) \quad (2.3)$$

which can be further generalized as

$$P\left( \bigcup_{k=1}^{K} A_k \right) = \sum_{k=1}^{K} P(A_k) - \sum_{i<j} P(A_i, A_j) + \sum_{i<j<k} P(A_i, A_j, A_k) - \cdots + (-1)^K P(A_1, A_2, \ldots, A_K) \quad (2.4)$$
If all are mutually exclusive, all but the first summation term on the right-hand side of Eq. (2.3) vanish, and it reduces to Eq. (2.2).

**Example 2.1** There are two tributaries in a watershed. From past experience, the probability that water in tributary 1 will overflow during a major storm event is 0.5, whereas the probability that tributary 2 will overflow is 0.4. Furthermore, the probability that both tributaries will overflow is 0.3. What is the probability that at least one tributary will overflow during a major storm event?

**Solution** Define $E_i =$ event that tributary $i$ overflows for $i = 1, 2$. From the problem statements, the following probabilities are known: $P(E_1) = 0.5$, $P(E_2) = 0.4$, and $P(E_1, E_2) = 0.3$.

The probability having at least one tributary overflowing is the probability of event $E_1$ or $E_2$ occurring, that is, $P(E_1 \cup E_2)$. Since the overflow of one tributary does not preclude the overflow of the other tributary, $E_1$ and $E_2$ are not mutually exclusive. Therefore, the probability that at least one tributary will overflow during a major storm event can be computed, according to Eq. (2.3), as

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1, E_2) = 0.5 + 0.4 - 0.3 = 0.6$$

### 2.2.2 Statistical independence

If two events are *statistically independent* of each other, the occurrence of one event has no influence on the occurrence of the other. Therefore, events $A$ and $B$ are independent if and only if $P(A, B) = P(A)P(B)$. The probability of joint occurrence of $K$ independent events can be generalized as

$$P\left(\bigcap_{k=1}^{K} A_k\right) = P(A_1) \times P(A_2) \times \ldots \times P(A_K) = \prod_{k=1}^{K} P(A_k) \quad (2.5)$$

It should be noted that the mutual exclusiveness of two events does not, in general, imply independence, and vice versa, unless one of the events is an impossible event. If the two events $A$ and $B$ are independent, then $A, A', B$, and $B'$ all are independent, but not necessarily mutually exclusive, events.

**Example 2.2** Referring to Example 2.1, the probabilities that tributaries 1 and 2 overflow during a major storm event are 0.5 and 0.4, respectively. For simplicity, assume that the occurrences of overflowing in the two tributaries are independent of each other. Determine the probability of at least one tributary overflowing in a major storm event.

**Solution** Use the same definitions for events $E_1$ and $E_2$. The problem is to determine $P(E_1 \cup E_2)$ by

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1, E_2)$$

Note that in this example the probability of joint occurrences of both tributaries overflowing, that is, $P(E_1, E_2)$, is not given directly by the problem statement, as in Example 2.1. However, it can be determined from knowing that the occurrences of
overflows in the tributaries are independent events, according to Eq. (2.5), as

\[ P(E_1, E_2) = P(E_1)P(E_2) = (0.5)(0.4) = 0.2 \]

Then the probability that at least one tributary would overflow during a major storm event is

\[ P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1, E_2) = 0.5 + 0.4 - 0.2 = 0.7 \]

### 2.2.3 Conditional probability

The *conditional probability* is the probability that a conditional event would occur. The conditional probability \( P(A \mid B) \) can be computed as

\[ P(A \mid B) = \frac{P(A, B)}{P(B)} \quad \text{(2.6)} \]

in which \( P(A \mid B) \) is the occurrence probability of event \( A \) given that event \( B \) has occurred. It represents a reevaluation of the occurrence probability of event \( A \) in the light of the information that event \( B \) has occurred. Intuitively, \( A \) and \( B \) are two independent events if and only if \( P(A \mid B) = P(A) \). In many cases it is convenient to compute the joint probability \( P(A, B) \) by

\[ P(A, B) = P(B)P(A \mid B) \quad \text{or} \quad P(A, B) = P(A)P(B \mid A) \]

The probability of the joint occurrence of \( K \) dependent events can be generalized as

\[ P\left( \bigcap_{k=1}^{K} A_k \right) = P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_2, A_1) \times \cdots \times P(A_K \mid A_{K-1}, \ldots, A_2, A_1) \quad \text{(2.7)} \]

**Example 2.3** Referring to Example 2.2, the probabilities that tributaries 1 and 2 would overflow during a major storm event are 0.5 and 0.4, respectively. After examining closely the assumption about the independence of overflow events in the two tributaries, its validity is questionable. Through an analysis of historical overflow events, it is found that the probability of tributary 2 overflowing is 0.6 if tributary 1 overflows. Determine the probability that at least one tributary would overflow in a major storm event.

**Solution** Let \( E_1 \) and \( E_2 \) be the events that tributary 1 and 2 overflow, respectively. From the problem statement, the following probabilities can be identified:

\[ P(E_1) = 0.5 \quad P(E_2) = 0.4 \quad P(E_2 \mid E_1) = 0.6 \]

in which \( P(E_2 \mid E_1) \) is the conditional probability representing the likelihood that tributary 2 would overflow given that tributary 1 has overflowed. The probability of at least one tributary overflowing during a major storm event can be computed by

\[ P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1, E_2) \]
in which the probability of joint occurrence of both tributaries overflowing, that
is, \( P(E_1, E_2) \), can be obtained from the given conditional probability, according to
Eq. (2.7), as

\[ P(E_1, E_2) = P(E_2 | E_1) P(E_1) = (0.6)(0.5) = 0.3 \]

The probability that at least one tributary would overflow during a major storm event
can be obtained as

\[ P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1, E_2) = 0.5 + 0.4 - 0.3 = 0.6 \]

### 2.2.4 Total probability theorem and Bayes’ theorem

The probability of the occurrence of an event \( E \), in general, cannot be deter-
dined directly or easily. However, the event \( E \) may occur along with other
attribute events \( A_k \). Referring to Fig. 2.2, event \( E \) could occur jointly with \( K \)
mutually exclusive (\( A_j \cap A_k = \emptyset \) for \( j \neq k \) and collectively exhaustive (\( A_1 \cup A_2 \cup \cdots \cup A_K \) = \( S \) attributes \( A_k, k = 1, 2, \ldots, K \). Then the probability of the occur-
rence of an event \( E \), regardless of the attributes, can be computed as

\[ P(E) = \sum_{k=1}^{K} P(E, A_k) = \sum_{k=1}^{K} P(E | A_k) P(A_k) \] (2.8)

Equation (2.8) is called the total probability theorem.

**Example 2.4** Referring to Fig. 2.3, two upstream storm sewer branches (\( I_1 \) and \( I_2 \))
merge to a sewer main (\( I_3 \)). Assume that the flow-carrying capacities of the two up-
stream sewer branches \( I_1 \) and \( I_2 \) are equal. However, hydrologic characteristics of
the contributing drainage basins corresponding to \( I_1 \) and \( I_2 \) are somewhat differ-
ent. Therefore, during a major storm event, the probabilities that sewers \( I_1 \) and \( I_2 \)
will exceed their capacities (surcharge) are 0.5 and 0.4, respectively. For simplicity,
assume that the occurrences of surcharge events in the two upstream sewer branches
are independent of each other. If the flow capacity of the downstream sewer main \( I_3 \) is

![Figure 2.2 Schematic diagram of total probability theorem.](image-url)
the same as its two upstream branches, what is the probability that the flow capacity of the sewer main $I_3$ will be exceeded? Assume that when both upstream sewers are carrying less than their full capacities, the probability of downstream sewer main $I_3$ exceeding its capacity is 0.2.

**Solution** Let $E_1$, $E_2$, and $E_3$, respectively, be events that sewer $I_1$, $I_2$, and $I_3$ exceed their respective flow capacity. From the problem statements, the following probabilities can be identified: $P(E_1) = 0.50$, $P(E_2) = 0.40$, and $P(E_3 | E'_1, E'_2) = 0.2$.

To determine $P(E_3)$, one first considers the basic events occurring in the two upstream sewer branches that would result in surcharge in the downstream sewer main $E_3$. There are four possible attribute events that can be defined from the flow conditions of the two upstream sewers leading to surcharge in the downstream sewer main. They are $A_1 = (E_1, E_2)$, $A_2 = (E'_1, E_2)$, $A_3 = (E_1, E'_2)$, and $A_4 = (E'_1, E'_2)$. Furthermore, the four events $A_1$, $A_2$, $A_3$, and $A_4$ are mutually exclusive.

Since the four attribute events $A_1$, $A_2$, $A_3$, and $A_4$ contribute to the occurrence of event $E_3$, the probability of the occurrence of $E_3$ can be calculated, according to Eq. (2.8), as

$$P(E_3) = P(E_3, A_1) + P(E_3, A_2) + P(E_3, A_3) + P(E_3, A_4)$$
$$= P(E_3 | A_1)P(A_1) + P(E_3 | A_2)P(A_2) + P(E_3 | A_3)P(A_3) + P(E_3 | A_4)P(A_4)$$

To solve this equation, each of the probability terms on the right-hand side must be identified. First, the probability of the occurrence of $A_1$, $A_2$, $A_3$, and $A_4$ can be determined as the following:

$$P(A_1) = P(E_1, E_2) = P(E_1) \times P(E_2) = (0.5)(0.4) = 0.2$$

The reason that $P(E_1, E_2) = P(E_1) \times P(E_2)$ is due to the independence of events $E_1$ and $E_2$. Since $E_1$ and $E_2$ are independent events, then $E_1$, $E'_1$, $E_2$, and $E'_2$ are also
independent events. Therefore,
\[
P(A_2) = P(E'_1, E_2) = P(E'_1) \times P(E_2) = (1 - 0.5)(0.4) = 0.2 \\
P(A_3) = P(E_1, E'_2) = P(E_1) \times P(E'_2) = (0.5)(1 - 0.4) = 0.3 \\
P(A_4) = P(E'_1, E'_2) = P(E'_1) \times P(E'_2) = (1 - 0.5)(1 - 0.4) = 0.3
\]

The next step is to determine the values of the conditional probabilities, that is, \(P(E_3 | A_1), P(E_2 | A_2), P(E_3 | A_3), \) and \(P(E_3 | A_4).\) The value of \(P(E_3 | A_4) = P(E_3 | E'_1, E'_2) = 0.2\) is given by the problem statement. On the other hand, the values of the remaining three conditional probabilities can be determined from an understanding of the physical process. Note that from the problem statement the downstream sewer main has the same conveyance capacity as the two upstream sewers. Hence any upstream sewer exceeding its flow-carrying capacity would result in surcharge in the downstream sewer main. Thus the remaining three conditional probabilities can be easily determined as
\[
P(E_3 | A_1) = P(E_3 | E_1, E_2) = 1.0 \\
P(E_3 | A_2) = P(E_3 | E'_1, E_2) = 1.0 \\
P(E_3 | A_3) = P(E_3 | E_1, E'_2) = 1.0
\]

Putting all relevant information into the total probability formula given earlier, the probability that the downstream sewer main \(I_3\) would be surcharged in a major storm is
\[
P(E_3) = P(E_3 | A_1)P(A_1) + P(E_3 | A_2)P(A_2) + P(E_3 | A_3)P(A_3) + P(E_3 | A_4)P(A_4) \\
= (1.0)(0.2) + (1.0)(0.2) + (1.0)(0.3) + (0.2)(0.3) \\
= 0.76
\]

The total probability theorem describes the occurrence of an event \(E\) that may be affected by a number of attribute events \(A_k, k = 1, 2, \ldots, K.\) In some situations, one knows \(P(E|A_k)\) and would like to determine the probability that a particular event \(A_k\) contributes to the occurrence of event \(E.\) In other words, one likes to find \(P(A_k|E).\) Based on the definition of the conditional probability (Eq. 2.6) and the total probability theorem (Eq. 2.8), \(P(A_k|E)\) can be computed as
\[
P(A_k | E) = \frac{P(A_k, E)}{P(E)} = \frac{P(E | A_k)P(A_k)}{\sum_{k=1}^{K} P(E | A_k)P(A_k)} \quad \text{for } k = 1, 2, \ldots, K \quad (2.9)
\]

Equation (2.9) is called Bayes' theorem, and \(P(A_k)\) is the prior probability, representing the initial belief of the likelihood of occurrence of attribute event \(A_k.\) \(P(E | A_k)\) is the likelihood function, and \(P(A_k | E)\) is the posterior probability, representing the new evaluation of \(A_k\) being responsible in the light of the occurrence of event \(E.\) Hence Bayes' theorem can be used to update and revise the calculated probability as more information becomes available.
Example 2.5  Referring to Example 2.4, if surcharge is observed in the downstream storm sewer main \( I_3 \), what is the probability that the incident is caused by simultaneous surcharge of both upstream sewer branches?

Solution  From Example 2.4, \( A_1 \) represents the event that both upstream storm sewer branches exceed their flow-carrying capacities. The problem is to find the conditional probability of \( A_1 \), given that event \( E_3 \) has occurred, that is, \( P(A_1 \mid E_3) \). This conditional probability can be expressed as

\[
P(A_1 \mid E_3) = \frac{P(A_1, E_3)}{P(E_3)} = \frac{P(E_3 \mid A_1)P(A_1)}{P(E_3)}
\]

From Example 2.4, the numerator and denominator of the preceding conditional probability can be computed as

\[
P(A_1 \mid E_3) = \frac{P(E_3 \mid A_1)P(A_1)}{P(E_3)} = \frac{(1.0)(0.2)}{0.76} = 0.263
\]

The original assessment of the probability is 20 percent that both upstream sewer branches would exceed their flow-carrying capacities. After an observation of downstream surcharge from a new storm event, the probability of surcharge occurring in both upstream sewers is revised to 26.3 percent.

2.3 Random Variables and their Distributions

In analyzing the statistical features of infrastructural system responses, many events of interest can be defined by the related random variables. A random variable is a real-value function defined on the sample space. In other words, a random variable can be viewed as a mapping from the sample space to the real line, as shown in Fig. 2.4. The standard convention is to denote a random variable by an upper-case letter, whereas a lower-case letter is used to represent the realization of the corresponding random variable. For example, one may use \( Q \) to represent flow magnitude, a random variable, whereas \( q \) is used to represent the values that \( Q \) takes. A random variable can be discrete or continuous. Examples of discrete random variables encountered in hydrosystems infrastructural designs are the number of storm events occurring in a specified time period, the number of overtopping events per year for a levee system, and so on. On the other hand, examples of continuous random variables are flow rate, rainfall intensity, water-surface elevation, roughness factor, and pollution concentration, among others.

2.3.1 Cumulative distribution function and probability density function

The cumulative distribution function (CDF), or simply distribution function (DF), of a random variable \( X \) is defined as

\[
F_X(x) = P(X \leq x)
\]
The CDF $F_x(x)$ is the nonexceedance probability, which is a nondecreasing function of the argument $x$, that is, $F_x(a) \leq F_x(b)$, for $a < b$. As the argument $x$ approaches the lower bounds of the random variable $X$, the value of $F_x(x)$ approaches zero, that is, $\lim_{x \to -\infty} F_x(x) = 0$; on the other hand, the value of $F_x(x)$ approaches unity as its argument approaches the upper bound of $X$, that is, $\lim_{x \to \infty} F_x(x) = 1$. With $a < b$, $P(a < X \leq b) = F_x(b) - F_x(a)$.

For a discrete random variable $X$, the probability mass function (PMF), is defined as

$$p_x(x) = P(X = x) \quad (2.11)$$

The PMF of any discrete random variable, according to axioms (1) and (2) in Sec. 2.1, must satisfy two conditions: (1) $p_x(x_k) \geq 0$, for all $x_k$'s, and (2) $\sum_{all \ k} p_x(x_k) = 1$. The PMF of a discrete random variable and its associated CDF are sketched schematically in Fig. 2.5. As can be seen, the CDF of a discrete random variable is a staircase function.

For a continuous random variable, the probability density function (PDF) $f_x(x)$ is defined as

$$f_x(x) = \frac{dF_x(x)}{dx} \quad (2.12)$$
The PDF of a continuous random variable $f_x(x)$ is the slope of its corresponding CDF. Graphic representations of a PDF and a CDF are shown in Fig. 2.6. Similar to the discrete case, any PDF of a continuous random variable must satisfy two conditions: (1) $f_x(x) \geq 0$ and (2) $\int f_x(x) \, dx = 1$. Given the PDF of a random variable $X$, its CDF can be obtained as

$$F_x(x) = \int_{-\infty}^{x} f_x(u) \, du$$

in which $u$ is the dummy variable. It should be noted that $f_x(\cdot)$ is not a probability; it only has meaning when it is integrated between two points. The probability of a continuous random variable taking on a particular value is zero, whereas this may not be the case for discrete random variables.
Example 2.6  The time to failure $T$ of a pump in a water distribution system is a continuous random variable having the PDF of

$$f_t(t) = \exp(-t/1250)/\beta \quad \text{for } t \geq 0, \; \beta > 0$$

in which $t$ is the elapsed time (in hours) before the pump fails, and $\beta$ is the parameter of the distribution function. Determine the constant $\beta$ and the probability that the operating life of the pump is longer than 200 h.

Solution  The shape of the PDF is shown in Fig. 2.7. If the function $f_t(t)$ is to serve as a PDF, it has to satisfy two conditions: (1) $f_t(t) \geq 0$, for all $t$, and (2) the area under $f_t(t)$ must equal unity. The compliance of the condition (1) can be proved easily. The value of the constant $\beta$ can be determined through condition (2) as

$$1 = \int_0^\infty f_t(t) \, dt = \int_0^\infty \frac{e^{-t/1250}}{\beta} \, dt = \left[ \frac{-1250e^{-t/1250}}{\beta} \right]_0^\infty = \frac{1250}{\beta}$$
Therefore, the constant $\beta = 1250 \text{ h/failure}$. This particular PDF is called the exponential distribution (see Sec. 2.6.3). To determine the probability that the operational life of the pump would exceed 200 h, one calculates $P(T \geq 200)$:

$$P(T \geq 200) = \int_{200}^{\infty} \frac{e^{-t/1250}}{1250} \, dt = \left[ -e^{-t/1250} \right]_{200}^{\infty} = e^{-200/1250} = 0.852$$

### 2.3.2 Joint, conditional, and marginal distributions

The joint distribution and conditional distribution, analogous to the concepts of joint probability and conditional probability, are used for problems involving multiple random variables. For example, flood peak and flood volume often are considered simultaneously in the design and operation of a flood-control reservoir. In such cases, one would need to develop a joint PDF of flood peak and flood volume. For illustration purposes, the discussions are limited to problems involving two random variables.

The joint PMF and joint CDF of two discrete random variables $X$ and $Y$ are defined, respectively, as

$$p_{x,y}(x, y) = P(X = x, Y = y) \quad (2.14a)$$

$$F_{x,y}(u, v) = P(X \leq u, Y \leq v) = \sum_{x \leq u} \sum_{y \leq v} p_{x,y}(x, y) \quad (2.14b)$$

Schematic diagrams of the joint PMF and joint CDF of two discrete random variables are shown in Fig. 2.8.
The joint PDF of two continuous random variables $X$ and $Y$, denoted as $f_{x,y}(x, y)$, is related to its corresponding joint CDF as

$$ f_{x,y}(x, y) = \frac{\partial^2 [F_{x,y}(x, y)]}{\partial x \partial y} $$ \hfill (2.15a)

$$ F_{x,y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{x,y}(u, v) \, du \, dv $$ \hfill (2.15b)

Similar to the univariate case, $F_{x,y}(-\infty, -\infty) = 0$ and $F_{x,y}(\infty, \infty) = 1$. Two random variables $X$ and $Y$ are statistically independent if and only if $f_{x,y}(x, y) = f_x(x) \times f_y(y)$ and $F_{x,y}(x, y) = F_x(x) \times F_y(y)$. Hence a problem involving multiple independent random variables is, in effect, a univariate problem in which each individual random variable can be treated separately.
If one is interested in the distribution of one random variable regardless of all others, the **marginal distribution** can be used. Given the joint PDF $f_{x,y}(x, y)$, the **marginal PDF** of a random variable $X$ can be obtained as

$$f_x(x) = \int_{-\infty}^{\infty} f_{x,y}(x, y) \, dy$$

(2.16)

For continuous random variables, the **conditional PDF** for $X \mid Y$, similar to the conditional probability shown in Eq. (2.6), can be defined as

$$f_x(x \mid y) = \frac{f_{x,y}(x, y)}{f_y(y)}$$

(2.17)

in which $f_y(y)$ is the marginal PDF of random variable $Y$. The conditional PMF for two discrete random variables similarly can be defined as

$$p_x(x \mid y) = \frac{p_{x,y}(x, y)}{p_y(y)}$$

(2.18)

Figure 2.9 shows the joint and marginal PDFs of two continuous random variables $X$ and $Y$. It can be shown easily that when the two random variables are statistically independent, $f_x(x \mid y) = f_x(x)$.

Equation (2.17) alternatively can be written as

$$f_{x,y}(x, y) = f_x(x \mid y) \times f_y(y)$$

(2.19)

which indicates that a joint PDF between two correlated random variables can be formulated by multiplying a conditional PDF and a suitable marginal PDF.

**Figure 2.9** Joint and marginal probability density function (PDFs) of two continuous random variables. *(After Ang and Tang, 1975.)*
Note that the marginal distributions can be obtained from the joint distribution function, but not vice versa.

**Example 2.7** Suppose that $X$ and $Y$ are two random variables that can only take values in the intervals $0 \leq x \leq 2$ and $0 \leq y \leq 2$. Suppose that the joint CDF of $X$ and $Y$ for these intervals has the form of $F_{x,y}(x, y) = cxy(x^2 + y^2)$. Find (a) the joint PDF of $X$ and $Y$, (b) the marginal PDF of $X$, (c) the conditional PDF $f_{y|x}(y|x = 1)$, and (d) $P(Y \leq 1 | x = 1)$.

**Solution** First, one has to find the constant $c$ so that the function $F_{x,y}(x, y)$ is a legitimate CDF. It requires that the value of $F_{x,y}(x, y) = 1$ when both arguments are at their respective upper bounds. That is,

$$F_{x,y}(x = 2, y = 2) = 1 = c(2)(2)(2^2 + 2^2)$$

Therefore, $c = 1/32$. The resulting joint CDF is shown in Fig. 2.10a.

![Joint CDF](image-a)

![Joint PDF](image-b)

**Figure 2.10** (a) Joint cumulative distribution function (CDF) and (b) probability density function (PDF) for Example 2.7.
(a) To derive the joint PDF, Eq. (2.15a) is applied, that is,

\[ f_{x,y}(x, y) = \frac{\partial}{\partial x} \left\{ \frac{\partial}{\partial y} \left[ xy(x^2 + y^2)^{3/2} \right] \right\} = \frac{\partial}{\partial x} \left( \frac{x^3 + 3xy^2}{32} \right) = \frac{3(x^2 + y^2)}{32} \text{ for } 0 \leq x, y \leq 2 \]

A plot of joint PDF is shown in Fig. 2.10b.

(b) To find the marginal distribution of \( X \), Eq. (2.16) can be used:

\[ f_x(x) = \int_0^2 \frac{3(x^2 + y^2)}{32} \, dx = \frac{4 + 3x^2}{16} \text{ for } 0 \leq x \leq 2 \]

(c) The conditional distribution \( f_y(y | x) \) can be obtained by following Eq. (2.17) as

\[ f_y(y | x = 1) = \frac{f_{x,y}(x = 1, y)}{f_x(x = 1)} = \frac{3(1 + y^2)}{4 + 3(1)^2} = \frac{3(1 + y^2)}{14} \text{ for } 0 \leq y \leq 2 \]

(d) The conditional probability \( P(Y \leq 1 | X = 1) \) can be computed as

\[ P(Y \leq 1 | X = 1) = \int_0^1 f_y(y | x = 1) \, dy = \int_0^1 \frac{3(1 + y^2)}{14} \, dy = \frac{2}{7} \]

2.4 Statistical Properties of Random Variables

In statistics, the term population is synonymous with the sample space, which describes the complete assemblage of all the values representative of a particular random process. A sample is any subset of the population. Furthermore, parameters in a statistical model are quantities that are descriptive of the population. In this book, Greek letters are used to denote statistical parameters. Sample statistics, or simply statistics, are quantities calculated on the basis of sample observations.

2.4.1 Statistical moments of random variables

In practical statistical applications, descriptors commonly used to show the statistical properties of a random variable are those indicative of (1) the central tendency, (2) the dispersion, and (3) the asymmetry of a distribution. The frequently used descriptors in these three categories are related to the statistical moments of a random variable. Currently, two types of statistical moments are used in hydro systems engineering applications: product-moments and L-moments. The former is a conventional one with a long history of practice, whereas the latter has been receiving great attention recently from water resources engineers in analyzing hydrologic data (Stedinger et al., 1993; Rao and Hamed 2000). To be consistent with the current general practice and usage, the terms moments and statistical moments in this book refer to the conventional product-moments unless otherwise specified.
Product-moments. The \( r \)th-order product-moment of a random variable \( X \) about any reference point \( X = x_0 \) is defined, for the continuous case, as

\[
E[(X - x_0)^r] = \int_{-\infty}^{\infty} (x - x_0)^r f_x(x) \, dx = \int_{-\infty}^{\infty} (x - x_0)^r \, dF_x(x) \tag{2.20a}
\]

whereas for the discrete case,

\[
E[(X - x_0)^r] = \sum_{k=1}^{K} (x_k - x_0)^r p_x(x_k) \tag{2.20b}
\]

where \( E[\cdot] \) is a statistical expectation operator. In practice, the first three moments (\( r = 1, 2, 3 \)) are used to describe the central tendency, variability, and asymmetry of the distribution of a random variable. Without losing generality, the following discussions consider continuous random variables. For discrete random variables, the integral sign is replaced by the summation sign. Here it is convenient to point out that when the PDF in Eq. (2.20a) is replaced by a conditional PDF, as described in Sec. 2.3, the moments obtained are called the conditional moments.

Since the expectation operator \( E[\cdot] \) is for determining the average value of the random terms in the brackets, the sample estimator for the product-moments for \( \mu'_r = E(X^r) \), based on \( n \) available data \( (x_1, x_2, \ldots, x_n) \), can be written as

\[
\hat{\mu}'_r = \sum_{i=1}^{n} w_i(n) x_i^r \tag{2.21}
\]

where \( w_i(n) \) is a weighting factor for sample observation \( x_i \), which depends on sample size \( n \). Most commonly, \( w_i(n) = 1/n \), for all \( i = 1, 2, \ldots, n \). The last column of Table 2.1 lists the formulas applied in practice for computing some commonly used statistical moments.

Two types of product-moments are used commonly: moments about the origin, where \( x_0 = 0 \), and central moments, where \( x_0 = \mu_x \), with \( \mu_x = E[X] \). The \( r \)th-order central moment is denoted as \( \mu_r = E[(X - \mu_x)^r] \), whereas the \( r \)th-order moment about the origin is denoted as \( \mu'_r = E(X^r) \). It can be shown easily, through the binomial expansion, that the central moments \( \mu_r = E[(X - \mu_x)^r] \) can be obtained from the moments about the origin as

\[
\mu_r = \sum_{i=0}^{r} (-1)^i C_{r,i} \mu_x^i \mu'_{r-i} \tag{2.22}
\]

where \( C_{r,i} = \binom{r}{i} \) is a binomial coefficient, with \( ! \) representing factorial, that is, \( r! = r \times (r-1) \times (r-2) \times \cdots \times 2 \times 1 \). Conversely, the moments about the origin can be obtained from the central moments in a similar fashion as

\[
\mu'_r = \sum_{i=0}^{r} C_{r,i} \mu_x^i \mu_{r-i} \tag{2.23}
\]
<table>
<thead>
<tr>
<th>Moment</th>
<th>Measure of</th>
<th>Definition</th>
<th>Continuous variable</th>
<th>Discrete variable</th>
<th>Sample estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>Central location</td>
<td>Mean, expected value ( E(X) = \mu_x )</td>
<td>( \mu_x = \int_{-\infty}^{\infty} xf_x(x) , dx )</td>
<td>( \mu_x = \sum_{x \in X} x \cdot p(x) )</td>
<td>( \bar{x} = \sum x_i / n )</td>
</tr>
<tr>
<td>Second</td>
<td>Dispersion</td>
<td>Variance, ( \text{Var}(X) = \mu_2 = \sigma_x^2 )</td>
<td>( \sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_x(x) , dx )</td>
<td>( \sigma_x^2 = \sum_{x \in X} (x_k - \mu_x)^2 P_x(x_k) )</td>
<td>( s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard deviation, ( \sigma_x )</td>
<td>( \sigma_x = \sqrt{\text{Var}(X)} )</td>
<td>( \sigma_x = \sqrt{\sum_{x \in X} (x_k - \mu_x)^2 P_x(x_k)} )</td>
<td>( s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coefficient of variation, ( \Omega_x )</td>
<td>( \Omega_x = \sigma_x / \mu_x )</td>
<td>( \Omega_x = \sigma_x / \mu_x )</td>
<td>( C_v = s / \bar{x} )</td>
</tr>
<tr>
<td>Third</td>
<td>Asymmetry</td>
<td>Skewness</td>
<td>( \mu_3 = \int_{-\infty}^{\infty} (x - \mu_x)^3 f_x(x) , dx )</td>
<td>( \mu_3 = \sum_{x \in X} (x_k - \mu_x)^3 P_x(x_k) )</td>
<td>( m_3 = \frac{n}{n-1}(n-2) \sum (x_i - \bar{x})^3 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skewness coefficient, ( \gamma_x )</td>
<td>( \gamma_x = \mu_3 / \sigma_x^3 )</td>
<td>( \gamma_x = \mu_3 / \sigma_x^3 )</td>
<td>( g = m_3 / s^3 )</td>
</tr>
<tr>
<td>Fourth</td>
<td>Peakedness</td>
<td>Kurtosis, ( \kappa_x )</td>
<td>( \mu_4 = \int_{-\infty}^{\infty} (x - \mu_x)^4 f_x(x) , dx )</td>
<td>( \mu_4 = \sum_{x \in X} (x_k - \mu_x)^4 P_x(x_k) )</td>
<td>( m_4 = \frac{n}{n-1}(n-2)(n-3) \sum (x_i - \bar{x})^4 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Excess coefficient, ( \varepsilon_x )</td>
<td>( \kappa_x = \mu_4 / \sigma_x^4 )</td>
<td>( \kappa_x = \mu_4 / \sigma_x^4 )</td>
<td>( k = m_4 / s^4 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \varepsilon_x = \kappa_x - 3 )</td>
<td>( \varepsilon_x = \kappa_x - 3 )</td>
<td>( \varepsilon_x = \kappa_x - 3 )</td>
</tr>
</tbody>
</table>
Equation (2.22) enables one to compute central moments from moments about the origin, whereas Eq. (2.23) does the opposite. Derivations for the expressions of the first four central moments and the moments about the origin are left as exercises (Problems 2.10 and 2.11).

The main disadvantages of the product-moments are (1) that estimation from sample observations is sensitive to the presence of extraordinary values (called outliers) and (2) that the accuracy of sample product-moments deteriorates rapidly with an increase in the order of the moments. An alternative type of moments, called L-moments, can be used to circumvent these disadvantages.

Example 2.8 (after Tung and Yen, 2005) Referring to Example 2.6, determine the first two moments about the origin for the time to failure of the pump. Then calculate the first two central moments.

Solution From Example 2.6, the random variable \( T \) is the time to failure having an exponential PDF as

\[
f_t(t) = \left(\frac{1}{\beta}\right) \exp \left(-\frac{t}{1250}\right) \quad \text{for } t \geq 0, \quad \beta > 0
\]

in which \( t \) is the elapsed time (in hours) before the pump fails, and \( \beta = 1250 \text{ h/failure} \).

The moments about the origin, according to Eq. (2.20a), are

\[E(T^r) = \mu'_r = \int_0^\infty t^r \left(\frac{e^{-t/\beta}}{\beta}\right) dt\]

Using integration by parts, the results of this integration are

for \( r = 1 \), \( \mu'_1 = E(T) = \mu_t = \beta = 1250 \text{ h} \)

for \( r = 2 \), \( \mu'_2 = E(T^2) = 2\beta^2 = 3,125,000 \text{ h}^2 \)

Based on the moments about the origin, the central moments can be determined, according to Eq. (2.22) or Problem (2.10), as

for \( r = 1 \), \( \mu_1 = E(T - \mu_t) = 0 \)

for \( r = 2 \), \( \mu_2 = E[(T - \mu_t)^2] = \mu'_2 - \mu^2 = 2\beta^2 - \beta^2 = \beta^2 = 1,562,500 \text{ h}^2 \)

L-moments. The \( r \)th-order L-moments are defined as (Hosking, 1986, 1990)

\[
\lambda_r = \frac{1}{r} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} E(X_{r-j:n}) \quad r = 1, 2, \ldots \quad \text{(2.24)}
\]

in which \( X_{j:n} \) is the \( j \)th-order statistic of a random sample of size \( n \) from the distribution \( F_x(x) \), namely, \( X_1 \leq X_2 \leq \cdots \leq X_{j} \leq \cdots \leq X_{n} \). The “L” in L-moments emphasizes that \( \lambda_r \) is a linear function of the expected order statistics. Therefore, sample L-moments can be made a linear combination of the ordered data values. The definition of the L-moments given in Eq. (2.24) may appear to be mathematically perplexing; the computations, however, can be simplified greatly through their relations with the probability-weighted moments,
which are defined as (Greenwood et al., 1979)

\[ M_{r,p,q} = E\{X^r[F_x(x)]^p[1-F_x(x)]^q\} = \int_{-\infty}^{\infty} x^r[F_x(x)]^p[1-F_x(x)]^q \, dF_x(x) \]  

(2.25)

Compared with Eq. (2.20a), one observes that the conventional product-moments are a special case of the probability-weighted moments with \( p = q = 0 \), that is, \( M_{r,0,0} = \mu'_r \). The probability-weighted moments are particularly attractive when the closed-form expression for the CDF of the random variable is available.

To work with the random variable linearly, \( M_{1,p,q} \) can be used. In particular, two types of probability-weighted moments are used commonly in practice, that is,

\[ \alpha_r = M_{1,0,r} = E\{X(1 - F_x(X))^r\} \quad r = 0, 1, 2, \ldots \]  

(2.26a)

\[ \beta_r = M_{1,r,0} = E\{X F_x(X)^r\} \quad r = 0, 1, 2, \ldots \]  

(2.26b)

In terms of \( \alpha_r \) or \( \beta_r \), the \( r \)th-order L-moment \( \lambda_r \) can be obtained as (Hosking, 1986)

\[ \lambda_{r+1} = (-1)^r \sum_{j=0}^{r} p_{r,j}^* \alpha_j = \sum_{j=0}^{r} p_{r,j}^* \beta_j \quad r = 0, 1, \ldots \]  

(2.27)

in which

\[ p_{r,j}^* = (-1)^{r-j} \binom{r}{j} \frac{(r+j)!}{(j!)^2(r-j)!} \]

For example, the first four L-moments of random variable \( X \) are

\[ \lambda_1 = \beta_0 = \mu'_1 = \mu_x \]  

(2.28a)

\[ \lambda_2 = 2\beta_1 - \beta_0 \]  

(2.28b)

\[ \lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0 \]  

(2.28c)

\[ \lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \]  

(2.28d)

To estimate sample \( \alpha \)- and \( \beta \)-moments, random samples are arranged in ascending or descending order. For example, arranging \( n \) random observations in ascending order, that is, \( X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(j)} \leq \cdots \leq X_{(n)} \), the \( r \)th-order \( \beta \)-moment \( \beta_r \) can be estimated as

\[ \hat{\beta}_r = \frac{1}{n} \sum_{i=1}^{n} X_{(i)} \hat{F}(X_{(i)})^r \]  

(2.29)

where \( \hat{F}(X_{(i)}) \) is an estimator for \( F(X_{(i)}) = P(X \leq X_{(i)}) \), for which many plotting-position formulas have been used in practice (Stedinger et al., 1993).
The one that is used often is the Weibull plotting-position formula, that is, 
\( \hat{F}(X_{(i)}) = \frac{i}{n+1} \).

L-moments possess several advantages over conventional product-moments. Estimators of L-moments are more robust against outliers and are less biased. They approximate asymptotic normal distributions more rapidly and closely. Although they have not been used widely in reliability applications as compared with the conventional product-moments, L-moments could have a great potential to improve reliability estimation. However, before more evidence becomes available, this book will limit its discussions to the uses of conventional product-moments.

Example 2.9 (after Tung and Yen, 2005) Referring to Example 2.8, determine the first two L-moments, that is, \( \lambda_1 \) and \( \lambda_2 \), of random time to failure \( T \).

Solution To determine \( \lambda_1 \) and \( \lambda_2 \), one first calculates \( \beta_0 \) and \( \beta_1 \), according to Eq. (2.26b), as

\[
\beta_0 = E(T[F_t(T)]^0) = E(T) = \mu_t = \beta
\]

\[
\beta_1 = E(T[F_t(T)]^1) = \int_0^\infty t F_t(t) f_t(t) \, dt = \int_0^\infty [t(1 - e^{-t/\beta})(e^{-t/\beta}/\beta)] \, dt = \frac{3}{4} \beta
\]

From Eq. (2.28), the first two L-moments can be computed as

\[
\lambda_1 = \beta_0 = \mu_t = \beta \quad \lambda_2 = 2\beta_1 - \beta_0 = \frac{6\beta}{4} - \beta = \frac{\beta}{2}
\]

2.4.2 Mean, mode, median, and quantiles

The central tendency of a continuous random variable \( X \) is commonly represented by its expectation, which is the first-order moment about the origin:

\[
E(X) = \mu_x = \int_{-\infty}^{\infty} x f_x(x) \, dx = \int_{-\infty}^{\infty} x \, dF_x(x) = \int_{-\infty}^{\infty} [1 - F_x(x)] \, dx \quad (2.30)
\]

This expectation is also known as the mean of a random variable. It can be seen easily that the mean of a random variable is the first-order L-moment \( \lambda_1 \). Geometrically, the mean or expectation of a random variable is the location of the centroid of the PDF or PMF. The second and third integrations in Eq. (2.30) indicate that the mean of a random variable is the shaded area shown in Fig. 2.11.

The following two operational properties of the expectation are useful:

1. The expectation of the sum of several random variables (regardless of their dependence) equals the sum of the expectation of the individual random
variable, that is,

\[ E \left( \sum_{k=1}^{K} a_k X_k \right) = \sum_{k=1}^{K} a_k \mu_k \] (2.31)

in which \( \mu_k = E(X_k) \), for \( k = 1, 2, \ldots, K \).

2. The expectation of multiplication of several independent random variables equals the product of the expectation of the individual random variables, that is,

\[ E \left( \prod_{k=1}^{K} X_k \right) = \prod_{k=1}^{K} \mu_k \] (2.32)

Two other types of measures of central tendency of a random variable, namely, the median and mode, are sometimes used in practice. The median of a random variable is the value that splits the distribution into two equal halves. Mathematically, the median \( x_{md} \) of a continuous random variable satisfies

\[ F_X(x_{md}) = \int_{-\infty}^{x_{md}} f_X(x) \, dx = 0.5 \] (2.33)

The median, therefore, is the 50th quantile (or percentile) of random variable \( X \). In general, the 100\( p \)th quantile of a random variable \( X \) is a quantity \( x_p \) that satisfies

\[ P(X \leq x_p) = F_X(x_p) = p \] (2.34)

The mode is the value of a random variable at which the value of a PDF is peaked. The mode \( x_{mo} \) of a random variable \( X \) can be obtained by solving the
following equation:

\[
\left[ \frac{\partial f_x(x)}{\partial x} \right]_{x = x_{mo}} = 0 \quad (2.35)
\]

Referring to Fig. 2.12, a PDF could be unimodal with a single peak, bimodal with two peaks, or multimodal with multiple peaks. Generally, the mean, median, and mode of a random variable are different unless the PDF is symmetric and unimodal. Descriptors for the central tendency of a random variable are summarized in Table 2.1.

**Example 2.10 (after Tung and Yen, 2005)** Refer to Example 2.8, the pump reliability problem. Find the mean, mode, median, and 10 percent quantile for the random time to failure \( T \).

**Solution** The mean of the time to failure, called the *mean time to failure* (MTTF), is the first-order moment about the origin, which is \( \mu_t = 1250 \text{ h} \) as calculated previously in Example 2.8. From the shape of the PDF for the exponential distribution as shown in Fig. 2.7, one can immediately identify that the mode, representing the most likely time of pump failure, is at the beginning of pump operation, that is, \( t_{mo} = 0 \text{ h} \).

\[f_x(x)\]

\[f_x(x)\]

**Figure 2.12** Unimodal (a) and bimodal (b) distributions.
To determine the median time to failure of the pump, one can first derive the expression for the CDF from the given exponential PDF as

\[ F_t(t) = P(T \leq t) = \int_0^t \frac{e^{-u/1250}}{1250} du = 1 - e^{-t/1250} \quad \text{for } t \geq 0 \]

in which \( u \) is a dummy variable. Then the median time to failure \( t_{\text{md}} \) can be obtained, according to Eq. (2.33), by solving

\[ F_t(t_{\text{md}}) = 1 - \exp(-t_{\text{md}}/1250) = 0.5 \]

which yields \( t_{\text{md}} = 866.43 \) h.

Similarly, the 10 percent quantile \( t_{0.1} \), namely, the elapsed time over which the pump would fail with a probability of 0.1, can be found in the same way as the median except that the value of the CDF is 0.1, that is,

\[ F_t(t_{0.1}) = 1 - \exp(-t_{0.1}/1250) = 0.1 \]

which yields \( t_{0.1} = 131.7 \) h.

### 2.4.3 Variance, standard deviation, and coefficient of variation

The spreading of a random variable over its range is measured by the **variance**, which is defined for the continuous case as

\[ \text{Var}(X) = \mu_2 = \sigma_x^2 = E[(X - \mu_x)^2] = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_x(x) \, dx \quad (2.36) \]

The variance is the second-order central moment. The positive square root of the variance is called the **standard deviation** \( \sigma_x \), which is often used as a measure of the degree of uncertainty associated with a random variable.

The standard deviation has the same units as the random variable. To compare the degree of uncertainty of two random variables with different units, a dimensionless measure \( \Omega_x = \sigma_x/\mu_x \), called the **coefficient of variation**, is useful. By its definition, the coefficient of variation indicates the variation of a random variable relative to its mean. Similar to the standard deviation, the second-order L-moment \( \lambda_2 \) is a measure of dispersion of a random variable. The ratio of \( \lambda_2 \) to \( \lambda_1 \), that is, \( \tau_2 = \lambda_2/\lambda_1 \), is called the **L-coefficient of variation**.

Three important properties of the variance are

1. \( \text{Var}(a) = 0 \) when \( a \) is a constant. \quad (2.37)
2. \( \text{Var}(X) = E(X^2) - E^2(X) = \mu'_2 - \mu_2^2 \) \quad (2.38)
3. The variance of the sum of several independent random variables equal the sum of variance of the individual random variables, that is,

\[ \text{Var} \left( \sum_{k=1}^{K} a_k X_k \right) = \sum_{k=1}^{K} a_k^2 \sigma_k^2 \quad (2.39) \]
where \( a_k \) is a constant, and \( \sigma_k \) is the standard deviation of random variable \( X_k, k = 1, 2, \ldots, K \).

**Example 2.11 (modified from Mays and Tung, 1992)** Consider the mass balance of a surface reservoir over a 1-month period. The end-of-month storage \( S \) can be computed as

\[
S_{m+1} = S_m + P_m + I_m - E_m - r_m
\]

in which the subscript \( m \) is an indicator for month, \( S_m \) is the initial storage volume in the reservoir, \( P_m \) is the precipitation amount on the reservoir surface, \( I_m \) is the surface-runoff inflow, \( E_m \) is the total monthly evaporation amount from the reservoir surface, and \( r_m \) is the controlled monthly release volume from the reservoir.

It is assumed that at the beginning of the month, the initial storage volume and total monthly release are known. The monthly total precipitation amount, surface-runoff inflow, and evaporation are uncertain and are assumed to be independent random variables. The means and standard deviations of \( P_m, I_m, \) and \( E_m \) from historical data for month \( m \) are estimated as

\[
\begin{align*}
E(P_m) &= 1000 \text{ m}^3, \\
E(I_m) &= 8000 \text{ m}^3, \\
E(E_m) &= 3000 \text{ m}^3 \\
\sigma(P_m) &= 500 \text{ m}^3, \\
\sigma(I_m) &= 2000 \text{ m}^3, \\
\sigma(E_m) &= 1000 \text{ m}^3
\end{align*}
\]

Determine the mean and standard deviation of the storage volume in the reservoir by the end of the month if the initial storage volume is 20,000 m\(^3\) and the designated release for the month is 10,000 m\(^3\).

**Solution** From Eq. (2.31), the mean of the end-of-month storage volume in the reservoir can be determined as

\[
E(S_{m+1}) = S_m + E(P_m) + E(I_m) - E(E_m) - r_m = 20,000 + 1000 + 8000 - 3000 - 10,000 = 16,000 \text{ m}^3
\]

Since the random hydrologic variables are statistically independent, the variance of the end-of-month storage volume in the reservoir can be obtained, from Eq. (2.39), as

\[
\text{Var}(S_{m+1}) = \text{Var}(P_m) + \text{Var}(I_m) + \text{Var}(E_m) = [(0.5)^2 + (2)^2 + (1)^2] \times (1000 \text{ m}^3)^2 = 5.25 \times (1000 \text{ m}^3)^2
\]

The standard deviation and coefficient of variation of \( S_{m+1} \) then are

\[
\begin{align*}
\sigma(S_{m+1}) &= \sqrt{5.25} \times 1000 = 2290 \text{ m}^3 \\
\Omega(S_{m+1}) &= 2290/16,000 = 0.143
\end{align*}
\]

### 2.4.4 Skewness coefficient and kurtosis

The asymmetry of the PDF of a random variable is measured by the *skewness coefficient* \( \gamma_3 \), defined as

\[
\gamma_3 = \frac{\mu_3}{\mu_2^{1.5}} = \frac{E[(X - \mu)^3]}{\sigma_x^3}
\]
The skewness coefficient is dimensionless and is related to the third-order central moment. The sign of the skewness coefficient indicates the degree of symmetry of the probability distribution function. If $\gamma_x = 0$, the distribution is symmetric about its mean. When $\gamma_x > 0$, the distribution has a long tail to the right, whereas $\gamma_x < 0$ indicates that the distribution has a long tail to the left. Shapes of distribution functions with different values of skewness coefficients and the relative positions of the mean, median, and mode are shown in Fig. 2.13.

![Figure 2.13](image)

**Figure 2.13** Relative locations of mean, median, and mode for (a) positively skewed, (b) symmetric and (c) negatively skewed distributions.
Similarly, the degree of asymmetry can be measured by the *L-skewness coefficient* $\tau_3$, defined as

$$\tau_3 = \lambda_3 / \lambda_2$$  \hspace{1cm} (2.41)

The value of the L-skewness coefficient for all feasible distribution functions must lie within the interval of $[-1, 1]$ (Hosking, 1986).

Another indicator of the asymmetry is the *Pearson skewness coefficient*, defined as

$$\gamma_1 = \frac{\mu_x - \mu_{mo}}{\sigma_x}$$  \hspace{1cm} (2.42)

As can be seen, the Pearson skewness coefficient does not require computing the third-order moment. In practice, product-moments higher than the third order are used less because they are unreliable and inaccurate when estimated from a small number of samples. Equations used to compute the sample product-moments are listed in the last column of Table 2.1.

*Kurtosis* $\kappa_x$ is a measure of the peakedness of a distribution. It is related to the fourth-order central moment of a random variable as

$$\kappa_x = \frac{\mu_4}{\mu_2^2} = \frac{E[(X - \mu_x)^4]}{\sigma_x^4}$$  \hspace{1cm} (2.43)

with $\kappa_x > 0$. For a random variable having a normal distribution (Sec. 2.6.1), its kurtosis is equal to 3. Sometimes the *coefficient of excess*, defined as $\varepsilon_x = \kappa_x - 3$, is used. For all feasible distribution functions, the skewness coefficient and kurtosis must satisfy the following inequality relationship (Stuart and Ord, 1987)

$$\gamma_x^2 + 1 \leq \kappa_x$$  \hspace{1cm} (2.44)

By the definition of L-moments, the *L-kurtosis* is defined as

$$\tau_4 = \lambda_4 / \lambda_2$$  \hspace{1cm} (2.45)

Similarly, the relationship between the L-skewness and L-kurtosis for all feasible probability distribution functions must satisfy (Hosking, 1986)

$$\frac{5\tau_3^2 - 1}{4} \leq \tau_4 < 1$$  \hspace{1cm} (2.46)

Royston (1992) conducted an analysis comparing the performance of sample skewness and kurtosis defined by the product-moments and L-moments. Results indicated that the L-skewness and L-kurtosis have clear advantages
over the conventional product-moments in terms of being easy to interpret, fairly robust to outliers, and less unbiased in small samples.

2.4.5 Covariance and correlation coefficient

When a problem involves two dependent random variables, the degree of linear dependence between the two can be measured by the correlation coefficient $\rho_{x,y}$, which is defined as

$$\text{Corr}(X, Y) = \rho_{x,y} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} \quad (2.47)$$

where $\text{Cov}(X, Y)$ is the covariance between random variables $X$ and $Y$, defined as

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E(XY) - \mu_x \mu_y \quad (2.48)$$

Various types of correlation coefficients have been developed in statistics for measuring the degree of association between random variables. The one defined by Eq. (2.47) is called the Pearson product-moment correlation coefficient, or correlation coefficient for short in this and general use.

It can be shown easily that $\text{Cov}(X_1', X_2') = \text{Corr}(X_1, X_2)$, with $X_1'$ and $X_2'$ being the standardized random variables. In probability and statistics, a random variable can be standardized as

$$X' = \frac{(X - \mu_x)}{\sigma_x} \quad (2.49)$$

Hence a standardized random variable has zero mean and unit variance. Standardization will not affect the skewness coefficient and kurtosis of a random variable because they are dimensionless.

Figure 2.14 graphically illustrates several cases of the correlation coefficient. If the two random variables $X$ and $Y$ are statistically independent, then $\text{Corr}(X, Y) = \text{Cov}(X, Y) = 0$ (Fig. 2.14c). However, the reverse statement is not necessarily true, as shown in Fig. 2.14d. If the random variables involved are not statistically independent, Eq. (2.70) for computing the variance of the sum of several random variables can be generalized as

$$\text{Var} \left( \sum_{k=1}^{K} a_k X_k \right) = \sum_{k=1}^{K} a_k^2 \sigma_k^2 + 2 \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} a_k a_{k'} \text{Cov}(X_k, X_{k'}) \quad (2.50)$$

Example 2.12 (after Tung and Yen, 2005) Perhaps the assumption of independence of $P_m, I_m,$ and $E_m$ in Example 2.11 may not be reasonable in reality. One examines the historical data closely and finds that correlations exist among the three hydrologic random variables. Analysis of data reveals that $\text{Corr}(P_m, I_m) = 0.8, \text{Corr}(P_m, E_m) = -0.4,$ and $\text{Corr}(I_m, E_m) = -0.3.$ Recalculate the standard deviation associated with the end-of-month storage volume.
Figure 2.14 Different cases of correlation between two random variables: (a) perfectly linearly correlated in opposite directions; (b) strongly linearly correlated in a positive direction; (c) uncorrelated in linear fashion; (d) perfectly correlated in nonlinear fashion but uncorrelated linearly.

**Solution** By Eq. (2.50), the variance of the reservoir storage volume at the end of the month can be calculated as

\[
\text{Var}(S_{m+1}) = ar(P_m) + \text{Var}(I_m) + \text{Var}(E_m) + 2\text{Cov}(P_m, I_m) \\
\quad - 2\text{Cov}(P_m, E_m) - 2\text{Cov}(I_m, E_m) \\
= \text{Var}(P_m) + \text{Var}(I_m) + \text{Var}(E_m) + 2\text{Corr}(P_m, I_m)\sigma(P_m)\sigma(I_m) \\
\quad - 2\text{Corr}(P_m, E_m)\sigma(P_m)\sigma(E_m) - 2\text{Corr}(I_m, E_m)\sigma(I_m)\sigma(E_m) \\
= (500)^2 + (2000)^2 + (1000)^2 + 2(0.8)(500)(2000) \\
\quad - 2(-0.4)(500)(1000) - 2(-0.3)(2000)(1000) \\
= 8.45(1000 \text{ m}^3)^2
\]
The corresponding standard deviation of the end-of-month storage volume is
\[ \sigma(S_{m+1}) = \sqrt{8.45 \times 1000} = 2910 \text{ m}^3 \]

In this case, consideration of correlation increases the standard deviation by 27 percent compared with the uncorrelated case in Example 2.11.

**Example 2.13** Referring to Example 2.7, compute correlation coefficient between \( X \) and \( Y \).

**Solution** Referring to Eqs. (2.47) and (2.48), computation of the correlation coefficient requires the determination of \( \mu_x, \mu_y, \sigma_x, \) and \( \sigma_y \) from the marginal PDFs of \( X \) and \( Y \):
\[ f_x(x) = \frac{4 + 3x^2}{16} \text{ for } 0 \leq x \leq 2 \]
\[ f_y(y) = \frac{4 + 3y^2}{16} \text{ for } 0 \leq y \leq 2 \]
as well as \( E(XY) \) from their joint PDF obtained earlier:
\[ f_{x,y}(x, y) = \frac{3(x^2 + y^2)}{32} \text{ for } 0 \leq x, y \leq 2 \]

From the marginal PDFs, the first two moments of \( X \) and \( Y \) about the origin can be obtained easily as
\[ \mu_x = E(X) = \int_0^2 x f_x(x) dx = \frac{5}{4} = E(Y) = \mu_y \]
\[ E(X^2) = \int_0^2 x^2 f_x(x) dx = \frac{28}{15} = E(Y^2) \]
Hence the variances of \( X \) and \( Y \) can be calculated as
\[ \text{Var}(X) = E(X^2) - (\mu_x)^2 = \frac{73}{140} \]
\[ \text{Var}(Y) = E(Y^2) - (\mu_y)^2 = \frac{73}{140} \]

To calculate \( \text{Cov}(X, Y) \), one could first compute \( E(XY) \) from the joint PDF as
\[ E(XY) = \int_0^2 \int_0^2 xy f_{x,y}(x, y) dx dy = \frac{3}{2} \]
Then the covariance of \( X \) and \( Y \), according to Eq. (2.48), is
\[ \text{Cov}(X, Y) = E(XY) - \mu_x \mu_y = -1/16 \]

The correlation between \( X \) and \( Y \) can be obtained as
\[ \text{Corr}(X, Y) = \rho_{x,y} = \frac{-1/16}{73/240} = -0.205 \]

### 2.5 Discrete Univariate Probability Distributions

In the reliability analysis of hydrosystems engineering problems, several probability distributions are used frequently. Based on the nature of the random variable, probability distributions are classified into discrete and continuous types. In this section, two discrete distributions, namely, the binomial distribution and the Poisson distribution, that are used commonly in hydrosystems reliability analysis, are described. Section 2.6 describes several frequently used univariate continuous distributions. For the distributions discussed in this chapter and others not included herein, their relationships are shown in Fig. 2.15.
Figure 2.15 Relationships among univariate distributions. (After Leemis, 1986.)
Computations of probability and quantiles for the great majority of the distribution functions described in Secs. 2.5 and 2.6 are available in Microsoft Excel.

2.5.1 Binomial distribution

The binomial distribution is applicable to random processes with only two types of outcomes. The state of components or subsystems in many hydrosystems can be classified as either functioning or failed, which is a typical example of a binary outcome. Consider an experiment involving a total of \( n \) independent trials with each trial having two possible outcomes, say, success or failure. In each trial, if the probability of having a successful outcome is \( p \), the probability of having \( x \) successes in \( n \) trials can be computed as

\[
p_x(x) = C_{n,x} p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n
\]

where \( C_{n,x} \) is the binomial coefficient, and \( q = 1 - p \), the probability of having a failure in each trial. Computationally, it is convenient to use the following recursive formula for evaluating the binomial PMF (Drane et al., 1993):

\[
p_x(x \mid n, p) = \left( \frac{n + 1 - x}{x} \right) \left( \frac{p}{q} \right) p_x(x - 1 \mid n, p) = R_B(x) p_x(x - 1 \mid n, p) \quad (2.52)
\]

for \( x = 0, 1, 2, \ldots, n \), with the initial probability \( p_x(x = 0 \mid n, p) = q^n \). A simple recursive scheme for computing the binomial cumulative probability is given by Tietjen (1994).

A random variable \( X \) having a binomial distribution with parameters \( n \) and \( p \) has the expectation \( E(X) = np \) and variance \( \text{Var}(X) = npq \). Shape of the PMF of a binomial random variable depends on the values of \( p \) and \( q \). The skewness coefficient of a binomial random variable is \( (q - p) / \sqrt{npq} \). Hence the PMF is positively skewed if \( p < q \), symmetric if \( p = q = 0.5 \), and negatively skewed if \( p > q \). Plots of binomial PMFs for different values of \( p \) with a fixed \( n \) are shown in Fig. 2.16. Referring to Fig. 2.15, the sum of several independent binomial random variables, each with a common parameter \( p \) and different \( n_k \)s, is still a binomial random variable with parameters \( p \) and \( \Sigma_k n_k \).

Example 2.14 A roadway-crossing structure, such as a bridge or a box or pipe culvert, is designed to pass a flood with a return period of 50 years. In other words, the annual probability that the roadway-crossing structure would be overtopped is a 1-in-50 chance or \( 1/50 = 0.02 \). What is the probability that the structure would be overtopped over an expected service life of 100 years?

Solution In this example, the random variable \( X \) is the number of times the roadway-crossing structure will be overtopped over a 100-year period. One can treat each year as an independent trial from which the roadway structure could be overtopped or not overtopped. Since the outcome of each “trial” is binary, the binomial distribution is applicable.
The event of interest is the overtopping of the roadway structure. The probability of such an event occurring in each trial (namely, each year), is 0.02. A period of 100 years represents 100 trials. Hence, in the binomial distribution model, the parameters are \( p = 0.02 \) and \( n = 100 \). The probability that overtopping occurs in a period of 100 years can be calculated, according to Eq. (2.51), as

\[
P(\text{overtopping occurs in an 100-year period}) = P(X \geq 1 | n = 100, p = 0.02) = \sum_{x=1}^{100} p_x(x) = \sum_{x=1}^{100} C_{100,x}(0.02)^x(0.98)^{100-x}
\]
This equation for computing the overtopping probability requires evaluations of 100 binomial terms, which could be very cumbersome. In this case, one could solve the problem by looking at the other side of the coin, i.e., the nonoccurrence of overtopping events. In other words,

\[
P(\text{overtopping occurs in a 100-year period}) = P(\text{overtopping occurs at least once in a 100-year period}) = 1 - P(\text{no overtopping occurs in a 100-year period}) = 1 - (0.98)^{100}
\]

Calculation of the overtopping risk, as illustrated in this example, is made under an implicit assumption that the occurrence of floods is a stationary process. In other words, the flood-producing random mechanism for the watershed under consideration does not change with time. For a watershed undergoing changes in hydrologic characteristics, one should be cautious about the estimated risk.

The preceding example illustrates the basic application of the binomial distribution to reliability analysis. A commonly used alternative is the Poisson distribution described in the next section. More detailed descriptions of these two distributions in time-dependent reliability analysis of hydrosystems infrastructural engineering are given in Sec. 4.7.

### 2.5.2 Poisson distribution

The **Poisson distribution** has the PMF as

\[
p_x(\nu) = \frac{e^{-\nu} \nu^x}{x!} \quad \text{for } x = 0, 1, 2, \ldots
\]

(2.53)

where the parameter \( \nu > 0 \) represents the mean of a Poisson random variable. Unlike the binomial random variables, Poisson random variables have no upper bound. A recursive formula for calculating the Poisson PMF is (Drane et al., 1993)

\[
p_x(\nu) = \left( \frac{\nu}{x} \right) p_{x-1}(\nu) = R_p(x)p_x(\nu) \quad \text{for } x = 1, 2, \ldots
\]

(2.54)

with \( p_0(\nu) = e^{-\nu} \) and \( R_p(x) = \nu/x \). When \( \nu \to \infty \) and \( \nu \to 0 \) while \( np = \nu = \text{constant} \), the term \( R_B(x) \) in Eq. (2.52) for the binomial distribution becomes \( R_p(x) \) for the Poisson distribution. Tietjen (1994) presents a simple recursive scheme for computing the Poisson cumulative probability.

For a Poisson random variable, the mean and the variance are identical to \( \nu \). Plots of Poisson PMFs corresponding to different values of \( \nu \) are shown in Fig. 2.17. As shown in Fig. 2.15, Poisson random variables also have the same reproductive property as binomial random variables. That is, the sum of several independent Poisson random variables, each with a parameter \( \nu_k \), is still a Poisson random variable with a parameter \( \nu_1 + \nu_2 + \cdots + \nu_k \). The skewness
Figure 2.17  Probability mass functions of Poisson random variables with different parameter values.

The coefficient of a Poisson random variable is $1/\sqrt{\nu}$, indicating that the shape of the distribution approaches symmetry as $\nu$ gets large.

The Poisson distribution has been applied widely in modeling the number of occurrences of a random event within a specified time or space interval. Equation (2.2) can be modified as

$$P_x(x | \lambda, t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$  \hspace{1cm} (2.55)

in which the parameter $\lambda$ can be interpreted as the average rate of occurrence of the random event in a time interval $(0, t)$. 
Example 2.15 Referring to Example 2.14, the use of binomial distribution assumes, implicitly, that the overtopping occurs, at most, once each year. The probability is zero for having more than two overtopping events annually. Relax this assumption and use the Poisson distribution to reevaluate the probability of overtopping during a 100-year period.

Solution Using the Poisson distribution, one has to determine the average number of overtopping events in a period of 100 years. For a 50-year event, the average rate of overtopping is \( \lambda = 0.02 \) \text{/year}. Therefore, the average number of overtopping events in a period of 100 years can be obtained as \( \nu = (0.02)(100) = 2 \) overtoppings. The probability of overtopping in an 100-year period, using a Poisson distribution, is

\[
P(\text{overtopping occurs in a 100-year period})
= P(\text{overtopping occurs at least once in a 100-year period})
= 1 - P(\text{no overtopping occurs in a 100-year period})
= 1 - p(X = 0 \mid \nu = 2) = 1 - e^{-2}
= 1 - 0.1353 = 0.8647
\]

Comparing with the result from Example 2.14, use of the Poisson distribution results in a slightly smaller risk of overtopping.

To relax the restriction of equality of the mean and variance for the Poisson distribution, Consul and Jain (1973) introduced the generalized Poisson distribution (GPD) having two parameters \( \theta \) and \( \lambda \) with the probability mass function as

\[
p_x(x \mid \theta, \lambda) = \frac{\theta(\theta + x\lambda)^{n-1}e^{-(\theta + x\lambda)}}{x!} \quad \text{for } x = 0, 1, 2, \ldots; \lambda \geq 0 \quad (2.56)
\]

The parameters \((\theta, \lambda)\) can be determined by the first two moments (Consul, 1989) as

\[
E(X) = \frac{\theta}{1 - \lambda} \quad \text{Var}(X) = \frac{\theta}{(1 - \lambda)^3} \quad (2.57)
\]

The variance of the GPD model can be greater than, equal to, or less than the mean depending on whether the second parameter \( \lambda \) is positive, zero, or negative. The values of the mean and variance of a GPD random variable tend to increase as \( \theta \) increases. The GPD model has greater flexibility to fit various types of random counting processes, such as binomial, negative binomial, or Poisson, and many other observed data.

2.6 Some Continuous Univariate Probability Distributions

Several continuous PDFs are used frequently in reliability analysis. They include normal, lognormal, gamma, Weibull, and exponential distributions. Other distributions, such as beta and extremal distributions, also are used sometimes.
The relations among the various continuous distributions considered in this chapter and others are shown in Fig. 2.15.

### 2.6.1 Normal (Gaussian) distribution

The normal distribution is a well-known probability distribution involving two parameters: the mean and variance. A normal random variable having the mean $\mu_x$ and variance $\sigma^2_x$ is denoted herein as $X \sim N(\mu_x, \sigma_x)$ with the PDF

$$f_{N_x}(x | \mu_x, \sigma^2_x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_x}{\sigma_x} \right)^2 \right]$$

for $-\infty < x < \infty$ (2.58)

The relationship between $\mu_x$ and $\sigma_x$ and the L-moments are $\mu_x = \lambda_1$ and $\sigma_x = \sqrt{\pi} \lambda_2$.

The normal distribution is bell-shaped and symmetric with respect to the mean $\mu_x$. Therefore, the skewness coefficient of a normal random variable is zero. Owing to the symmetry of the PDF, all odd-order central moments are zero. The kurtosis of a normal random variable is $\kappa_x = 3.0$. Referring to Fig. 2.15, a linear function of several normal random variables also is normal. That is, the linear combination of $K$ normal random variables $W = a_1X_1 + a_2X_2 + \cdots + a_KX_K$, with $X_k \sim N(\mu_k, \sigma_k)$, for $k = 1, 2, \ldots, K$, is also a normal random variable with the mean $\mu_w$ and variance $\sigma^2_w$, respectively, as

$$\mu_w = \sum_{k=1}^{K} a_k \mu_k$$
$$\sigma^2_w = \sum_{k=1}^{K} \sigma^2_k + 2 \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} a_k a_{k'} \text{Cov}(X_k, X_{k'})$$

The normal distribution sometimes provides a viable alternative to approximate the probability of a nonnormal random variable. Of course, the accuracy of such an approximation depends on how closely the distribution of the nonnormal random variable resembles the normal distribution. An important theorem relating to the sum of independent random variables is the central limit theorem, which loosely states that the distribution of the sum of a number of independent random variables, regardless of their individual distributions, can be approximated by a normal distribution, as long as none of the variables has a dominant effect on the sum. The larger the number of random variables involved in the summation, the better is the approximation. Because many natural processes can be thought of as the summation of a large number of independent component processes, none dominating the others, the normal distribution is a reasonable approximation for these overall processes. Finally, Dowson and Wragg (1973) have shown that when only the mean and variance are specified, the maximum entropy distribution on the interval $(-\infty, +\infty)$ is the normal distribution. That is, when only the first two moments are specified, the use of the normal distribution implies more information about the nature of the underlying process specified than any other distributions.

Probability computations for normal random variables are made by first transforming the original variable to a standardized normal variable $Z$ by
Eq. (2.49), that is,

\[ Z = \frac{(X - \mu_x)}{\sigma_x} \]

in which \( Z \) has a mean of zero and a variance of one. Since \( Z \) is a linear function of the normal random variable \( X \), \( Z \) is therefore normally distributed, that is, \( Z \sim N(\mu_z = 0, \sigma_z = 1) \). The PDF of \( Z \), called the standard normal distribution, can be obtained easily as

\[ \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad \text{for } -\infty < z < \infty \]  

(2.59)

The general expressions for the product-moments of the standard normal random variable are

\[ E(Z^{2r}) = \frac{(2r)!}{2^r r!} \quad \text{and} \quad E(Z^{2r+1}) = 0 \quad \text{for } r \geq 1 \]  

(2.60)

Computations of probability for \( X \sim N(\mu_x, \sigma_x) \) can be made as

\[ P(X \leq x) = P\left(\frac{X - \mu_x}{\sigma_x} \leq \frac{x - \mu_x}{\sigma_x}\right) = P(Z \leq z) = \Phi(z) \]  

where \( z = (x - \mu_x)/\sigma_x \), and \( \Phi(z) \) is the standard normal CDF defined as

\[ \Phi(z) = \int_{-\infty}^{z} \phi(z) \, dz \]  

(2.62)

Figure 2.18 shows the shape of the PDF of the standard normal random variable.

The integral result of Eq. (2.62) is not analytically available. A table of the standard normal CDF, such as Table 2.2 or similar, can be found in many statistics textbooks (Abramowitz and Stegun, 1972; Haan, 1977; Blank, 1980).

![Figure 2.18](image-url)  

**Figure 2.18** Probability density of the standard normal variable.
For numerical computation purposes, several highly accurate approximations are available for determining $\Phi(z)$. One such approximation is the polynomial approximation (Abramowitz and Stegun, 1972)

$$
\Phi(z) = 1 - \phi(z) \left( b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5 \right)
$$  \hspace{1cm} \text{for } z \geq 0 \hspace{1cm} (2.63)

where $t = \frac{z}{\sqrt{2}}$, and $b_1 = -0.31938153$, $b_2 = -0.356563782$, $b_3 = 1.781477937$, $b_4 = -1.821255978$, and $b_5 = 1.33027443$. The maximum absolute error of the approximation is $7.5 \times 10^{-8}$, which is sufficiently accurate for most practical applications. Note that Eq. (2.63) is applicable to the non-negative-valued $z$. For $z < 0$, the value of standard normal CDF can be computed as $\Phi(z) = 1 - \Phi(|z|)$ by the symmetry of $\phi(z)$.

Approximation equations, such as
Eq. (2.63), can be programmed easily for probability computations without needing the table of the standard normal CDF.

Equally practical is the inverse operation of finding the standard normal quantile \( z_p \) with the specified probability level \( p \). The standard normal CDF table can be used, along with some mechanism of interpolation, to determine \( z_p \). However, for practical algebraic computations with a computer, the following rational approximation can be used (Abramowitz and Stegun, 1972):

\[
z_p = t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3} \quad \text{for } 0.5 < p \leq 1 \quad (2.64)
\]

in which \( p = \Phi(z_p) \), \( t = \sqrt{-2 \ln(1 - p)} \), \( c_0 = 2.515517 \), \( c_1 = 0.802853 \), \( c_2 = 0.010328 \), \( d_1 = 1.432788 \), \( d_2 = 0.189269 \), and \( d_3 = 0.001308 \). The corresponding maximum absolute error by this rational approximation is \( 4 \times 10^{-4} \). Note that Eq. (2.64) is valid for the value of \( \Phi(z_p) \) that lies between \([0.5, 1]\). When \( p < 0.5 \), one can still use Eq. (2.64) by letting \( t = \sqrt{-2 \times \ln(p)} \) and attaching a negative sign to the computed quantile value. Vedder (1995) proposed a simple approximation for computing the standard normal cumulative probabilities and standard normal quantiles.

**Example 2.16** Referring to Example 2.14, determine the probability of more than five overtopping events over a 100-year period using a normal approximation.

**Solution** In this problem, the random variable \( X \) of interest is the number of overtopping events in a 100-year period. The exact distribution of \( X \) is binomial with parameters \( n = 100 \) and \( p = 0.02 \) or the Poisson distribution with a parameter \( \nu = 2 \). The exact probability of having more than five occurrences of overtopping in 100 years can be computed as

\[
P(X > 5) = P(X \geq 6) = \sum_{x=6}^{100} \binom{100}{x} (0.02)^x (0.98)^{100-x}
\]

where

\[
P(X \leq 5) = 1 - P(X > 5) = 1 - \sum_{x=6}^{5} \binom{100}{x} (0.02)^x (0.98)^{100-x}
\]

Thus

\[
P(X > 5) = 1 - 0.9845 = 0.0155
\]

As can be seen, there are a total of six terms to be summed up on the right-hand side. Although the computation of probability by hand is within the realm of a reasonable task, the following approximation is viable. Using a normal probability approximation, the mean and variance of \( X \) are

\[
\mu_x = np = (100)(0.02) = 2.0 \quad \sigma_x^2 = npq = (100)(0.02)(0.98) = 1.96
\]

The preceding binomial probability can be approximated as

\[
P(X \geq 6) = 1 - P(X < 5.5) = 1 - P[Z < (5.5 - 2.0)/\sqrt{1.96}] = 1 - \Phi(2.5) = 1 - 0.9938 = 0.062
\]
DeGroot (1975) showed that when \( np^{1.5} > 1.07 \), the error of using the normal distribution to approximate the binomial probability did not exceed 0.05. The error in the approximation gets smaller as the value of \( np^{1.5} \) becomes larger. For this example, \( np^{1.5} = 0.283 \leq 1.07 \), and the accuracy of approximation was not satisfactory as shown.

Example 2.17 (adopted from Mays and Tung, 1992) The annual maximum flood magnitude in a river has a normal distribution with a mean of 6000 ft\(^3\)/s and standard deviation of 4000 ft\(^3\)/s. (a) What is the annual probability that the flood magnitude would exceed 10,000 ft\(^3\)/s? (b) Determine the flood magnitude with a return period of 100 years.

Solution (a) Let \( Q \) be the random annual maximum flood magnitude. Since \( Q \) has a normal distribution with a mean \( \mu_Q = 6000 \) ft\(^3\)/s and standard deviation \( \sigma_Q = 4000 \) ft\(^3\)/s, the probability of the annual maximum flood magnitude exceeding 10,000 ft\(^3\)/s is

\[
P(Q > 10,000) = 1 - P[Z \leq (10,000 - 6000)/4000] = 1 - \Phi(1.00) = 1 - 0.8413 = 0.1587
\]

(b) A flood event with a 100-year return period represents the event the magnitude of which has, on average, an annual probability of 0.01 being exceeded. That is, \( P(Q \geq q_{100}) = 0.01 \), in which \( q_{100} \) is the magnitude of the 100-year flood. This part of the problem is to determine \( q_{100} \) from

\[
P(Q \leq q_{100}) = 1 - P(Q \geq q_{100}) = 0.99
\]

because

\[
P(Q \leq q_{100}) = P[Z \leq (q_{100} - \mu_Q)/\sigma_Q]) = P[Z \leq (q_{100} - 6000)/4000] = \Phi(q_{100} - 6000)/4000] = 0.99
\]

From Table 2.2 or Eq. (2.64), one can find that \( \Phi(2.33) = 0.99 \). Therefore,

\[
(q_{100} - 6000)/4000 = 2.33
\]

which gives that the magnitude of the 100-year flood event as \( q_{100} = 15,320 \) ft\(^3\)/s.

2.6.2 Lognormal distribution

The lognormal distribution is a commonly used continuous distribution for positively valued random variables. Lognormal random variables are closely related to normal random variables, by which a random variable \( X \) has a lognormal distribution if its logarithmic transform \( Y = \ln(X) \) has a normal distribution with mean \( \mu_{\ln x} \) and variance \( \sigma_{\ln x}^2 \). From the central limit theorem, if a natural process can be thought of as a multiplicative product of a large number of an independent component processes, none dominating the others, the lognormal
distribution is a reasonable approximation for these natural processes. The PDF of a lognormal random variable is

\[ f_{LN}(x | \mu_{ln,x}, \sigma_{ln,x}^2) = \frac{1}{\sqrt{2\pi \sigma_{ln,x}^2}} \exp \left\{ -\frac{1}{2} \left[ \frac{\ln(x) - \mu_{ln,x}}{\sigma_{ln,x}} \right]^2 \right\} \quad \text{for } x > 0 \quad (2.65) \]

which can be derived from the normal PDF. Statistical properties of a lognormal random variable in the original scale can be computed from those of log-transformed variables as

\[ \mu_x = \lambda_1 = \exp \left( \mu_{ln,x} + \frac{\sigma_{ln,x}^2}{2} \right) \quad (2.66a) \]

\[ \sigma_{x}^2 = \mu_x^2 \exp \left( \sigma_{ln,x}^2 - 1 \right) \quad (2.66b) \]

\[ \Omega_x^2 = \exp \left( \sigma_{ln,x}^2 \right) - 1 \quad (2.66c) \]

\[ \gamma_x = \Omega_x^3 + 3\Omega_x \quad (2.66d) \]

From Eq. (2.66d), one realizes that the shape of a lognormal PDF is always positively skewed (Fig. 2.19). Equations (2.66a) and (2.66b) can be derived easily by the moment-generating function (Tung and Yen, 2005, Sec. 4.2). Conversely, the statistical moments of \( \ln(X) \) can be computed from those of \( X \) by

\[ \mu_{ln,x} = \frac{1}{2} \ln \left[ \frac{\mu_x^2}{1 + \Omega_x^2} \right] = \ln(\mu_x) - \frac{1}{2} \sigma_{ln,x}^2 \quad (2.67a) \]

\[ \sigma_{ln,x}^2 = \ln \left( 1 + \Omega_x^2 \right) \quad (2.67b) \]

It is interesting to note from Eq. (2.67b) that the variance of a log-transformed variable is dimensionless.

In terms of the L-moments, the second-order L-moment for a two- and three-parameter lognormal distribution is (Stedinger et al., 1993)

\[ \lambda_2 = \exp \left( \mu_{ln,x} + \frac{\sigma_{ln,x}^2}{2} \right) \text{erf} \left( \frac{\sigma_{ln,x}}{\sqrt{2}} \right) = \exp \left( \mu_{ln,x} + \frac{\sigma_{ln,x}^2}{2} \right) \left[ 2\Phi \left( \frac{\sigma_{ln,x}}{\sqrt{2}} \right) - 1 \right] \quad (2.68) \]

in which \( \text{erf}(\cdot) \) is an error function the definitional relationship of which, with \( \Phi(x) \) is

\[ \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} \, du = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2/2} \, dz = 2\Phi(\sqrt{2}x) - 1 \quad (2.69) \]

Hence the L-coefficient of variation is \( \tau_2 = 2\Phi(\sigma_{ln,x}/\sqrt{2}) - 1 \). The relationship between the third- and fourth-order L-moment ratios can be approximated by the following polynomial function with accuracy within \( 5 \times 10^{-4} \) for \( |\tau_2| < 0.9 \) (Hosking, 1991):

\[ \tau_4 = 0.12282 + 0.77518\tau_2^2 + 0.12279\tau_2^3 - 0.13638\tau_2^4 + 0.11386\tau_2^5 \quad (2.70) \]
Since the sum of normal random variables is normally distributed, the product of lognormal random variables also is lognormally distributed (see Fig. 2.15). This useful reproductive property of lognormal random variables can be stated as if $X_1, X_2, \ldots, X_K$ are independent lognormal random variables, then $W = b_0 \prod_{k=1}^{K} X_k$ has a lognormal distribution with mean and variance as

$$
\mu_{\ln w} = \ln(b_0) + \sum_{k=1}^{K} b_k \mu_{\ln x_k} \quad \sigma_{\ln w}^2 = \sum_{k=1}^{K} b_k^2 \sigma_{\ln x_k}^2
$$

In the case that two lognormal random variables are correlated with a correlation coefficient $\rho_{x,y}$ in the original scale, then the covariance terms in the
log-transformed space must be included in calculating \( \sigma^2_{\ln w} \). Given \( \rho_{x,y} \), the correlation coefficient in the log-transformed space can be computed as

\[
\text{Corr}(\ln X, \ln Y) = \rho_{\ln x, \ln y} = \frac{\ln(1 + \rho_{x,y}\Omega_x\Omega_y)}{\sqrt{\ln(1 + \Omega_x^2) \times \ln(1 + \Omega_y^2)}}
\] (2.71)

Derivation of Eq. (2.71) can be found in Tung and Yen (2005).

**Example 2.18** Re-solve Example 2.17 by assuming that the annual maximum flood magnitude in the river follows a lognormal distribution.

**Solution** (a) Since \( Q \) has a lognormal distribution, \( \ln Q \) is normally distributed with a mean and variance that can be computed from Eqs. (2.67a) and (2.67b), respectively, as

\[
\Omega_Q = \frac{4000}{6000} = 0.667
\]

\[
\sigma^2_{\ln Q} = \ln(1 + 0.667^2) = 0.368
\]

\[
\mu_{\ln Q} = \ln(6000) - \frac{0.368}{2} = 8.515
\]

The probability of the annual maximum flood magnitude exceeding 10,000 ft\(^3\)/s is

\[
P(Q > 10,000) = P[\ln Q > \ln(10,000)] = 1 - P[Z \leq (9.210 - 8.515)/\sqrt{0.368}]
\]

\[
= 1 - \Phi(1.146) = 1 - 0.8741 = 0.1259
\]

(b) A 100-year flood \( q_{100} \) represents the event the magnitude of which corresponds to \( P(Q \geq q_{100}) = 0.01 \), which can be determined from

\[
P(Q \leq q_{100}) = 1 - P(Q \geq q_{100}) = 0.99
\]

because

\[
P(Q \leq q_{100}) = P[\ln Q \leq \ln(q_{100})] = P[Z \leq (\ln(q_{100}) - \mu_{\ln Q})/\sigma_{\ln Q}]
\]

\[
= P[Z \leq (\ln(q_{100}) - 8.515)/\sqrt{0.368}] = \Phi(\ln(q_{100}) - 8.515)/\sqrt{0.368} = 0.99
\]

From Table 2.2 or Eq. (2.64), one can find that \( \Phi(2.33) = 0.99 \). Therefore,

\[
[\ln(q_{100}) - 8.515]/\sqrt{0.368} = 2.33
\]

which yields \( \ln(q_{100}) = 9.9284 \). The magnitude of the 100-year flood event then is

\[
q_{100} = \exp(9.9284) = 20,500 \text{ ft}^3/\text{s}.
\]

### 2.6.3 Gamma distribution and variations

The **gamma distribution** is a versatile continuous distribution associated with a positive-valued random variable. The **two-parameter gamma distribution** has
a PDF defined as

\[
   f_G(x | \alpha, \beta) = \frac{1}{\beta \Gamma(\alpha)} (x/\beta)^{\alpha-1} e^{x/\beta} \quad \text{for } x > 0
\]

(2.72)

in which \( \beta > 0 \) and \( \alpha > 0 \) are the parameters and \( \Gamma(\bullet) \) is a gamma function defined as

\[
   \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \, dt
\]

(2.73)

The mean, variance, and skewness coefficient of a gamma random variable having a PDF as Eq. (2.72) are

\[
   \mu_x = \lambda_1 = \alpha \beta \quad \sigma_x^2 = \alpha \beta^2 \quad \gamma_x = 2/\sqrt{\alpha}
\]

(2.74)

In terms of L-moments, the second-order L-moment is

\[
   \lambda_2 = \frac{\beta \Gamma(\alpha + 0.5)}{\sqrt{\pi} \Gamma(\alpha)}
\]

(2.75)

and the relationship between the third- and fourth-order L-moment ratios can be approximated as (Hosking, 1991)

\[
   \tau_4 = 0.1224 + 0.30115 \tau_3^2 + 0.95812 \tau_3^4 - 0.57488 \tau_3^6 + 0.19383 \tau_3^8
\]

(2.76)

In the case that the lower bound of a gamma random variable is a positive quantity, the preceding two-parameter gamma PDF can be modified to a three-parameter gamma PDF as

\[
   f_G(x | \xi, \alpha, \beta) = \frac{1}{\beta \Gamma(\alpha)} \left( \frac{x - \xi}{\beta} \right)^{\alpha-1} e^{-(x-\xi)/\beta} \quad \text{for } x > \xi
\]

(2.77)

where \( \xi \) is the lower bound. The two-parameter gamma distribution can be reduced to a simpler form by letting \( Y = X/\beta \), and the resulting one-parameter gamma PDF (called the standard gamma distribution) is

\[
   f_G(y | \alpha) = \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y} \quad \text{for } y > 0
\]

(2.78)

Tables of the cumulative probability of the standard gamma distribution can be found in Dudewicz (1976). Shapes of some gamma distributions are shown in Fig. 2.20 to illustrate its versatility. If \( \alpha \) is a positive integer in Eq. (2.78), the distribution is called an Erlang distribution.

When \( \alpha = 1 \), the two-parameter gamma distribution reduces to an exponential distribution with the PDF

\[
   f_{\text{EXP}}(x | \beta) = e^{-x/\beta} / \beta \quad \text{for } x > 0
\]

(2.79)

An exponential random variable with a PDF as Eq. (2.79) has the mean and standard deviation equal to \( \beta \) (see Example 2.8). Therefore, the coefficient of
variation of an exponential random variable is equal to unity. The exponential
distribution is used commonly for describing the life span of various electronic
and mechanical components. It plays an important role in reliability mathem-
tics using time-to-failure analysis (see Chap. 5).

Two variations of the gamma distribution are used frequently in hydrologic
frequency analysis, namely, the Pearson and log-Pearson type 3 distributions.
In particular, the log-Pearson type 3 distribution is recommended for use by
the U.S. Water Resources Council (1982) as the standard distribution for flood
frequency analysis. A Pearson type 3 random variable has the PDF

\[
f_{P3}(x \mid \xi, \alpha, \beta) = \frac{1}{|\beta| \Gamma(\alpha)} \left( \frac{x - \xi}{\beta} \right)^{\alpha - 1} e^{-\left[\frac{\ln(x) - \xi}{\beta}\right]^{\alpha - 1}} e^{-\left[\frac{\ln(x) - \xi}{\beta}\right]}
\]

(2.80)

with \( \alpha > 0, x \geq \xi \) when \( \beta > 0 \) and with \( \alpha > 0, x \leq \xi \) when \( \beta < 0 \). When \( \beta > 0 \),
the Pearson type 3 distribution is identical to the three-parameter gamma
distribution. However, the Pearson type 3 distribution has the flexibility to
model negatively skewed random variables corresponding to \( \beta < 0 \). Therefore,
the skewness coefficient of the Pearson type 3 distribution can be computed,
from modifying Eq. (2.74), as \( \text{sign}(\beta)2/\sqrt{\alpha} \).

Similar to the normal and lognormal relationships, the PDF of a log-Pearson
type 3 random variable is

\[
f_{LP3}(x \mid \xi, \alpha, \beta) = \frac{1}{x|\beta| \Gamma(\alpha)} \left[ \frac{\ln(x) - \xi}{\beta} \right]^{\alpha - 1} e^{-\left[\frac{\ln(x) - \xi}{\beta}\right]}
\]

(2.81)

with \( \alpha > 0, x \geq e^\xi \) when \( \beta > 0 \) and with \( \alpha > 0, x \leq e^\xi \) when \( \beta < 0 \). Numerous
studies can be found in the literature about Pearson type 3 and log-Pearson
type 3 distributions. Kite (1977), Stedinger et al. (1993), and Rao and Hamed (2000) provide good summaries of these two distributions.

Evaluation of the probability of gamma random variables involves computations of the gamma function, which can be made by using the following recursive formula:

\[
\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)
\]

When the argument \( \alpha \) is an integer number, then \( \Gamma(\alpha) = (\alpha - 1)! \). However, when \( \alpha \) is a real number, the recursive relation would lead to \( \Gamma(\alpha') \) as the smallest term, with \( 1 < \alpha' < 2 \). The value of \( \Gamma(\alpha') \) can be determined by a table of the gamma function or by numerical integration on Eq. (2.73). Alternatively, the following approximation could be applied to accurately estimate the value of \( \Gamma(\alpha') \) (Abramowitz and Stegun, 1972):

\[
\Gamma(\alpha') = \Gamma(x + 1) = 1 + \sum_{i=1}^{5} a_i x^i \quad \text{for } 0 < x < 1
\]

in which \( a_1 = -0.577191652 \), \( a_2 = 0.988205891 \), \( a_3 = -0.897056937 \), \( a_4 = 0.4245549 \), and \( a_5 = -0.1010678 \). The maximum absolute error associated with Eq. (2.83) is \( 5 \times 10^{-5} \).

### 2.6.4 Extreme-value distributions

Hydrosystems engineering reliability analysis often focuses on the statistical characteristics of extreme events. For example, the design of flood-control structures may be concerned with the distribution of the largest events over the recorded period. On the other hand, the establishment of a drought-management plan or water-quality management scheme might be interested in the statistical properties of minimum flow over a specified period. Statistics of extremes are concerned with the statistical characteristics of \( X_{\text{max},n} = \max\{X_1, X_2, \ldots, X_n\} \) and/or \( X_{\text{min},n} = \min\{X_1, X_2, \ldots, X_n\} \) in which \( X_1, X_2, \ldots, X_n \) are observations of random processes. In fact, the exact distributions of extremes are functions of the underlying (or parent) distribution that generates the random observations \( X_1, X_2, \ldots, X_n \) and the number of observations. Of practical interest are the asymptotic distributions of extremes. Asymptotic distribution means that the resulting distribution is the limiting form of \( F_{\text{max},n}(y) \) or \( F_{\text{min},n}(y) \) as the number of observations \( n \) approaches infinity. The asymptotic distributions of extremes turn out to be independent of the sample size \( n \) and the underlying distribution for random observations. That is,

\[
\lim_{n \to \infty} F_{\text{max},n}(y) = F_{\text{max}}(y) \quad \lim_{n \to \infty} F_{\text{min},n}(y) = F_{\text{min}}(y)
\]

Furthermore, these asymptotic distributions of the extremes largely depend on the tail behavior of the parent distribution in either direction toward the extremes. The center portion of the parent distribution has little significance for defining the asymptotic distributions of extremes. The work on statistics of extremes was pioneered by Fisher and Tippett (1928) and later was extended
Fundamentals of Probability and Statistics for Reliability Analysis  67

by Gnedenko (1943). Gumbel (1958), who dealt with various useful applications of $X_{\text{max},n}$ and $X_{\text{min},n}$ and other related issues.

Three types of asymptotic distributions of extremes are derived based on the different characteristics of the underlying distribution (Haan, 1977):

**Type I.** Parent distributions are unbounded in the direction of extremes, and all statistical moments exist. Examples of this type of parent distribution are normal (for both largest and smallest extremes), lognormal, and gamma distributions (for the largest extreme).

**Type II.** Parent distributions are unbounded in the direction of extremes, but all moments do not exist. One such distribution is the Cauchy distribution (Sec. 2.6.5). Thus the type II extremal distribution has few applications in practical engineering analysis.

**Type III.** Parent distributions are bounded in the direction of the desired extreme. Examples of this type of underlying distribution are the beta distribution (for both largest and smallest extremes) and the lognormal and gamma distributions (for the smallest extreme).

Owing to the fact that $X_{\text{min},n} = -\max\{-X_1, -X_2, \ldots, -X_n\}$, the asymptotic distribution functions of $X_{\text{max},n}$ and $X_{\text{min},n}$ satisfy the following relation (Leadbetter et al., 1983):

$$F_{\text{min}}(y) = 1 - F_{\text{max}}(-y)$$  \hspace{1cm} (2.84)

Consequently, the asymptotic distribution of $X_{\text{min}}$ can be obtained directly from that of $X_{\text{max}}$. Three types of asymptotic distributions of the extremes are listed in Table 2.3.

**Extreme-value type I distribution.** This is sometimes referred to as the Gumbel distribution, Fisher-Tippett distribution, and double exponential distribution. The CDF and PDF of the extreme-value type I (EV1) distribution have, respectively, the following forms:

$$F_{\text{EV1}}(x | \xi, \beta) = \exp \left\{ - \exp \left[ - \left( \frac{x - \xi}{\beta} \right) \right] \right\} \quad \text{for maxima}$$  \hspace{1cm} (2.85a)

$$= 1 - \exp \left\{ - \exp \left[ + \left( \frac{x - \xi}{\beta} \right) \right] \right\} \quad \text{for minima}$$

**TABLE 2.3 Three Types of Asymptotic Cumulative Distribution Functions (CDFs) of Extremes**

<table>
<thead>
<tr>
<th>Type</th>
<th>Maxima</th>
<th>Range</th>
<th>Minima</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$\exp(-e^{-y})$</td>
<td>$-\infty &lt; y &lt; \infty$</td>
<td>$1 - \exp(-e^y)$</td>
<td>$-\infty &lt; y &lt; \infty$</td>
</tr>
<tr>
<td>II</td>
<td>$\exp(-y^\alpha)$</td>
<td>$\alpha &lt; 0$, $y &gt; 0$</td>
<td>$1 - \exp[-(-y)^\alpha]$</td>
<td>$\alpha &lt; 0$, $y &lt; 0$</td>
</tr>
<tr>
<td>III</td>
<td>$\exp[-(-y)^\alpha]$</td>
<td>$\alpha &gt; 0$, $y &lt; 0$</td>
<td>$1 - \exp(-y^\alpha)$</td>
<td>$\alpha &gt; 0$, $y &gt; 0$</td>
</tr>
</tbody>
</table>

**NOTE:** $y = (x - \xi)/\beta$. 
 Chapter Two

Figure 2.21 Probability density function of extreme-value type I random variables.

\[ f_{EV1}(x \mid \xi, \beta) = \frac{1}{\beta} \exp \left\{ -\left(\frac{x - \xi}{\beta}\right) - \exp \left[ -\left(\frac{x - \xi}{\beta}\right) \right] \right\} \] for maxima

\[ = \frac{1}{\beta} \exp \left\{ +\left(\frac{x - \xi}{\beta}\right) - \exp \left[ +\left(\frac{x - \xi}{\beta}\right) \right] \right\} \] for minima

for \(-\infty < x, \xi < \infty\), and \(\beta \geq 0\). The shapes of the EV1 distribution are shown in Fig. 2.21, in which transformed random variable \(Y = (X - \xi)/\beta\) is used. As can be seen, the PDF associated with the largest extreme is a mirror image of the smallest extreme with respect to the vertical line passing through the common mode, which happens to be the parameter \(\xi\). The first three product-moments of an EV1 random variable are

\[ \mu_x = \lambda_1 = \xi + 0.5772\beta \] for the largest extreme
\[ = \xi - 0.5772\beta \] for the smallest extreme

\[ \sigma_x^2 = 1.645\beta^2 \] for both types

\[ \gamma_x = 1.13955 \] for the largest extreme
\[ = -1.13955 \] for the smallest extreme

The second- to fourth-order L-moments of the EV1 distribution for maxima are

\[ \lambda_2 = \beta \ln(2) \] \(\tau_3 = 0.1699\) \(\tau_4 = 0.1504\) (2.87)

Using the transformed variable \(Y = (X - \xi)/\beta\), the CDFs of the EV1 for the maxima and minima are shown in Table 2.3. Shen and Bryson (1979) showed
that if a random variable had an EV1 distribution, the following relationship is satisfied when $\xi$ is small:

$$x_{T_1} \approx \left[ \frac{\ln(T_1)}{\ln(T_2)} \right] x_{T_2}$$  \hspace{1cm} (2.88)

where $x_T$ is the quantile corresponding to the exceedance probability of $1/T$.

**Example 2.19** Repeat Example 2.17 by assuming that the annual maximum flood follows the EV1 distribution.

**Solution** Based on the values of a mean of 6000 ft$^3$/s and standard deviation of 4000 ft$^3$/s, the values of distributional parameters $\xi$ and $\beta$ can be determined as follows. For maxima, $\beta$ is computed from Eq. (2.86b) as

$$\beta = \frac{\sigma_Q}{\sqrt{1.645}} = \frac{4000}{1.2826} = 3118.72 \text{ ft}^3/\text{s}$$

and from Eq. (2.86a), one has

$$\xi = \mu_Q - 0.577\beta = 6000 - 0.577(3118.72) = 4200.50 \text{ ft}^3/\text{s}$$

(a) The probability of exceeding 10,000 ft$^3$/s, according to Eq. (2.85a), is

$$P(Q > 10,000) = 1 - F_{EV1}(10,000)$$

$$= 1 - \exp \left[ -\exp \left( -\frac{10,000 - 4200.50}{3118.72} \right) \right]$$

$$= 1 - \exp(-1.860)$$

$$= 1 - 0.8558 = 0.1442$$

(b) On the other hand, the magnitude of the 100-year flood event can be calculated as

$$y_{100} = q_{100} - \xi = -\ln\left[ -\ln(1 - 0.01) \right] = 4.60$$

Hence $q_{100} = 4200.50 + 4.60(3118.7) = 18,550 \text{ ft}^3/\text{s}$.

**Extreme-value type III distribution.** For the extreme-value type III (EV3) distribution, the corresponding parent distributions are bounded in the direction of the desired extreme (see Table 2.3). For many hydrologic and hydraulic random variables, the lower bound is zero, and the upper bound is infinity. For this reason, the EV3 distribution for the maxima has limited applications. On the other hand, the EV3 distribution of the minima is used widely for modeling the smallest extremes, such as drought or low-flow condition. The EV3 distribution for the minima is also known as the *Weibull distribution*, having a PDF defined as

$$f_W(x \mid \xi, \alpha, \beta) = \frac{\alpha}{\beta} \left( \frac{x - \xi}{\beta} \right)^{\alpha-1} \exp \left[ -\left( \frac{x - \xi}{\beta} \right)^{\alpha} \right]$$

for $x \geq \xi$ and $\alpha, \beta > 0$  \hspace{1cm} (2.89)
When $\xi = 0$ and $\alpha = 1$, the Weibull distribution reduces to the exponential distribution. Figure 2.22 shows that the versatility of the Weibull distribution function depends on the parameter values. The CDF of Weibull random variables can be derived as

$$F_W(x | \xi, \alpha, \beta) = 1 - \exp \left[ - \left( \frac{x - \xi}{\beta} \right)^\alpha \right]$$ (2.90)

The mean and variance of a Weibull random variable can be derived as

$$\mu_x = \lambda_1 = \xi + \beta \Gamma \left( 1 + \frac{1}{\beta} \right)$$ (2.91a)

$$\sigma_x^2 = \beta^2 \left[ \Gamma \left( 1 + \frac{2}{\alpha} \right) - \Gamma^2 \left( 1 + \frac{1}{\alpha} \right) \right]$$ (2.91b)

and the second-order L-moment is

$$\lambda_2 = \beta (1 - 2^{-1/\alpha}) \Gamma \left( 1 + \frac{1}{\alpha} \right)$$ (2.92)

Generalized extreme-value distribution. The generalized extreme-value (GEV) distribution provides an expression that encompasses all three types of extreme-value distributions. The CDF of a random variable corresponding to the maximum with a GEV distribution is

$$F_{GEV}(x | \xi, \alpha, \beta) = \exp \left\{ - \left[ 1 - \frac{\alpha(x - \xi)}{\beta} \right]^{1/\alpha} \right\} \quad \text{for } \alpha \neq 0$$ (2.93)

Figure 2.22  Probability density functions of a Weibull random variable.
When \( \alpha = 0 \), Eq. (2.93) reduces to Eq. (2.85a) for the Gumbel distribution. For \( \alpha < 0 \), it corresponds to the EV2 distribution having a lower bound \( x > \xi + \beta/\alpha \), whereas, on the other hand, for \( \alpha > 0 \), it corresponds to the EV3 distribution having an upper bound \( x < \xi + \beta/\alpha \). For \( |\alpha| < 0.3 \), the shape of the GEV distribution is similar to the Gumbel distribution, except that the right-hand tail is thicker for \( \alpha < 0 \) and thinner for \( \alpha > 0 \) (Stedinger et al., 1993).

The first three moments of the GEV distribution, respectively, are

\[
\mu_x = \lambda_1 = \xi + \left( \frac{\beta}{\alpha} \right) \left[ 1 - \Gamma(1 + \alpha) \right]
\]

\[
\sigma_x^2 = \left( \frac{\beta}{\alpha} \right)^2 \left[ \Gamma(1 + 2\alpha) - \Gamma^2(1 + \alpha) \right]
\]

\[
\gamma_x = \text{sign}(\alpha) \frac{-\Gamma(1 + 3\alpha) + 3\Gamma(1 + 2\alpha) \Gamma(1 + \alpha) - 2\Gamma^3(1 + \alpha)}{[\Gamma(1 + 2\alpha) - \Gamma^2(1 + \alpha)]^{1.5}}
\]

where \( \text{sign}(\alpha) \) is +1 or −1 depending on the sign of \( \alpha \). From Eqs. (2.94b) and (2.94c) one realizes that the variance of the GEV distribution exists when \( \alpha > -0.5 \), and the skewness coefficient exists when \( \alpha > -0.33 \). The GEV distribution recently has been used frequently in modeling the random mechanism of hydrologic extremes, such as precipitation and floods.

The relationships between the L-moments and GEV model parameters are

\[
\lambda_2 = \frac{\beta}{\alpha} (1 - 2^{-\alpha}) \Gamma(1 + \alpha)
\]

\[
\tau_3 = \frac{2(1 - 3^{-\alpha})}{(1 - 2^{-\alpha})} - 3
\]

\[
\tau_4 = \frac{1 - 5(4^{-\alpha}) + 10(3^{-\alpha}) - 6(2^{-\alpha})}{1 - 2^{-\alpha}}
\]

### 2.6.5 Beta distributions

The beta distribution is used for describing random variables having both lower and upper bounds. Random variables in hydrosystems that are bounded on both limits include reservoir storage and groundwater table for unconfined aquifers. The nonstandard beta PDF is

\[
f_{NB}(x | a, b, \alpha, \beta) = \frac{1}{B(\alpha, \beta)(b - a)^{\alpha-1}}(b - x)^{\beta-1}(x - a)^{\alpha-1}(b - x)^{\beta-1}
\]

for \( a \leq x \leq b \)

(2.96)

in which \( a \) and \( b \) are the lower and upper bounds of the beta random variable, respectively; \( \alpha > 0, \beta > 0 \); and \( B(\alpha, \beta) \) is a beta function defined as

\[
B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}
\]

(2.97)
Using the new variable $Y = (X - a)/(b - a)$, the nonstandard beta PDF can be reduced to the standard beta PDF as

$$f_B(y | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1} (1 - y)^{\beta-1} \quad \text{for } 0 < y < 1$$ (2.98)

The beta distribution is also a very versatile distribution that can have many shapes, as shown in Fig. 2.23. The mean and variance of the standard beta random variable $Y$, respectively, are

$$\mu_y = \frac{\alpha}{\alpha + \beta} \quad \sigma^2_y = \frac{\alpha \beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$$ (2.99)

When $\alpha = \beta = 1$, the beta distribution reduces to a uniform distribution as

$$f_U(x) = \frac{1}{b - a} \quad \text{for } a \leq x \leq b$$ (2.100)

### 2.6.6 Distributions related to normal random variables

The normal distribution has been playing an important role in the development of statistical theories. This subsection briefly describes two distributions related to the functions of normal random variables.

![Shapes of standard beta probability density functions. (After Johnson and Kotz, 1972.)](image)
χ² (chi-square) distribution. The sum of the squares of $K$ independent standard normal random variables results in a χ² (chi-square) random variable with $K$ degrees of freedom, denoted as $χ^2_K$. In other words,

$$
\sum_{k=1}^{K} Z_k^2 \sim χ^2_K
$$

(2.101)

in which the $Z_k$s are independent standard normal random variables. The PDF of a $χ^2$ random variable with $K$ degrees of freedom is

$$
f_{χ^2}(x \mid K) = \frac{1}{2^{K/2}Γ(K/2)}x^{(K/2-1)}e^{-x/2} \text{ for } x > 0
$$

(2.102)

Comparing Eq. (2.102) with Eq. (2.72), one realizes that the $χ^2$ distribution is a special case of the two-parameter gamma distribution with $α = K/2$ and $β = 2$. The mean, variance, and skewness coefficient of a $χ^2_K$ random variable, respectively, are

$$
µ_x = K \quad σ_x^2 = 2K \quad γ_x = 2/\sqrt{K/2}
$$

Thus, as the value of $K$ increases, the $χ^2$ distribution approaches a symmetric distribution. Figure 2.24 shows a few $χ^2$ distributions with various degrees of freedom. If $X_1, X_2, \ldots, X_K$ are independent normal random variables with the common mean $µ_x$ and variance $σ_x^2$, the $χ^2$ distribution is related to the sample of normal random variables as follows:

1. The sum of $K$ squared standardized normal variables $Z_k = (X_k - X)/σ_x$, $k = 1, 2, \ldots, K$, has a $χ^2$ distribution with $(K - 1)$ degrees of freedom.
2. The quantity $(K - 1)S^2/σ_x^2$ has a $χ^2$ distribution with $(K - 1)$ degrees of freedom in which $S^2$ is the unbiased sample variance computed according to Table 2.1.

![Figure 2.24 Shapes of chi-square probability density functions (PDFs) where d.f. refers to the degrees of freedom.](image)
**t-distribution.** A random variable having a t-distribution results from the ratio of the standard normal random variable to the square root of the \( \chi^2 \) random variable divided by its degrees of freedom, that is,

\[
T_K = \frac{Z}{\sqrt{\chi^2_K / K}}
\]

in which \( T_K \) is a t-distributed random variable with \( K \) degrees of freedom. The PDF of \( T_K \) can be expressed as

\[
f_T(x | K) = \frac{\Gamma[(K + 1)/2]}{\sqrt{\pi K} \Gamma(K/2)} \left(1 + \frac{x^2}{K}\right)^{-(K+1)/2} \quad \text{for } -\infty < x < \infty
\]

A t-distribution is symmetric with respect to the mean \( \mu_x = 0 \) when \( K \geq 1 \). Its shape is similar to the standard normal distribution, except that the tails of the PDF are thicker than \( \phi(z) \). However, as \( K \to \infty \), the PDF of a t-distributed random variable approaches the standard normal distribution. Figure 2.25 shows some PDFs for t-random variables of different degrees of freedom. It should be noted that when \( K = 1 \), the t-distribution reduces to the Cauchy distribution, for which all product-moments do not exist. The mean and variance of a t-distributed random variable with \( K \) degrees of freedom are

\[
\mu_x = 0 \quad \sigma_x^2 = K/(K - 2) \quad \text{for } K \geq 3
\]

When the population variance of normal random variables is known, the sample mean \( \bar{X} \) of \( K \) normal random samples from \( N(\mu_x, \sigma_x^2) \) has a normal distribution with mean \( \mu_x \) and variance \( \sigma_x^2 / K \). However, when the population variance is unknown but is estimated by \( S^2 \) according to Table 2.1, then the quantity \( \sqrt{K(\bar{X} - \mu_x)}/S \), which is the standardized sample mean using the sample variance, has a t-distribution with \( (K - 1) \) degrees of freedom.

![Figure 2.25](image-url)
2.7 Multivariate Probability Distributions

Multivariate probability distributions are extensions of univariate probability distributions that jointly account for more than one random variable. Bivariate and trivariate distributions are special cases where two and three random variables, respectively, are involved. The fundamental basis of multivariate probability distributions is described in Sec. 2.3.2. In general, the availability of multivariate distribution models is significantly less than that for univariate cases. Owing to their frequent use in multivariate modeling and reliability analysis, two multivariate distributions, namely, multivariate normal and multivariate lognormal, are presented in this section. Treatments of some multivariate nonnormal random variables are described in Secs. 4.5 and 7.5. For other types of multivariate distributions, readers are referred to Johnson and Kotz (1976) and Johnson (1987).

Several ways can be used to construct a multivariate distribution (Johnson and Kotz, 1976; Hutchinson and Lai, 1990). Based on the joint distribution discussed in Sec. 2.2.2, the straightforward way of deriving a joint PDF involving K multivariate random variables is to extend Eq. (2.19) as

\[ f(x) = f_1(x_1) \times f_2(x_2 | x_1) \times \cdots \times f_K(x_1, x_2, \ldots, x_{K-1}) \]  \hspace{1cm} (2.105)

in which \( x = (x_1, x_2, \ldots, x_K)^T \) is a vector containing variates of K random variables with the superscript \( t \) indicating the transpose of a matrix or vector. Applying Eq. (2.105) requires knowledge of the conditional PDFs of the random variables, which may not be easily obtainable.

One simple way of constructing a joint PDF of two random variables is by mixing. Morgenstern (1956) suggested that the joint CDF of two random variables could be formulated, according to their respective marginal CDFs, as

\[ F_{1,2}(x_1, x_2) = F_1(x_1)F_2(x_2)[1 + \theta[1 - F_1(x_1)][1 - F_2(x_2)]] \]  \hspace{1cm} (2.106)

in which \( F_k(x_k) \) is the marginal CDF of the random variable \( X_k \), and \( \theta \) is a weighting constant. When the two random variables are independent, the weighting constant \( \theta = 0 \). Furthermore, the sign of \( \theta \) indicates the positive-ness or negativeness of the correlation between the two random variables. This equation was later extended by Farlie (1960) to

\[ F_{1,2}(x_1, x_2) = F_1(x_1)F_2(x_2)[1 + \theta f_1(x_1)f_2(x_2)] \]  \hspace{1cm} (2.107)

in which \( f_k(x_k) \) is the marginal PDF of the random variable \( X_k \). Once the joint CDF is obtained, the joint PDF can be derived according to Eq. (2.15a).

Constructing a bivariate PDF by the mixing technique is simple because it only requires knowledge about the marginal distributions of the involved random variables. However, it should be pointed out that the joint distribution obtained from Eq. (2.106) or Eq. (2.107) does not necessarily cover the entire range of the correlation coefficient \([-1, 1]\) for the two random variables.
under consideration. This is illustrated in Example 2.20. Liu and Der Kiureghian (1986) derived the range of the valid correlation coefficient value for the bivariate distribution by mixing, according to Eq. (2.106), from various combinations of marginal PDFs, and the results are shown in Table 2.4.

Nataf (1962), Mardia (1970a, 1970b), and Vale and Maurelli (1983) proposed other ways to construct a bivariate distribution for any pair of random variables. This was done by finding the transforms $Z_k = t(X_k)$, for $k = 1, 2$, such that $Z_1$ and $Z_2$ are standard normal random variables. Then a bivariate normal distribution is ascribed to $Z_1$ and $Z_2$. One such transformation is 

$$z_k = \Phi^{-1}[F_k(x_k)],$$

for $k = 1, 2$. A detailed description of such a normal transformation is given in Sec. 4.5.3.

Example 2.20 Consider two correlated random variables $X$ and $Y$, each of which has a marginal PDF of an exponential distribution type as

$$f_x(x) = e^{-x} \quad \text{for } x \geq 0 \quad f_y(y) = e^{-y} \quad \text{for } y \geq 0$$

To derive a joint distribution for $X$ and $Y$, one could apply the Morgenstern formula. The marginal CDFs of $X$ and $Y$ can be obtained easily as

$$F_x(x) = 1 - e^{-x} \quad \text{for } x \geq 0 \quad F_y(y) = 1 - e^{-y} \quad \text{for } y \geq 0$$

According to Eq. (2.106), the joint CDF of $X$ and $Y$ can be expressed as

$$F_{x,y}(x, y) = (1 - e^{-x})(1 - e^{-y})(1 + \theta e^{-x-y}) \quad \text{for } x, y \geq 0$$

Then the joint PDF of $X$ and $Y$ can be obtained, according to Eq. (2.7a), as

$$f_{x,y}(x, y) = e^{-x-y}[1 + \theta(2e^{-x} - 1)(2e^{-y} - 1)] \quad \text{for } x, y \geq 0$$

<table>
<thead>
<tr>
<th>Marginal distribution</th>
<th>N</th>
<th>U</th>
<th>SE</th>
<th>SR</th>
<th>T1L</th>
<th>T1S</th>
<th>LN</th>
<th>GM</th>
<th>T2L</th>
<th>T3S</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>0.318</td>
<td>0.326</td>
<td>0.282</td>
<td>0.316</td>
<td>0.305</td>
<td>0.305</td>
<td>0.312</td>
<td>0.303</td>
<td>0.292</td>
<td>0.318</td>
</tr>
<tr>
<td>U</td>
<td>0.326</td>
<td>0.333</td>
<td>0.289</td>
<td>0.324</td>
<td>0.312</td>
<td>0.312</td>
<td>0.270</td>
<td>0.303</td>
<td>0.292</td>
<td>0.333</td>
</tr>
<tr>
<td>SE</td>
<td>0.282</td>
<td>0.289</td>
<td>0.25</td>
<td>0.28</td>
<td>0.27</td>
<td>0.27</td>
<td>0.270</td>
<td>0.292</td>
<td>0.292</td>
<td>0.289</td>
</tr>
<tr>
<td>SR</td>
<td>0.316</td>
<td>0.324</td>
<td>0.25</td>
<td>0.314</td>
<td>0.303</td>
<td>0.303</td>
<td>0.303</td>
<td>0.292</td>
<td>0.292</td>
<td>0.314</td>
</tr>
<tr>
<td>T1L</td>
<td>0.305</td>
<td>0.312</td>
<td>0.27</td>
<td>0.303</td>
<td>0.292</td>
<td>0.292</td>
<td>0.292</td>
<td>0.292</td>
<td>0.292</td>
<td>0.305</td>
</tr>
<tr>
<td>T1S</td>
<td>0.305</td>
<td>0.312</td>
<td>0.27</td>
<td>0.303</td>
<td>0.292</td>
<td>0.292</td>
<td>0.292</td>
<td>0.292</td>
<td>0.292</td>
<td>0.312</td>
</tr>
<tr>
<td>LN</td>
<td>&lt;0.318</td>
<td>&lt;0.326</td>
<td>&lt;0.282</td>
<td>&lt;0.316</td>
<td>&lt;0.305</td>
<td>&lt;0.305</td>
<td>&lt;0.305</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.318</td>
</tr>
<tr>
<td>GM</td>
<td>&lt;0.318</td>
<td>&lt;0.326</td>
<td>&lt;0.282</td>
<td>&lt;0.316</td>
<td>&lt;0.305</td>
<td>&lt;0.305</td>
<td>&lt;0.305</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.318</td>
</tr>
<tr>
<td>T2L</td>
<td>&lt;0.305</td>
<td>&lt;0.312</td>
<td>&lt;0.270</td>
<td>&lt;0.303</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.305</td>
</tr>
<tr>
<td>T3S</td>
<td>&lt;0.305</td>
<td>&lt;0.312</td>
<td>&lt;0.270</td>
<td>&lt;0.303</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
<td>&lt;0.292</td>
</tr>
</tbody>
</table>

**NOTE:** N = normal; U = uniform; SE = shifted exponential; SR = shifted Rayleigh; T1L = type I largest value; T1S = type I smallest value; LN = lognormal; GM = gamma; T2L = type II largest value; T3S = type III smallest value.

**SOURCE:** After Lin and Der Kiureghian (1986).
To compute the correlation coefficient between $X$ and $Y$, one first computes the covariance of $X$ and $Y$ as $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$, in which $E(XY)$ is computed by

$$E(XY) = \int_0^\infty \int_0^\infty xyf_{x,y}(x, y) \, dx \, dy = 1 + \frac{\theta}{4}$$

Referring to Eq. (2.79), since the exponential random variables $X$ and $Y$ currently considered are the special cases of $\beta = 1$, therefore, $\mu_x = \mu_y = 1$ and $\sigma_x = \sigma_y = 1$. Consequently, the covariance of $X$ and $Y$ is $\theta/4$, and the corresponding correlation coefficient is $\theta/4$. Note that the weighing constant $\theta$ lies between $[-1, 1]$. The preceding bivariate exponential distribution obtained from the Morgenstern formula could only be valid for $X$ and $Y$ having a correlation coefficient in the range $[-1/4, 1/4]$.

### 2.7.1 Multivariate normal distributions

A bivariate normal distribution has a PDF defined as

$$f_{x_1, x_2}(x_1, x_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho_{12}^2}} \exp \left[ -\frac{Q}{2(1 - \rho_{12}^2)} \right]$$

for $-\infty < x_1, x_2 < \infty$, in which

$$Q = \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho_{12} \left( \frac{x_1 - \mu_1}{\sigma_1} \right) \left( \frac{x_2 - \mu_2}{\sigma_2} \right)$$

where $\mu$ and $\sigma$ are, respectively, the mean and standard deviation, the subscripts 1 and 2 indicate the random variables $X_1$ and $X_2$, respectively, and $\rho_{12}$ is the correlation coefficient of the two random variables. Plots of the bivariate normal PDF in a three-dimensional form are shown in Fig. 2.26. The contour curves of the bivariate normal PDF of different correlation coefficients are shown in Fig. 2.27.

The marginal PDF of $X_k$ can be derived, according to Eq. (2.8), as

$$f_k(x_k) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x_k - \mu_k}{\sigma_k} \right)^2 \right]$$

for $-\infty < x_k < \infty$ for $k = 1$ and 2. As can be seen, the two random variables having a bivariate normal PDF are, individually, normal random variables. It should be pointed out that given two normal marginal PDFs, one can construct a bivariate PDF that is not in the form of a bivariate normal as defined by Eq. (2.108).

According to Eq. (2.17), the conditional normal PDF of $X_1 | x_2$ can be obtained as

$$f_{x_1 | x_2}(x_1 | x_2) = \frac{1}{\sigma_1 \sqrt{2\pi (1 - \rho_{12}^2)}} \exp \left\{ -\frac{1}{2} \left[ \frac{(x_1 - \mu_1) - \rho_{12} \sigma_1 \sigma_2 (x_2 - \mu_2)}{\sigma_1 \sqrt{1 - \rho_{12}^2}} \right]^2 \right\}$$

(2.109)
Figure 2.26 Three-dimensional plots of bivariate standard normal probability density functions. (After Johnson and Kotz, 1976.)
for $-\infty < x_1 < \infty$. Based on Eq. (2.109), the conditional expectation and variance of the normal random variable $X_1 \mid x_2$ can be obtained as

$$E(X_1 \mid x_2) = \mu_1 + \rho_{12}(\sigma_1/\sigma_2)(x_2 - \mu_2) \quad (2.110)$$

$$\text{Var}(X_1 \mid x_2) = \sigma^2_1 (1 - \rho^2_{12}) \quad (2.111)$$

Expressions of the conditional PDF, expectation, and variance for $X_2 \mid x_1$ can be obtained immediately by exchanging the subscripts in Eqs. (2.109) through (2.111).
For the general case involving $K$ correlated normal random variables, the multivariate normal PDF is

$$f(x) = \frac{|C_x|^{1/2}}{(2\pi)^{K/2}} \exp\left[-\frac{1}{2} (x - \mu_x)^t C_x^{-1} (x - \mu_x)\right] \quad \text{for } -\infty < x < \infty \quad (2.112)$$

in which $\mu_x = (\mu_1, \mu_2, \ldots, \mu_K)^t$, a $K \times 1$ column vector of the mean values of the variables, with the superscript $t$ indicating the transpose of a matrix or vector, and $C_x$ is a $K \times K$ covariance matrix:

$$Cov(X) = C_x = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_{KK} \end{bmatrix}$$

This covariance matrix is symmetric, that is, $\sigma_{jk} = \sigma_{kj}$, for $j \neq k$, where $\sigma_{jk} = Cov(X_j, X_k)$. In matrix notation, the covariance matrix for a vector of random variables can be expressed as

$$C_x = E[(X - \mu_x)(X - \mu_x)^t] \quad (2.113)$$

In terms of standard normal random variables, $Z_k = (X_k - \mu_k)/\sigma_k$, the standardized multivariate normal PDF, can be expressed as

$$\phi(z) = \frac{|R_z|^{1/2}}{(2\pi)^{K/2}} \exp\left(-\frac{1}{2} z^t R_z^{-1} z\right) \quad \text{for } -\infty < z < \infty \quad (2.114)$$

in which $R_z = C_z = E(ZZ^t)$ is a $K \times K$ correlation matrix:

$$Corr(X) = Cov(Z) = R_z = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ \rho_{21} & 1 & \cdots & \rho_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{K1} & \rho_{K2} & \cdots & 1 \end{bmatrix}$$

with $\rho_{jk} = Cov(Z_j, Z_k)$ being the correlation coefficient between each pair of normal random variables $X_j$ and $X_k$. For bivariate standard normal variables,
the following relationships of cross-product moments are useful (Hutchinson and Lai, 1990):

\[ E[Z_{1}^{2m}Z_{2}^{2n}] = \frac{(2m)!(2n)!}{(2m+n)!} \sum_{j=0}^{\min(m,n)} \frac{(2\rho_{12})^{2j}}{(m-j)!(n-j)!(2j)!} \]

\[ E[Z_{1}^{2m+1}Z_{2}^{2n+1}] = \frac{(2m+1)!(2n+1)!}{2^{m+n}} \rho_{12} \sum_{j=0}^{\min(m,n)} \frac{(2\rho_{12})^{2j}}{(m-j)!(n-j)!(2j+1)!} \]

\[ E[Z_{1}^{2m+1}Z_{2}^{2n}] = E[Z_{1}^{2m}Z_{2}^{2n+1}] = 0 \] (2.115)

for \( m \) and \( n \) being positive integer numbers.

2.7.2 Computation of multivariate normal probability

Evaluation of the probability of multivariate normal random variables involves multidimensional integration as

\[ \Phi(z | R_{x}) = P(Z_{1} \leq z_{1}, Z_{2} \leq z_{2}, \ldots, Z_{K} \leq z_{K} | R_{x}) = \int_{-\infty}^{z_{1}} \int_{-\infty}^{z_{2}} \cdots \int_{-\infty}^{z_{K}} \phi(z | R_{x}) \, dz \Delta y \Delta x \] (2.116)

Accurate evaluation for \( \Phi(z | R_{x}) \) is generally difficult and often is resolved by approximations.

Bivariate normal probability. For a bivariate case, Fortran computer programs for computing the lower left volume under the density surface, that is, \( \Phi(a, b | \rho) = P(Z_{1} \leq a, Z_{2} \leq b | \rho) \), have been developed by Donnelly (1973) and Baughman (1988). The double integral for evaluating the bivariate normal probability can be reduced to a single integral as shown in Eq. (2.111). Several approximations have been derived (Johnson and Kotz, 1976). Derivation of bounds for the bivariate normal probability is presented in Sec. 2.7.3. For a bivariate normal probability, exact solutions have been obtained in the form of figures such as Fig. 2.28 for computing the upper-right volume under the bivariate standard normal density surface, that is, \( L(a, b | \rho) = P(Z_{1} \geq a, Z_{2} \geq b | \rho) \), in which \( L(a, b | \rho) \) can be expressed in terms of \( L(a, 0 | \rho) \) as

\[ L(a, b | \rho) = L\left(a, 0 \bigg| \frac{(\rho a - b)(\text{sign} a)}{\sqrt{a^{2} - 2\rho ab + b^{2}}} \right) + L\left(b, 0 \bigg| \frac{(\rho b - a)(\text{sign} b)}{\sqrt{a^{2} - 2\rho ab + b^{2}}} \right) \]

\[ = \begin{cases} 0, & \text{if } ab > 0 \text{ or } ab = 0 \text{ and } a + b \geq 0 \\ \frac{1}{2}, & \text{otherwise} \end{cases} \] (2.117)
Chapter Two

The relationship between $\Phi(a, b \mid \rho)$ and $L(a, b \mid \rho)$ is

$$
\Phi(a, b \mid \rho) = -1 + \Phi(a) + \Phi(b) + L(a, b \mid \rho)
$$  \hspace{1cm} (2.118)

From the definitions of $\Phi(a, b \mid \rho)$ and $L(a, b \mid \rho)$ and the symmetry of the bivariate normal PDF, the following relations are in order:

$$
\Phi(a, \infty \mid \rho) = \Phi(a) \quad \Phi(\infty, b \mid \rho) = \Phi(b)
$$  \hspace{1cm} (2.119a)

$$
L(a, -\infty \mid \rho) = 1 - \Phi(a) \quad L(-\infty, b \mid \rho) = 1 - \Phi(b)
$$  \hspace{1cm} (2.119b)

Figure 2.28 Bivariate normal cumulative distribution function. (AfterAbramowitz and Stegun, 1972.)
\( L(a, 0 \mid r) \) for \( 0 \leq a \leq 1 \) and \( 0 \leq r \leq 1 \).

Values for \( a < 0 \) can be obtained using \( L(a, 0 \mid -r) = 0.5 - L(-a, 0 \mid r) \).

Figure 2.28 (Continued)

\[
L(a, b \mid \rho) = L(b, a \mid \rho) \quad (2.119c)
\]

\[
L(-a, b \mid \rho) + L(a, b \mid -\rho) = 1 - \Phi(b) \quad (2.119d)
\]

\[
L(-h, -k \mid \rho) - L(k, h \mid \rho) = 1 - \Phi(h) - \Phi(k) \quad (2.119e)
\]

Example 2.21 Consider two correlated normal random variables \( X_1 \) and \( X_2 \) with

their statistical properties being

\[
E(X_1) = 10 \quad \text{Var}(X_1) = 9 \quad E(X_2) = 5 \quad \text{Var}(X_2) = 4 \quad \text{Cov}(X_1, X_2) = 3.6
\]

Compute \( P(X_1 \leq 13, X_2 \leq 3) \).
Figure 2.28 (Continued)
Solution Based on the given information, the correlation coefficient between $X_1$ and $X_2$ is

$$\rho_{1,2} = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}} = \frac{3.6}{\sqrt{9} \sqrt{4}} = 0.6$$

Then

$$P(X_1 \leq 13, X_2 \leq 3) = P\left(Z_1 \leq \frac{13 - 10}{3}, Z_2 \leq \frac{3 - 5}{2} \mid \rho_{1,2} = 0.6\right)$$

$$= P(Z_1 \leq 1, Z_2 \leq 1 \mid \rho_{1,2} = 0.6)$$

$$= \Phi(a = 1, b = 1 \mid \rho_{1,2} = 0.6)$$

By Eq. (2.118),

$$\Phi(1, -1 \mid 0.6) = -1 + \Phi(1) + \Phi(-1) + L(1, -1 \mid 0.6)$$  \hspace{1cm} (a)$$

Since $ab = -1 < 0$, according to Eq. (2.117),

$$L(1, -1 \mid 0.6) = L\left(1, 0 \left| \frac{1.6}{\sqrt{3.2}} \right\rangle + L\left(-1, 0 \left| \frac{1.6}{\sqrt{3.2}} \right\rangle - \frac{1}{2}\right)$$

$$= L(1, 0 \mid 0.894) + L(-1, 0 \mid 0.894) - 0.5$$

From Fig. 2.28b, $L(1, 0 \mid 0.894) = 0.159$. Since $L(-1, 0 \mid 0.894) = 0.5 - L(1, 0 \mid -0.894)$, and according to Fig. 2.28a, by extrapolation, $L(1, 0 \mid -0.894) = 0.004$,

$$L(-1, 0 \mid 0.894) = 0.5 - 0.004 = 0.496$$

Consequently, $L(1, -1 \mid 0.6) = 0.159 + 0.496 - 0.5 = 0.155$.

According to (a), the bivariate normal probability is $\Phi(1, -1 \mid 0.6) = L(1, -1 \mid 0.6) = 0.155$. Alternatively, the bivariate normal probability can be computed by Eq. (2.121), and the result of the numerical integration for this example is 0.1569.

**Multivariate normal probability.** Johnson and Kotz (1976) show that if the correlation coefficient $\rho_{ij}$ can be expressed as $\rho_{ij} = \lambda_i \lambda_j$ for all $i$ and $j$ and $|\lambda_j| \leq 1$, then each correlated standard normal random variable $Z_k$ can be represented by

$$Z_k = \lambda_k Z'_0 + \sqrt{1 - \lambda_k^2} Z'_k$$ \hspace{1cm} for $k = 1, 2, \ldots, K$$

where $Z'_0, Z'_1, Z'_2, \ldots, Z'_K$ are *independent* standard normal variables. The inequality $Z_k \leq z_k$ can be expressed as

$$Z'_k \leq \frac{z_k - \lambda_k z'_0}{\sqrt{1 - \lambda_k^2}}$$
Then the multivariate normal probability can be calculated as

\[
\Phi(z|R_x) = \int_{-\infty}^{\infty} \phi(u) \prod_{k=1}^{K} \Phi \left( \frac{z_k - \lambda_k u}{\sqrt{1 - \lambda^2_k}} \right) du
\] (2.120)

As can be seen from Eq. (2.120), the computation of the multivariate normal probability is reduced from multiple integrals to a single integral for which the result can be obtained accurately by various numerical integration techniques. (see Appendix 4A) Under the special case of equicorrelation, that is, \( \rho_{ij} = \rho \), for all \( i \) and \( j \), the multivariate normal probability can be computed as

\[
\Phi(z|R_x) = \int_{-\infty}^{\infty} \phi(u) \prod_{k=1}^{K} \Phi \left( \frac{z_k - \sqrt{\rho} u}{\sqrt{1 - \rho}} \right) du
\] (2.121)

This equation, in particular, can be applied to evaluate the bivariate normal probability.

Under the general unequal correlation case, evaluation of multivariate normal probability by Eq. (2.120) requires solving for \( K \lambda' \)‘s based on \( K(K-1)/2 \) different values of \( \rho \) in the correlation matrix \( R_x \). This may not necessarily be a trivial task. Ditlevsen (1984) proposed an accurate algorithm by expanding \( \Phi(z|R_x) \) in a Taylor series about an equal correlation \( \rho_{ij} = \rho > 0 \), for \( i \neq j \). The equicorrelation \( \rho \) is determined in such a way that the first-order expansion term \( d \Phi(z | \rho) \) vanishes. The term \( d \Phi(z | \rho) \) can be expressed as

\[
d \Phi(z | \rho) = \frac{1}{1-\rho} \int_{-\infty}^{\infty} \phi(u) \left[ \prod_{k=1}^{K} \Phi \left( \frac{z_k - \sqrt{\rho} u}{\sqrt{1 - \rho}} \right) \right] \left[ \sum_{k=1}^{K} \sum_{j=k+1}^{K} a_k(u) a_j(u) \Delta \rho_{kj} \right] du
\] (2.122)

where

\[
a_k(u) = \phi \left( \frac{z_k - \sqrt{\rho} u}{\sqrt{1 - \rho}} \right) / \Phi \left( \frac{z_k - \sqrt{\rho} u}{\sqrt{1 - \rho}} \right)
\] (2.123)

and \( \Delta \rho_{ij} = \rho_{ij} - \rho \). In computing the value of \( d \Phi(z | \rho) \), numerical integration generally is required. However, one should be careful about the possible numerical overflow associated with the computation of \( a_k(u) \) as the value of \( u \) gets large. It can be shown that by the L’Hospital rule, \( \lim_{u \to \infty} a_k(u) = 0 \) and \( \lim_{u \to -\infty} a_k(u) = -u \).

Determination of the equicorrelation \( \rho \) for the expansion point can be made through the iterative procedure, as outlined in Fig. 2.29. A sensible starting value for \( \rho \) is the average of \( \rho_{ij} \):

\[
\rho = \frac{2}{K(K-1)} \sum_{i<j} \rho_{ij}
\] (2.124)
Given $\rho_{k,j}$, $z = (z_1, z_2, ..., z_K)^	op$.

Estimate $\rho_{old} = (\sum_{k<j} \rho_{kj})/[K(K-1)]$.

Compute $d\Phi(z)$.

Is $d\Phi(z) = 0$?

Yes

No

Compute $\Phi(z | \rho_{old})$.

Adjust $\rho_{new} = \rho_{old} (1 + d \Phi/\Phi)$.

Let $\rho_{old} = \rho_{new}$.

Compute $d^2\Phi$.

$\Phi(z | R_x) = \Phi(z | \rho_{old}) + d^2\Phi$.

Stop.

**Figure 2.29** Flowchart for determining the equicorrelation in a Taylor series expansion.

Once such $\rho$ is found, Ditlevsen (1984) suggests that the value of $\Phi(z | R_x)$ can be estimated accurately by a second-order approximation as

$$\Phi(z | R_x) \approx \Phi(z | \rho) + \frac{1}{2} d^2 \Phi(z | \rho)$$

(2.125)

in which $\Phi(z | \rho)$ is computed by Eq. (2.121), and the second-order error term is computed as

$$d^2 \Phi(z | \rho) = \frac{1}{4(1-\rho)^2} \int_{-\infty}^{\infty} \phi(u) \left[ \prod_{k=1}^{K} \Phi \left( \frac{z_k - \sqrt{\rho} u}{\sqrt{1-\rho}} \right) \right]
\sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{r=1}^{K} \sum_{s=1}^{K} a_i(u) a_j(u) a_r(u) a_s(u) (-b_r)^{h_{rs}} (-b_s)^{h_{rs}} \Delta \rho_{ij} \Delta \rho_{rs} du$$

(2.126)
where $\delta_{ir}$ is the Kronecker’s delta, having a value of 1 if $i = r$ and otherwise 0, and

$$b_k(u) = \frac{z_k \sqrt{\rho} u}{a_k(u) \sqrt{1 - \rho}} \quad (2.127)$$

As can be seen, the evaluation of $\Phi(\mathbf{z} | \mathbf{R}_x)$ is reduced from a multiple integral to a single integral, which can be executed efficiently and accurately by many numerical integration algorithms.

For the univariate normal distribution, an asymptotic expansion of $\Phi(z)$ is (Abramowitz and Stegun, 1972) for $z \to \infty$

$$
\Phi(-z) \approx \frac{\phi(z)}{z} \left[ 1 - \frac{1}{z^2} + \frac{1}{z^4} - \frac{1}{z^6} + \cdots + \frac{(-1)^n}{z^{2n}} \right] \quad (2.128)
$$

This expansion for $\Phi(z)$ is smaller than every summand with an odd number of terms and is larger than every summand with an even number of terms. The truncation error decreases as the number of terms increases. Note that Eq. (2.128) is particularly appropriate for evaluating the normal tail probability. The expansion has been generalized by Ruben (1964) for the multivariate normal distribution as

$$
\Phi(-z|\mathbf{R}_x) \approx \exp \left( -\frac{1}{2} \mathbf{z}^\top \mathbf{R}_x^{-1} \mathbf{z} \right) \sqrt{\frac{2\pi}{\prod_{k=1}^{K} a_k}} \quad \text{for } |\mathbf{z}| \to \infty \quad (2.129)
$$

in which the coefficients $a_k$ are elements in a vector $\mathbf{a}$ obtained from

$$\mathbf{a} = \mathbf{R}_x^{-1} \mathbf{z} \quad (2.130)$$

It should be noted that Eq. (2.130) is valid only when all coefficients $a_k$ are positive. The right-hand-side of Eq. (2.129) provides an upper bound for the multivariate normal probability.

### 2.7.3 Determination of bounds on multivariate normal probability

Instead of computing the exact value of $\Phi(\mathbf{z} | \mathbf{R}_x)$, several methods have been proposed to determine the bounds on the exact value of $\Phi(\mathbf{z} | \mathbf{R}_x)$. This section describes three such bounds.

#### Bounds of Rackwitz

The scheme of Rackwitz (1978) is based on the decomposition of a positive correlation coefficient $\rho_{ij} = \lambda_i \lambda_j$, for $i, j = 1, 2, \ldots, K$. The multivariate normal probability $\Phi(\mathbf{z} | \mathbf{R}_x)$ is obtained according to Eq. (2.120). Instead of solving for the exact values for all $K(K - 1)/2 \lambda$s, Rackwitz selects the smallest three values of $\mathbf{z} = (z_1, z_2, \ldots, z_K)^\top$ in $\Phi(\mathbf{z} | \mathbf{R}_x)$ and solves for the corresponding $\lambda$s that satisfy $\rho_{ij} = \lambda_i \lambda_j$, for $i, j = [1, 2, 3]$ with subscript $[i]$
representing the rank of \( z \)s in ascending order, that is, \( z[1] \leq z[2] \leq z[3] \leq \cdots \leq z[K−1] \leq z[K] \). For example, assume that all \( \rho_{ij} \)s are positive. Based on the three smallest \( z \)s, one can solve for \( \lambda[i] \) for \( i = 1, 2, 3 \) in terms of \( \rho[i][j] \) as

\[
\lambda[1] = \left( \frac{\rho[1][2] \rho[1][3]}{\rho[2][3]} \right)^{1/2} \\
\lambda[2] = \left( \frac{\rho[1][2] \rho[2][3]}{\rho[1][3]} \right)^{1/2} \\
\lambda[3] = \left( \frac{\rho[1][3] \rho[2][3]}{\rho[1][2]} \right)^{1/2}
\]

For the remaining \( \lambda \)s, their values can be computed as

\[
\lambda[i]_{U} = \max_{j<i} \left| \frac{\rho_{ij}}{\lambda_{j}U} \right| \quad i = [4], [5], \ldots, [K] \tag{2.132a}
\]

\[
\lambda[i]_{L} = \min_{j<i} \left| \frac{\rho_{ij}}{\lambda_{j}L} \right| \quad i = [4], [5], \ldots, [K] \tag{2.132b}
\]

The upper bound and lower bound of \( \Phi(z|R_x) \) can be obtained by Eq. (2.120) along with \( \lambda \)s computed by Eqs. (2.132a) and (2.132b), respectively.

**Bounds of Ditlevsen.** Ditlevsen (1979) proposed an approach for the bounds of the multivariate normal probability as follows:

\[
\Phi_U(z|R_x) = \Phi(z_1) - \sum_{k=2}^{K} \max \left\{ 0, \left[ \Phi(-z_k) - \sum_{j=1}^{k-1} \Phi(-z_k, -z_j | \rho_{kj}) \right] \right\} \tag{2.133a}
\]

\[
\Phi_L(z|R_x) = \Phi(z_1) - \sum_{k=2}^{K} \max \left\{ \Phi(-z_k) - \max_{j<k} \left[ \Phi(-z_k, -z_j | \rho_{kj}) \right] \right\} \tag{2.133b}
\]

in which \( \Phi_U(z|R_x) \) and \( \Phi_L(z|R_x) \) are the upper and lower bounds of the multivariate normal probability, respectively, and \( \Phi(-z_k, -z_j | \rho_{kj}) \) is the bivariate normal probability. Ditlevsen (1979) further simplified these bounds to involve the evaluation of only the univariate normal probability at the expense of having a more complicated algebraic expression where a narrow bound can be obtained under \( |\rho| < 0.6 \). For a larger correlation coefficient, Ditlevsen (1982) proposed a procedure using conditioning to obtain a narrow bound. The derivations of various probability bounds for system reliability are presented in Sec. 7.2.5

**Example 2.22** \( Z_1, Z_2, Z_3, Z_4, \) and \( Z_5 \) are correlated standard normal variables with the following correlation matrix:

\[
R_x = \begin{bmatrix}
  1.00 & 0.80 & 0.64 & 0.51 & 0.41 \\
  0.80 & 1.00 & 0.80 & 0.64 & 0.51 \\
  0.64 & 0.80 & 1.00 & 0.80 & 0.64 \\
  0.51 & 0.64 & 0.80 & 1.00 & 0.80 \\
  0.41 & 0.51 & 0.64 & 0.80 & 1.00
\end{bmatrix}
\]
Determine the multivariate probability \( P(Z_1 \leq -1, Z_2 \leq -2, Z_3 \leq 0, Z_4 \leq 2, Z_5 \leq 1) \) by Ditlevsen's approach using the Taylor series expansion. Also compute the bounds for the preceding multivariate normal probability using Rackwitz's and Ditlevsen's approaches.

**Solution** Using Ditlevsen's Taylor series expansion approach, the initial equicorrelation value can be used according to Eq. (2.124) as \( \rho = 0.655 \). The corresponding multivariate normal probability, based on Eq. (2.121), is \( \Phi(z | \rho = 0.655) = 0.01707 \). From Eq. (2.122), the first-order error, \( d \Phi(z | \rho = 0.655) \), is 0.003958. Results of iterations according to the procedure outlined in Fig. 2.29 are shown below:

| \( i \) | \( \rho \) | \( \Phi(z | \rho) \) | \( d \Phi(z | \rho) \) |
|-------|-------|---------|---------|
| 1     | 0.6550| 0.01707 | 0.3958 \times 10^{-2} |
| 2     | 0.8069| 0.02100 | -0.1660 \times 10^{-3} |
| 3     | 0.8005| 0.02086 | -0.3200 \times 10^{-4} |
| 4     | 0.7993| 0.02083 | -0.5426 \times 10^{-5} |

At \( \rho = 0.7993 \), the corresponding second-order error term in the Taylor series expansion, according to Eq. (2.126), is

\[
d^2 \Phi(z | \rho) = 0.01411
\]

Based on Eq. (2.125), the multivariate normal probability can be estimated as

\[
\Phi(z | \mathbf{R}) = \Phi(z | \rho = 0.7993) + 0.5d^2 \Phi(z | \rho = 0.7993)
\]

\[
= 0.02083 + 0.5(0.01411)
\]

\[
= 0.02789
\]

Using the Rackwitz approach for computing the bounds of the multivariate normal probability, the values of \( z \)'s are arranged in ascending order as \((z_1, z_2, z_3, z_4, z_5) = (-2, -1, 0, 1, 2)\) with the corresponding correlation matrix as

\[
\mathbf{R}_{kj} = \begin{bmatrix}
1.00 & 0.80 & 0.80 & 0.51 & 0.41 \\
0.80 & 1.00 & 0.64 & 0.41 & 0.51 \\
0.80 & 0.64 & 1.00 & 0.64 & 0.80 \\
0.51 & 0.41 & 0.80 & 1.00 & 0.80 \\
0.64 & 0.51 & 0.80 & 0.80 & 1.00
\end{bmatrix}
\]

The values of \( \lambda \)'s corresponding to the three smallest \( z \)'s, that is, -2, -1, and 0, are computed according to Eq. (2.131), and the results are

\[
\lambda_{[1]} = 1.00 \quad \lambda_{[2]} = 0.80 \quad \lambda_{[3]} = 0.80
\]

Using Eq. (2.132a), the values of the remaining \( \lambda \)'s for computing the upper bound are obtained as

\[
\lambda_{[4],U} = \max \left( \frac{\rho[41]}{\lambda_{[1]}}, \frac{\rho[42]}{\lambda_{[2]}}, \frac{\rho[43]}{\lambda_{[3]}}, \frac{\rho[44]}{\lambda_{[4]}} \right) = \max \left( 0.51 \times 0.41 \times 0.64 \times 1.00 \right) = 0.80
\]

\[
\lambda_{[5],U} = \max \left( \frac{\rho[51]}{\lambda_{[1]}}, \frac{\rho[52]}{\lambda_{[2]}}, \frac{\rho[53]}{\lambda_{[3]}}, \frac{\rho[54]}{\lambda_{[4]}}, \frac{\rho[55]}{\lambda_{[5]}} \right) = \max \left( 0.64 \times 0.51 \times 0.80 \times 0.80 \times 1.00 \right) = 1.00
\]
and, by the same token, for the lower bound are

\[ \lambda_{[4],L} = 0.51 \quad \lambda_{[5],L} = 0.6375 \]

Applying Eq. (2.120), along with \( \lambda_U = (1.0, 0.8, 0.8, 0.8, 1.0) \), one obtains the upper bound for the multivariate normal probability \( \Phi_U(z \mid R) = 0.01699 \). Similarly, using \( \lambda_L = (1.0, 0.8, 0.8, 0.51, 0.6375) \), the lower bound is obtained as \( \Phi_L(z \mid R) = 0.01697 \).

To use Eqs. (2.133a) and (2.133b) for computing the upper and lower bounds for the multivariate normal probability, the marginal probabilities and each pair of bivariate normal probabilities are computed first, according to Eq. (2.3). The results are

\[ \Phi(z_1) = 0.1587 \quad \Phi(z_2) = 0.02275 \quad \Phi(z_3) = 0.5000 \quad \Phi(z_4) = 0.9772 \quad \Phi(z_5) = 0.8413 \]

\[ \Phi(-z_1, -z_2) = 0.8395 \quad \Phi(-z_1, -z_3) = 0.4816 \]

\[ \Phi(-z_1, -z_4) = 0.0226 \quad \Phi(-z_1, -z_5) = 0.1523 \]

\[ \Phi(-z_2, -z_3) = 0.5000 \quad \Phi(-z_2, -z_4) = 0.0228 \quad \Phi(-z_2, -z_5) = 0.1585 \]

\[ \Phi(-z_3, -z_4) = 0.0227 \quad \Phi(-z_3, -z_5) = 0.1403 \quad \Phi(-z_4, -z_5) = 0.0209 \]

The lower and upper bounds of the multivariate probability can be obtained as 0.02070 and 0.02086, respectively.

### 2.7.4 Multivariate lognormal distributions

Similar to the univariate case, bivariate lognormal random variables have a PDF

\[
 f_{x_1, x_2}(x_1, x_2) = \frac{1}{2\pi x_1 x_2 \sigma_{\ln x_1} \sigma_{\ln x_2} \sqrt{1 - \rho_{12}^2}} \exp \left[ \frac{-Q'}{2(1 - \rho_{12}^2)} \right]
\] (2.134)

for \( x_1, x_2 > 0 \), in which

\[
 Q' = \frac{[\ln(x_1) - \mu_{\ln x_1}]^2}{\sigma_{\ln x_1}^2} + \frac{[\ln(x_2) - \mu_{\ln x_2}]^2}{\sigma_{\ln x_2}^2} - 2\rho'_{12} \frac{[\ln(x_1) - \mu_{\ln x_1}][\ln(x_2) - \mu_{\ln x_2}]}{\sigma_{\ln x_1} \sigma_{\ln x_2}}
\]

where \( \mu_{\ln x} \) and \( \sigma_{\ln x} \) are the mean and standard deviation of log-transformed random variables, subscripts 1 and 2 indicate the random variables \( X_1 \) and \( X_2 \), respectively, and \( \rho'_{12} = \text{Corr}(\ln X_1, \ln X_2) \) is the correlation coefficient of the two log-transformed random variables. After log-transformation is made, properties of multivariate lognormal random variables follow exactly as for the multivariate normal case. The relationship between the correlation coefficients in the original and log-transformed spaces can be derived using the moment-generating function (Tung and Yen, 2005, Sec. 4.2) as

\[
 \text{Corr}(X_1, X_2) = \rho_{12} = \frac{\exp(\rho'_{12} \sigma_{\ln x_1} \sigma_{\ln x_2}) - 1}{\sqrt{\exp(\sigma_{\ln x_1}^2) - 1} \sqrt{\exp(\sigma_{\ln x_2}^2) - 1}}
\] (2.135)
Example 2.23  Resolve Example 2.21 by assuming that both $X_1$ and $X_2$ are bivariate lognormal random variables.

Solution  Since $X_1$ and $X_2$ are lognormal variables,

$$P(X_1 \leq 13, X_2 \leq 3) = P[\ln(X_1) \leq \ln(13), \ln(X_2) \leq \ln(3)]$$

$$= P\left(Z_1 \leq \frac{\ln(13) - \mu_1'}{\sigma_1'}, Z_2 \leq \frac{\ln(3) - \mu_2'}{\sigma_2'} \mid \rho'\right)$$

in which $\mu_1'$, $\mu_2'$, $\sigma_1'$, and $\sigma_2'$ are the means and standard deviations of $\ln(X_1)$ and $\ln(X_2)$, respectively; $\rho'$ is the correlation coefficient between $\ln(X_1)$ and $\ln(X_2)$. The values of $\mu_1$, $\mu_2$, $\sigma_1$, and $\sigma_2$ can be computed, according to Eqs. (2.67a) and (2.67b), as

$$\sigma_1' = \sqrt{\ln(1 + 0.3^2)} = 0.294 \quad \sigma_2' = \sqrt{\ln(1 + 0.4^2)} = 0.385$$

$$\mu_1' = \ln(10) - \frac{1}{2}(0.294)^2 = 2.259 \quad \mu_2' = \ln(5) - \frac{1}{2}(0.385)^2 = 1.535$$

Based on Eq. (2.71), the correlation coefficient between $\ln(X_1)$ and $\ln(X_2)$ is

$$\rho' = \frac{\ln[1 + (0.6)(0.3)(0.4)]}{0.294 \cdot 0.385} = 0.623$$

Then

$$P(X_1 \leq 13, X_2 \leq 3 \mid \rho = 0.6) = P(Z_1 \leq 1.04, Z_2 \leq -1.13 \mid \rho' = 0.623)$$

$$= \Phi(a = 1.04, b = -1.13 \mid \rho' = 0.623)$$

From this point forward, the procedure for determining $\Phi(a = 1.04, b = -1.13 \mid \rho' = 0.623)$ is exactly identical to that of Example 2.21. The result from using Eq. (2.121) is 0.1285.

Problems

2.1 Referring to Example 2.4, solve the following problems:

(a) Assume that $P(E_1 \mid E_2) = 1.0$ and $P(E_2 \mid E_1) = 0.8$. What is the probability that the flow-carrying capacity of the sewer main is exceeded?

(b) If the flow capacity of the downstream sewer main is twice that of its two upstream branches, what is the probability that the flow capacity of the downstream sewer main is exceeded? Assume that if only branch 1 or branch 2 exceeds its corresponding capacity, the probability of flow in the sewer main exceeding its capacity is 0.15.

(c) Under the condition of (b), it is observed that surcharge occurred in the downstream sewer main. Determine the probabilities that (i) only branch 1 exceeds its capacity, (ii) only branch 2 is surcharged, and (iii) none of the sewer branches exceed their capacities.

2.2 Referring to Example 2.5, it is observed that surcharge occurred in the downstream sewer main. Determine the probabilities that (a) only branch 1 exceeds its flow-carrying capacity, (b) only branch 2 is surcharged, and (c) none of the sewer branches exceed their capacities.
2.3 A detention basin is designed to accommodate excessive surface runoff temporarily during storm events. The detention basin should not overflow, if possible, to prevent potential pollution of streams or other receiving water bodies.

For simplicity, the amount of daily rainfall is categorized as heavy, moderate, and light (including none). With the present storage capacity, the detention basin is capable of accommodating runoff generated by two consecutive days of heavy rainfall or three consecutive days of moderate rainfall. The daily rainfall amounts around the detention basin site are not entirely independent. In other words, the amount of rainfall on a given day would affect the rainfall amount on the next day.

Let random variable $X_t$ represent the amount of rainfall in any day $t$. The transition probability matrix indicating the conditional probability of the rainfall amount in a given day $t$, conditioned on the rainfall amount of the previous day, is shown in the following table.

<table>
<thead>
<tr>
<th>$X_{t+1}$</th>
<th>H</th>
<th>M</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>0.3</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>$X_t = M$</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>L</td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
</tr>
</tbody>
</table>

(a) For a given day, the amount of rainfall is light. What is the probability that the detention basin will overflow in the next three days? (After Mays and Tung, 1992.)

(b) Compute the probability that the detention basin will overflow in the next three days. Assume that at any given day of the month the probabilities for having the various rainfall amounts are $P(H) = 0.1$, $P(M) = 0.3$, $P(L) = 0.6$.

2.4 Before a section of concrete pipe of a special order can be accepted for installation in a culvert project, the thickness of the pipe needs to be inspected by state highway department personnel for specification compliance using ultrasonic reading. For this project, the required thickness of the concrete pipe wall must be at least 3 in. The inspection is done by arbitrarily selecting a point on the pipe surface and measuring the thickness at that point. The pipe is accepted if the thickness from the ultrasonic reading exceeds 3 in; otherwise, the entire section of the pipe is rejected. Suppose, from past experience, that 90 percent of all pipe sections manufactured by the factory were found to be in compliance with specifications. However, the ultrasonic thickness determination is only 80 percent reliable.

(a) What is the probability that a particular pipe section is well manufactured and will be accepted by the highway department?

(b) What is the probability that a pipe section is poorly constructed but will be accepted on the basis of ultrasonic test?

2.5 A quality-control inspector is testing the sample output from a manufacturing process for concrete pipes for a storm sewer project, wherein 95 percent of the items are satisfactory. Three pipes are chosen randomly for inspection. The successive
quality evaluations may be considered as independent. What is the probability that (a) none of the three pipes inspected are satisfactory and (b) exactly two are satisfactory?

2.6 Derive the PDF for a random variable having a triangular distribution with the lower bound \( a \), mode \( m \), and the upper bound \( b \), as shown in Fig. 2P.1.

2.7 Show that \( F_1(x_1) + F_2(x_2) - 1 \leq F_{1,2}(x_1, x_2) \leq \min\{F_1(x_1), F_2(x_2)\} \)

2.8 The Farlie-Gumbel-Morgenstern bivariate uniform distribution has the following joint CDF (Hutchinson and Lai, 1990):

\[
F_{x,y}(x, y) = xy[1 + \theta(1 - x)(1 - y)] \quad \text{for } 0 \leq x, y \leq 1
\]

with \(-1 \leq \theta \leq 1\). Do the following exercises: (a) derive the joint PDF, (b) obtain the marginal CDF and PDF of \( X \) and \( Y \), and (c) derive the conditional PDFs \( f_x(x|y) \) and \( f_y(y|x) \).

2.9 Refer to Problem 2.8. Compute (a) \( P(X \leq 0.5, Y \leq 0.5) \), (b) \( P(X \geq 0.5, Y \geq 0.5) \), and (c) \( P(X \geq 0.5 | Y = 0.5) \).

2.10 Apply Eq. (2.22) to show that the first four central moments in terms of moments about the origin are

\[
\begin{align*}
\mu_1 &= 0 \\
\mu_2 &= \mu'_2 - \mu_x^2 \\
\mu_3 &= \mu'_3 - 3\mu_x\mu'_2 + 2\mu_x^3 \\
\mu_4 &= \mu'_4 - 4\mu_x\mu'_3 + 6\mu_x^2\mu'_2 - 3\mu_x^4
\end{align*}
\]

![Figure 2P.1 Triangular distribution.](image)
2.11 Apply Eq. (2.23) to show that the first four moments about the origin could be expressed in terms of the first four central moments as

\[ \mu'_1 = \mu_x \]
\[ \mu'_2 = \mu_2 + \mu_x^2 \]
\[ \mu'_3 = \mu_3 + 3\mu_x\mu_2 + \mu_x^3 \]
\[ \mu'_4 = \mu_4 + 4\mu_x\mu_3 + 6\mu_x^2\mu_2 + \mu_x^4 \]

2.12 Based on definitions of \( \alpha \)- and \( \beta \)-moments, i.e., Eqs. (2.26a) and (2.26b), (a) derive the general expressions between the two moments, and (b) write out explicitly their relations for \( r = 0, 1, 2, \text{ and } 3 \).

2.13 Refer to Example 2.9. Continue to derive the expressions for the third and fourth L-moments of the exponential distribution.

2.14 A company plans to build a production factory by a river. You are hired by the company as a consultant to analyze the flood risk of the factory site. It is known that the magnitude of an annual flood has a lognormal distribution with a mean of 30,000 ft\(^3\)/s and standard deviation 25,000 ft\(^3\)/s. It is also known from a field investigation that the stage-discharge relationship for the channel reach is \( Q = 1500H^{1.4} \), where \( Q \) is flow rate (in ft\(^3\)/s) and \( H \) is water surface elevation (in feet) above a given datum. The elevation of a tentative location for the factory is 15 ft above the datum (after Mays and Tung, 1992). (a) What is the annual risk that the factory site will be flooded? (b) At this plant site, it is also known that the flood-damage function can be approximated as

\[
\text{Damage (in $1000)} = \begin{cases} 
0 & \text{if } H \leq 15 \text{ ft} \\
40(ln H + 8)(ln H - 2.7) & \text{if } H > 15 \text{ ft}
\end{cases}
\]

What is the annual expected flood damage? (Use the appropriate numerical approximation technique for calculations.)

2.15 Referring to Problem 2.6, assume that Manning’s roughness coefficient has a triangular distribution as shown in Fig. 2P.1. (a) Derive the expression for the mean and variance of Manning’s roughness. (b) Show that (i) for a symmetric triangular distribution, \( \sigma = (b - m)/\sqrt{6} \) and (ii) when the mode is at the lower or upper bound, \( \sigma = (b - a)/3\sqrt{2} \).

2.16 Suppose that a random variable \( X \) has a uniform distribution (Fig. 2P.2), with \( a \) and \( b \) being its lower and upper bounds, respectively. Show that (a) \( E(X) = \mu_x = (b + a)/2 \), (b) \( \text{Var}(X) = (b - a)^2/12 \), and (c) \( \Omega_x = (1 - a/\mu_x)/\sqrt{3} \).

2.17 Referring to the uniform distribution as shown in Fig. 2P.2, (a) derive the expression for the first two probability-weighted moments, and (b) derive the expressions for the L-coefficient of variation.

2.18 Refer to Example 2.8. Based on the conditional PDF obtained in part (c), derive the conditional expectation \( E(Y | x) \), and the conditional variance \( \text{Var}(Y | x) \).
Furthermore, plot the conditional expectation and conditional standard deviation of $Y$ on $x$ with respect to $x$.

2.19 Consider two random variables $X$ and $Y$ having the joint PDF of the following form:

$$f_{x,y}(x,y) = c \left( 5 - \frac{y}{2} + x^2 \right)$$

for $0 \leq x, y \leq 2$

(a) Determine the coefficient $c$. (b) Derive the joint CDF. (c) Find $f_x(x)$ and $f_y(y)$. (d) Determine the mean and variance of $X$ and $Y$. (e) Compute the correlation coefficient between $X$ and $Y$.

2.20 Consider the following hydrologic model in which the runoff $Q$ is related to the rainfall $R$ by

$$Q = a + bR$$

if $a > 0$ and $b > 0$ are model coefficients. Ignoring uncertainties of model coefficients, show that $\text{Corr}(Q, R) = 1.0$.

2.21 Suppose that the rainfall-runoff model in Problem 2.4.11 has a model error, and it can expressed as

$$Q = a + bR + \varepsilon$$

in which $\varepsilon$ is the model error term, which has a zero mean and standard deviation of $\sigma_\varepsilon$. Furthermore, the model error $\varepsilon$ is independent of the random rainfall $R$. Derive the expression for $\text{Corr}(Q, R)$.

2.22 Let $X = X_1 + X_3$ and $Y = X_2 + X_3$. Find $\text{Corr}(X, Y)$, assuming that $X_1$, $X_2$, and $X_3$ are statistically independent.

2.23 Consider two random variables $Y_1$ and $Y_2$ that each, individually, is a linear function of two other random variables $X_1$ and $X_2$ as follows:

$$Y_1 = a_{11}X_1 + a_{12}X_2 \quad Y_2 = a_{21}X_1 + a_{22}X_2$$

It is known that the mean and standard deviations of random variable $X_k$ are $\mu_k$ and $\sigma_k$, respectively, for $k = 1, 2$. (a) Derive the expression for the correlation coefficient.
coefficient between $Y_1$ and $Y_2$ under the condition that $X_1$ and $X_2$ are statistically independent. (b) Derive the expression for the correlation coefficient between $Y_1$ and $Y_2$ under the condition that $X_1$ and $X_2$ are correlated with a correlation coefficient $\rho$.

2.24 As a generalization to Problem 2.23, consider $M$ random variables $Y_1, Y_2, \ldots, Y_M$ that are linear functions of $K$ other random variables $X_1, X_2, \ldots, X_K$ in a vector form as follows:

$$Y_k = a_k^t X \quad \text{for } k = 1, 2, \ldots, M$$

in which $X = (X_1, X_2, \ldots, X_K)^t$, a column vector of $K$ random variables $X$s and $a_k^t = (a_{1k}, a_{2k}, \ldots, a_{Kk})$, a row vector of coefficients for the random variable $Y_k$. In matrix form, the preceding system of linear equations can be written as $y = A^t X$.

Given that the mean and standard deviations of the random variable $X_k$ are $\mu_k$ and $\sigma_k$, respectively, for $k = 1, 2, \ldots, K$, (a) derive the expression for the correlation matrix between $Y$s assuming that the random variable $X$s are statistically independent, and (b) derive the expression for the correlation coefficient between $Y$s under the condition that the random variable $X$s are correlated with a correlation matrix $R_x$.

2.25 A coffer dam is to be built for the construction of bridge piers in a river. In an economic analysis of the situation, it is decided to have the dam height designed to withstand floods up to 5000 ft$^3$/s. From flood frequency analysis it is estimated that the annual maximum flood discharge has a Gumbel distribution with the mean of 2500 ft$^3$/s and coefficient of variation of 0.25. (a) Determine the risk of flood water overtopping the coffer dam during a 3-year construction period. (b) If the risk is considered too high and is to be reduced by half, what should be the design flood magnitude?

2.26 Recompute the probability of Problem 2.25 by using the Poisson distribution.

2.27 There are five identical pumps at a pumping station. The PDFs of the time to failure of each pump are the same with an exponential distribution as Example 2.6, that is, $f(t) = 0.0008 \exp(-0.0008t)$ for $t \geq 0$.

The operation of each individual pump is assumed to be independent. The system requires at least two pumps to be in operation so as to deliver the required amount of water. Assuming that all five pumps are functioning, determine the reliability of the pump station being able to deliver the required amount of water over a 200-h period.

2.28 Referring to Example 2.14, determine the probability, by both binomial and Poisson distributions, that there would be more than five overtopping events over a period of 100 years. Compare the results with that using the normal approximation.

2.29 From a long experience of observing precipitation at a gauging station, it is found that the probability of a rainy day is 0.30. What is the probability that the next year would have at least 150 rainy days by looking up the normal probability table?
2.30 The well-known Thiem equation can be used to compute the drawdown in a confined and homogeneous aquifer as

\[ s_{ik} = \frac{\ln(r_{ok}/r_{ik})}{2\pi T} Q_k = \xi_{ik} Q_k \]

in which \( s_{ik} \) is drawdown at the \( i \)th observation location resulting from a pumpage of \( Q_k \) at the \( k \)th production well, \( r_{ok} \) is the radius of influence of the \( k \)th production well, \( r_{ik} \) is the distance between the \( i \)th observation point and the \( k \)th production well, and \( T \) is the transmissivity of the aquifer. The overall effect of the aquifer drawdown at the \( i \)th observation point, when more than one production well is in operation, can be obtained, by the principle of linear superposition, as the sum of the responses caused by all production wells in the field, that is,

\[ s_i = \sum_{k=1}^{K} \xi_{ik} Q_k = K \sum_{k=1}^{K} \xi_{ik} Q_k \]

where \( K \) is the total number of production wells in operation. Consider a system consisting of two production wells and one observation well. The locations of the three wells, the pumping rates of the two production wells, and their zones of influence are shown in Fig. 2P.3. It is assumed that the transmissivity of the aquifer has a lognormal distribution with the mean \( \mu_T = 4000 \text{ gpd/ft} \) and standard deviation \( \sigma_T = 2000 \text{ gpd/ft} \) (after Mays and Tung, 1992). (a) Prove that the total drawdown in the aquifer field also is lognormally distributed. (b) Compute the exact values of the mean and variance of the total drawdown at the observation point when \( Q_1 = 10,000 \text{ gpd} \) and \( Q_2 = 15,000 \text{ gpd} \). (c) Compute the probability that the resulting drawdown at the observation point does not exceed 2 ft. (d) If the maximum allowable probability of the total drawdown exceeding 2 ft is 0.10, find out the maximum allowable total pumpage from the two production wells.

2.31 A frequently used surface pollutant washoff model is based on a first-order decay function (Sartor and Boyd, 1972):

\[ M_t = M_0 e^{-cRt} \]

where \( M_0 \) is the initial pollutant mass at time \( t = 0 \), \( R \) is runoff intensity (mm/h), \( c \) is the washoff coefficient (mm\(^{-1}\)), \( M_t \) is the mass of the pollutant remaining.
Consider that measured hydrologic quantity $Q$ is subject to measurement error $\varepsilon$ and that both are related to the true but unknown discharge $Q$ as (Cong and Xu, 1987)

$$\varepsilon = Q' - Q$$

It is common to assume that (i) $E(\varepsilon | q) = 0$, (ii) $\text{Var}(\varepsilon | q) = [\sigma(q)]^2$, and (iii) random error $\varepsilon$ is normally distributed, that is, $\varepsilon | q \sim N(\mu_x | q = 0, \sigma_x | q)$.

(a) Show that $E(Q' | q) = q$, $E[(Q' | Q - q)^2 | q] = 0$, and $\text{Var}(Q' | Q - q)^2 | q] = \sigma^2(q)$.

(b) Under $\sigma(q) = \sigma$, show that $E(Q') = E(Q)$, $\text{Var}(\varepsilon) = \sigma^2 E(Q^2)$, and $\text{Var}(Q') = (1 + \sigma^2) \text{Var}(Q) + \sigma^2 E^2(Q)$. (c) Suppose that it is required that 75 percent of measurements have relative errors in the range of $\pm 5$ percent (precision level). Determine the corresponding value of $\sigma(q)$ assuming that the measurement error is normally distributed.

Show that the valid range of the correlation coefficient obtained in Example 2.20 is correct also for the general case of exponential random variables with parameters $\beta_1$ and $\beta_2$ of the form of Eq. (2.79).

Referring to Example 2.20, derive the range of the correlation coefficient for a bivariate exponential distribution using Farlie's formula (Eq. 2.107).

The Pareto distribution is used frequently in economic analysis to describe the randomness of benefit, cost, and income. Consider two correlated Pareto random
variables, each of which has the following marginal PDFs:

\[ f_k(x_k) = \frac{a^\theta}{x_k^{a+1}} \quad x_k > \theta_k > 0 \quad a > 0, \quad \text{for } k = 1, 2 \]

Derive the joint PDF and joint CDF by Morgenstern’s formula. Furthermore, derive the expression for \( E(X_1|X_2) \) and the correlation coefficient between \( X_1 \) and \( X_2 \).

2.38 Repeat Problem 2.37 using Farlie’s formula.

2.39 Analyzing the stream flow data from several flood events, it is found that the flood peak discharge \( Q \) and the corresponding volume \( V \) have the following relationship:

\[ \ln(V) = a + b \times \ln(Q) + \varepsilon \]

in which \( a \) and \( b \) are constants, and \( \varepsilon \) is the model error term. Suppose that the model error term \( \varepsilon \) has a normal distribution with mean 0 and standard deviation \( \sigma_\varepsilon \). Then show that the conditional PDF of \( V | Q \), \( h(v|q) \), is a lognormal distribution. Furthermore, suppose that the peak discharge is a lognormal random variable. Show that the joint PDF of \( V \) and \( Q \) is bivariate lognormal.

2.40 Analyzing the stream flow data from 105 flood events at different locations in Wyoming, Wahl and Rankl (1993) found that the flood peak discharge \( Q \) (in ft\(^3\)/s) and the corresponding volume \( V \) (in acre-feet, AF) have the following relationship:

\[ \ln(V) = \ln(0.0655) + 1.011 \times \ln(Q) + \varepsilon \]

in which \( \varepsilon \) is the model error term with the assumed \( \sigma_\varepsilon = 0.3 \). A flood frequency analysis of the North Platte River near Walden, Colorado, indicated that the annual maximum flood has a lognormal distribution with mean \( \mu_Q = 1380 \text{ ft}^3/\text{s} \) and \( \sigma_Q = 440 \text{ ft}^3/\text{s} \). (a) Derive the joint PDF of \( V \) and \( Q \) for the annual maximum flood. (b) Determine the correlation coefficient between \( V \) and \( Q \). (c) Compute \( P(Q \geq 2000 \text{ ft}^3/\text{s}, V \geq 180 \text{ AF}) \).

2.41 Let \( X_2 = a_0 + a_1Z_1 + a_2Z_1^2 \) and \( X_2 = b_0 + b_1Z_2 + b_2Z_2^2 \) in which \( Z_1 \) and \( Z_2 \) are bivariate standard normal random variables with a correlation coefficient \( \rho \), that is, \( \text{Corr}(Z_1, Z_2) = \rho \). Derive the expression for \( \text{Corr}(X_1, X_2) \) in terms of polynomial coefficients and \( \rho \).

2.42 Let \( X_1 \) and \( X_2 \) be bivariate lognormal random variables. Show that

\[
\frac{\exp \left(-\sigma_{\ln x_1}^2 \sigma_{\ln x_2}^2 - 1 \right)}{\sqrt{\exp \left(\sigma_{\ln x_1}^2\right) - 1} \sqrt{\exp \left(\sigma_{\ln x_2}^2\right) - 1}} \leq \text{Corr}(X_1, X_2) \leq \frac{\exp \left(\sigma_{\ln x_1}^2 \sigma_{\ln x_2}^2 - 1 \right)}{\sqrt{\exp \left(\sigma_{\ln x_1}^2\right) - 1} \sqrt{\exp \left(\sigma_{\ln x_2}^2\right) - 1}}
\]

What does this inequality indicate?
2.43 Derive Eq. (2.71) from Eq. (2.135):

\[
\text{Corr}(\ln X_1, \ln X_2) = \rho_{12}' = \frac{\ln(1 + \rho_{12}\Omega_1\Omega_2)}{\sqrt{\ln(1 + \Omega_1^2)}\sqrt{\ln(1 + \Omega_2^2)}}
\]

where \(\rho_{12} = \text{Corr}(X_1, X_2)\) and \(\Omega_k = \text{coefficient of variation of } X_k, k = 1, 2\).

2.44 Develop a computer program using Ditlevsen’s expansion for estimating the multivariate normal probability.

2.45 Develop computer programs for multivariate normal probability bounds by Rackwitz’s procedure and Ditlevsen’s procedure, respectively.

References


Chapter Two


Chapter 3

Hydrologic Frequency Analysis

One of the basic questions in many hydrosystems infrastructural designs that an engineer must answer is, “What should be the capacity or size of a system?” The planning goal is not to eliminate all hydro-hazards but to reduce the frequency of their occurrences and thus the resulting damage. If such planning is to be correct, the probabilities of flooding must be evaluated correctly. The problem is made more complex because in many cases the “input” is controlled by nature rather than by humans. For example, variations in the amount, timing, and spatial distribution of precipitation are the underlying reasons for the need for probabilistic approaches for many civil and environmental engineering projects. Our understanding and ability to predict precipitation and its resulting effects such as runoff are far from perfect. How, then, can an engineer approach the problem of design when he or she cannot be certain of the hydrologic load that will be placed on the infrastructure under consideration?

An approach that is used often is a statistical or probabilistic one. Such an approach does not require a complete understanding of the hydrologic phenomenon involved but examines the relationship between magnitude and frequency of occurrence in the hope of finding some statistical regularity between these variables. In effect, the past is extrapolated into the future. This assumes that whatever complex physical interactions control nature, the process does not change with time, and so the historical record can be used as a basis for estimating future events. In other words, the data are assumed to satisfy statistical stationarity by which the underlying distributional properties do not change with time, and the historical data series is representative of the storms and watershed conditions to be experienced in the future. An example that violates this statistical stationarity is the progressive urbanization within a watershed that could result in a tendency of increasing peak flow over time.

The hydrologic data most commonly analyzed in this way are rainfall and stream flow records. Frequency analysis was first used for the study of stream flow records by Herschel and Freeman during the period from 1880 to 1890.
Chapter Three

The first comprehensive study was performed by Fuller (1914). Gumbel (1941, 1942) first applied a particular extreme-value probability distribution to flood flows, whereas Chow (1954) extended the work using this distribution. A significant contribution to the study of rainfall frequencies was made by Yarnell (1936). The study analyzed rainfall durations lasting from 5 minutes to 24 hours and determined their frequency of occurrence at different locations within the continental United States. A similar study was performed by the Miami Conservancy District of Ohio for durations extending from 1 to 6 days (Engineering Staff of Miami Conservancy District, 1937). An extremal probability distribution was applied to rainfall data at Chicago, Illinois, by Chow (1953), and more recent frequency analysis of rainfall data was performed by the U.S. National Weather Service (Hershfield, 1964; U.S. Weather Bureau, 1964; Miller et al., 1973; Frederick et al., 1977, Huff and Angel, 1989, 1992). Low stream flows and droughts also were studied statistically by Gumbel (1954, 1963), who applied an extremal distribution to model the occurrences of drought frequencies. In the United Kingdom, hydrologic frequency analysis usually follows the procedures described in the Flood Studies Report of 1975 (National Environment Research Council, 1975). In general, frequency analysis is a useful analytical tool for studying randomly occurring events and need not be limited to hydrologic studies. Frequency analysis also has been applied to water quality studies and to ocean wave studies.

Basic probability concepts and theories useful for frequency analysis are described in Chap. 2. In general, there is no physical rule that requires the use of a particular distribution in the frequency analysis of geophysical data. However, since the maximum or minimum values of geophysical events are usually of interest, extreme-value-related distributions have been found to be most useful.

3.1 Types of Geophysical Data Series

The first step in the frequency-analysis process is to identify the set of data or sample to be studied. The sample is called a data series because many events of interest occur in a time sequence, and time is a useful frame of reference. The events are continuous, and thus their complete description as a function of time would constitute an infinite number of data points. To overcome this, it is customary to divide the events into a series of finite time increments and consider the average magnitude or instantaneous values of the largest or smallest within each interval. In frequency analysis, geophysical events that make up the data series generally are assumed to be statistically independent in time. In the United States, the water year concept was developed to facilitate the independence of hydrologic flood series. Throughout the eastern, southern, and Pacific western areas of the United States, the season of lowest stream flow is late summer and fall (August–October) (U.S. Department of Agriculture, 1955). Thus, by establishing the water year as October 1 to September 30, the chance of having related floods in each year is minimized, and the assumption of independence in the flood data is supported. In case time dependence is present.
in the data series and should be accounted for, procedures developed in time series analysis (Salas et al., 1980; Salas, 1993) should be applied. This means that the events themselves first must be identified in terms of a beginning and an end and then sampled using some criterion. Usually only one value from each event is included in the data series. There are three basic types of data series extractable from geophysical events:

1. A **complete series**, which includes all the available data on the magnitude of a phenomenon. A complete data series is used most frequently for flow-duration studies to determine the amount of firm power available in a proposed hydropower project or to study the low-flow behavior in water quality management. Such a data series is usually very large, and since in some instances engineers are only interested in the extremes of the distribution (e.g., floods, droughts, wind speeds, and wave heights), other data series often are more practical. For geophysical events, data in a complete series often exhibit significant time dependence, which makes the frequency-analysis procedure described herein inappropriate.

2. An **extreme-value series** is one that contains the largest (or smallest) data value for each of a number of equal time intervals. If, for example, the largest data value in each year of record is used, the extreme-value series is called an *annual maximum series*. If the smallest value is used, the series is called an *annual minimum series*.

3. A **partial-duration series** consists of all data above or below a base value. For example, one might consider only floods in a river with a magnitude greater than $1,000 \text{ m}^3/\text{s}$. When the base value is selected so that the number of events included in the data series equals the number of years of record, the resulting series is called an *annual exceedance series*. This series contains the $n$ largest or $n$ smallest values in $n$ years of record.

The selection of geophysical data series is illustrated in Fig. 3.1. Figure 3.1a represents the original data; the length of each line indicates the magnitude of the event. Figure 3.1b shows an annual maximum series with the largest data value in each year being retained for analysis. Figure 3.1c shows the data values that would be included in an annual exceedance series. Since there are 15 years of record, the 15 largest data values are retained. Figure 3.1d and e illustrate for comparison the rank in descending order of the magnitude of the events in each of the two series. As shown in Fig. 3.1d and e the annual maximum series and the annual exceedance series form different probability distributions, but when used to estimate extreme floods with return periods of 10 years or more, the differences between the results from the two series are minimal, and the annual maximum series is the one used most commonly. Thus this chapter focuses on the annual maximum series in the following discussion and examples.

Another issue related to the selection of the data series for frequency analysis is the adequacy of the record length. Benson (1952) generated randomly
Figure 3.1 Different types of data series: (a) original data series; (b) annual maximum series; (c) annual exceedance series; (d) ranked complete series; (e) ranked annual maximum and exceedance series (Chow, 1964).
selected values from known probability distributions and determined the record length necessary to estimate various probability events with acceptable error levels of 10 and 25 percent. Benson’s results are listed in Table 3.1. Linsley et al. (1982, p. 358) reported that similar simulation-based studies at Stanford University found that 80 percent of the estimates of the 100-year flood based on 20 years of record were too high and that 45 percent of the overestimates

<table>
<thead>
<tr>
<th>Design probability</th>
<th>Return period (years)</th>
<th>10% error (years)</th>
<th>25% error (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>10</td>
<td>90</td>
<td>18</td>
</tr>
<tr>
<td>0.02</td>
<td>50</td>
<td>110</td>
<td>39</td>
</tr>
<tr>
<td>0.01</td>
<td>100</td>
<td>115</td>
<td>48</td>
</tr>
</tbody>
</table>

Source: After Benson (1952).
exceeded 30 percent. The U.S. Water Resources Council (1967) recommended that at least 10 years of data should be available before a frequency analysis can be done. However, the results described in this section indicate that if a frequency analysis is done using 10 years of record, a high degree of uncertainty can be expected in the estimate of high-return-period events.

The final issue with respect to the data series used for frequency analysis is related to the problem of data homogeneity. For low-magnitude floods, peak stage is recorded at the gauge, and the discharge is determined from a rating curve established by current meter measurements of flows including similar-magnitude floods. In this case, the standard error of the measurement usually is less than 10 percent of the estimated discharge. For high-magnitude floods, peak stage often is inferred from high-water marks, and the discharge is computed by indirect means. For indirectly determined discharges, the standard error probably is several times larger, on the order of 16 to 30 percent (Potter and Walker, 1981). This is known as the discontinuous measurement error (DME) problem. Potter and Walker (1981) demonstrated that, as a result of DME, the probability distribution of measured floods can be greatly distorted with respect to the parent population. This further contributes to the uncertainty in flood frequency analysis.

3.2 Return Period

Hydrosystems engineers have been using the concept of the return period (or sometimes called recurrence interval) as a substitute for probability because it gives some physical interpretation to the probability. The return period for a given event is defined as the period of time on the long-term average at which a given event is equaled or exceeded. Hence, on average, an event with a 2-year return period will be equaled or exceeded once in 2 years. The relationship between the probability and return period is given by

\[
T = \frac{1}{P(X \geq x_T)} = \frac{1}{1 - P(X < x_T)}
\]  

in which \(x_T\) is the value of the variate corresponding to a \(T\)-year return period. For example, if the probability that a flood will be equaled or exceeded in a single year is 0.1, that is, \(P(X \geq x_T) = 0.1\), the corresponding return period is \(1/P(X \geq x_T) = 1/0.1 = 10\) years. Note that \(P(X \geq x_T)\) must be the probability that the event is equaled or exceeded in any one year and is the same for each year regardless of the magnitudes that occurred in prior years. This is so because the events are independent, and the long-term probabilities are used without regard to the order in which they may occur. A common error or misconception is to assume, for example, that if the 100-year event occurs this year, it will not occur again for the next 100 years. In fact, it could occur again next year and then not be repeated for several hundred years. This misconception resulted in considerable public complaints when the Phoenix area experienced two 50-year and one 100-year floods in a span of 18 months in 1978–1979 and the Milwaukee area experienced 100-year floods in June 1997 and June 1998.
Hence it is more appropriate and less confusing to use the odds ratio; e.g., the 100-year event can be described as the value having 1-in-100 chance being exceeded in any one year (Stedinger et al., 1993). In the United States in recent years it has become common practice to refer to the 100-year flood as the 1 percent chance exceedance flood, and similar percent chance exceedance descriptions are used for other flood magnitudes (U.S. Army Corps of Engineers, 1996).

The most common time unit for return period is the year, although semiannual, monthly, or any other time period may be used. The time unit used to form the time series will be the unit assigned to the return period. Thus an annual series will have a return-period unit of years, and a monthly series will have return-period unit of months. However, one should be careful about compliance with the statistical independence assumption for the data series. Many geophysical data series exhibit serial correlation when the time interval is short, which can be dealt with properly only by time-series analysis procedures (Salas, 1993).

3.3 Probability Estimates for Data Series: Plotting Positions (Rank-order Probability)

As stated previously, the objective of frequency analysis is to fit geophysical data to a probability distribution so that a relationship between the event magnitude and its exceedance probability can be established. The first step in the procedure is to determine the type of data series (i.e., event magnitude) to be used. In order to fit a probability distribution to the data series, estimates of probability (or equivalent return period) must be assigned to each magnitude in the data series.

Consider a data series consisting of the entire population of \( N \) values for a particular variable. If this series were ranked according to decreasing magnitude, it could be stated that the probability of the largest variate being equaled or exceeded is \( \frac{1}{N} \), where \( N \) is the total number of variates. Similarly, the exceedance probability of the second largest variate is \( \frac{2}{N} \), and so forth. In general,

\[
P(X \geq x_{(m)}) = \frac{1}{T_m} = \frac{m}{N}
\]  

(3.2)

in which \( m \) is the rank of the data in descending order, \( x_{(m)} \) is the \( m \)th largest variate in a data series of size \( N \), and \( T_m \) is the return period associated with \( x_{(m)} \). In practice, the entire population is not used or available. However, the reasoning leading to Eq. (3.2) is still valid, except that the result is now only an estimate of the exceedance probability based on a sample. Equation (3.2), which shows the ranked-order probability, is called a plotting position formula because it provides an estimate of probability so that the data series can be plotted (magnitudes versus probability).

Equation (3.2) is appropriate for data series from the population. Some modifications are made to avoid theoretical inconsistency when it is applied to sample data series. For example, Eq. (3.2) yields an exceedance probability of 1.0 for the smallest variate, implying that all values must be equal or larger. Since only
a sample is used, there is a likelihood that at some future time an event with a lower value could occur. In application, if the lower values in the series are not of great interest, this weakness can be overlooked, and in fact, Eq. (3.2) is used in the analysis of the annual exceedance series. A number of plotting-position formulas have been introduced that can be expressed in a general form as

\[ P(X \geq x_m) = u_m = \frac{1}{T_m} = \frac{m - a}{n + 1 - b} \]  

(3.3)

in which \( a \geq 0 \) and \( b \geq 0 \) are constants, and \( n \) is the number of observations in the sample data series. Table 3.2 lists several plotting-position formulas that have been developed and used in frequency analysis. Perhaps the most popular plotting-position formula is the Weibull formula (with \( a = 0 \) and \( b = 0 \)):

\[ P(X \geq x_m) = u_m = \frac{1}{T_m} = \frac{m}{n + 1} \]  

(3.4)

As shown in Table 3.2, it is noted that although these formulas give different results for the highest values in the series, they yield very similar results for the middle to lowest values, as seen in the last two columns.

Plotting-position formulas in the form of Eq. (3.3) can be categorized into being probability-unbiased and quantile-unbiased. The probability-unbiased

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula ( P(X \geq x_m) )</th>
<th>( T_m = \frac{1}{P(X \geq x_m)} ) for ( n = 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>California (1923)</td>
<td>( \frac{m}{n} )</td>
<td>20.0 2.00</td>
</tr>
<tr>
<td>Hazen (1930)</td>
<td>( \frac{m - 0.5}{n} )</td>
<td>40.0 2.11</td>
</tr>
<tr>
<td>Weibull (1939)</td>
<td>( \frac{m}{n + 1} )</td>
<td>41.0 2.10</td>
</tr>
<tr>
<td>Leivikov (1955)</td>
<td>( \frac{m - 0.3}{n + 0.4} )</td>
<td>29.1 2.10</td>
</tr>
<tr>
<td>Blom (1958)</td>
<td>( \frac{m - 0.375}{n + 0.25} )</td>
<td>24.5 2.10</td>
</tr>
<tr>
<td>Tukey (1962)</td>
<td>( \frac{m - 0.333}{n + 0.333} )</td>
<td>30.5 2.10</td>
</tr>
<tr>
<td>Gringorten (1963)</td>
<td>( \frac{m - 0.433}{n + 0.12} )</td>
<td>35.9 2.10</td>
</tr>
<tr>
<td>Cunnane (1978)</td>
<td>( \frac{m - 0.4}{n + 0.2} )</td>
<td>33.7 2.10</td>
</tr>
<tr>
<td>Hosking et al. (1985)</td>
<td>( \frac{m - 0.35}{n} )</td>
<td>30.7 2.07</td>
</tr>
</tbody>
</table>
plotting-position formula is concerned with finding a probability estimate $u_{(m)}$ for the exceedance probability of the $m$th largest observation such that $E[G(X_{(m)})] = u_{(m)}$, in which $G(X_{(m)}) = P(X \geq X_{(m)})$. In other words, the probability-unbiased plotting position yields the average exceedance probability for the $m$th largest observation in a sample of size $n$. If the data are independent random samples regardless of the underlying distribution, the estimator $U_{(m)} = G(X_{(m)})$ will have a beta distribution with the mean $E(U_{(m)}) = m/(n+1)$. Hence the Weibull plotting-position formula is probability-unbiased. On the other hand, Cunnane (1978) proposed quantile-unbiased plotting positions such that average value of the $m$th largest observation should be equal to $G^{-1}(u_{(m)})$, that is, $E(X_{(m)}) = G^{-1}(u_{(m)})$. The quantile-unbiased plotting-position formula, however, depends on the assumed distribution $G(\cdot)$. For example, referring to Table 3.2, the Blom plotting-position formula gives nearly unbiased quantiles for the normal distribution, and the Gringorton formula gives nearly unbiased quantiles for the Gumbel distribution. Cunnane’s formula, however, produces nearly quantile-unbiased plotting positions for a range of distributions.

### 3.4 Graphic Approach

Once the data series is identified and ranked and the plotting position is calculated, a graph of magnitude $x$ versus probability $[P(X \geq x), P(X < x), \text{or } T]$ can be plotted and a distribution fitted graphically. To facilitate this procedure, it is common to use some specially designed probability graph paper rather than linear graph paper. The probability scale in those special papers is chosen such that the resulting probability plot is a straight line. By plotting the data using a particular probability scale and constructing a best-fit straight line through the data, a graphic fit is made to the distribution used in constructing the probability scale. This is a graphic approach to estimate the statistical parameters of the distribution.

Example 3.1 illustrates the graphic approach to the analysis of flood data. The general procedure is as follows:

1. Identify the sample data series to be used. If high-return-period values are of interest, either the annual maximum or exceedance series can be used. If low-return-period values are of interest, use an annual exceedance series.

2. Rank the data series in decreasing order, and compute exceedance probability or return period using the appropriate plotting-position formula.

3. Obtain the probability paper corresponding to the distribution one wishes to fit to the data series.

4. Plot the series, and draw a best-fit straight line through the data. An eyeball fit or a mathematical procedure, such as the least-squares method, can be used. Before doing the fit, make a judgment regarding whether or not to include the unusual observations that do not lie near the line (termed outliers).
5. Extend the line to the highest return-period value needed, and read all required return-period values off the line.

Example 3.1  The Boneyard Creek stream gauging station was located near the fire station on the campus of the University of Illinois at Urbana–Champaign. From the USGS Water Supply Papers, the partial duration data of peak discharges above 400 ft$^3$/s between the water years 1961 and 1975 were obtained and listed below. In addition, for the years when there was no flow in a year exceeding 400 ft$^3$/s, the peak flow for that year is given in parenthesis (e.g., 1961).

<table>
<thead>
<tr>
<th>Year</th>
<th>Discharge, ft$^3$/s</th>
<th>Year</th>
<th>Discharge, ft$^3$/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1961</td>
<td>(390)</td>
<td>1969</td>
<td>549, 454</td>
</tr>
<tr>
<td>1962</td>
<td>(374)</td>
<td>1970</td>
<td>414, 410</td>
</tr>
<tr>
<td>1963</td>
<td>(342)</td>
<td>1971</td>
<td>434, 524</td>
</tr>
<tr>
<td>1964</td>
<td>507</td>
<td>1972</td>
<td>505, 415, 406</td>
</tr>
<tr>
<td>1965</td>
<td>579, 406, 596</td>
<td>1973</td>
<td>428, 447, 407</td>
</tr>
<tr>
<td>1966</td>
<td>416</td>
<td>1974</td>
<td>468, 543, 441</td>
</tr>
<tr>
<td>1967</td>
<td>533</td>
<td>1975</td>
<td>591, 497</td>
</tr>
<tr>
<td>1968</td>
<td>505</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) List the ranked annual maximum series. Also compute and list the corresponding plotting positions (return period) and exceedance probability $P(X \geq x)$.

(b) Plot the annual maximum series on (i) Gumbel paper and (ii) lognormal paper.

(c) Construct a best-fit line through the nonlinear plots, and estimate the flows for return periods of 2, 10, 25, and 50 years.

Solution  $n = 15$

<table>
<thead>
<tr>
<th>Annual Maximum Discharge (ft$^3$/s)</th>
<th>Rank (m)</th>
<th>$T_m = \frac{n+1}{m}$ (years)</th>
<th>$P(X \geq x_m) = 1/T_m$</th>
<th>$P(X &lt; x_m) = 1 - 1/T_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>596</td>
<td>1</td>
<td>16.00</td>
<td>0.0625</td>
<td>0.9375</td>
</tr>
<tr>
<td>591</td>
<td>2</td>
<td>8.00</td>
<td>0.1250</td>
<td>0.8750</td>
</tr>
<tr>
<td>549</td>
<td>3</td>
<td>5.33</td>
<td>0.1875</td>
<td>0.8125</td>
</tr>
<tr>
<td>543</td>
<td>4</td>
<td>4.00</td>
<td>0.2500</td>
<td>0.7500</td>
</tr>
<tr>
<td>533</td>
<td>5</td>
<td>3.20</td>
<td>0.3125</td>
<td>0.6875</td>
</tr>
<tr>
<td>524</td>
<td>6</td>
<td>2.67</td>
<td>0.3750</td>
<td>0.6250</td>
</tr>
<tr>
<td>507</td>
<td>7</td>
<td>2.29</td>
<td>0.4375</td>
<td>0.5625</td>
</tr>
<tr>
<td>505</td>
<td>8</td>
<td>2.00</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>505</td>
<td>9</td>
<td>1.78</td>
<td>0.5625</td>
<td>0.4375</td>
</tr>
<tr>
<td>447</td>
<td>10</td>
<td>1.60</td>
<td>0.6250</td>
<td>0.3750</td>
</tr>
<tr>
<td>416</td>
<td>11</td>
<td>1.46</td>
<td>0.6875</td>
<td>0.3125</td>
</tr>
<tr>
<td>414</td>
<td>12</td>
<td>1.33</td>
<td>0.7500</td>
<td>0.2500</td>
</tr>
<tr>
<td>390</td>
<td>13</td>
<td>1.23</td>
<td>0.8125</td>
<td>0.1875</td>
</tr>
<tr>
<td>374</td>
<td>14</td>
<td>1.14</td>
<td>0.8750</td>
<td>0.1250</td>
</tr>
<tr>
<td>342</td>
<td>15</td>
<td>1.06</td>
<td>0.9375</td>
<td>0.0625</td>
</tr>
</tbody>
</table>

(b) Plots of the annual maximum flow series on the Gumbel and lognormal probability papers are shown in Fig. 3.2.
Figure 3.2 Probability plot for the annual maximum series for 1961–1975 on the Boneyard Creek at Urbana, IL: (a) Gumbel probability plot; (b) lognormal probability plot.

(c) The following table summarizes the results read from the plots:

<table>
<thead>
<tr>
<th>Return Period (years)</th>
<th>2</th>
<th>10</th>
<th>25</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gumbel</td>
<td>470</td>
<td>610</td>
<td>680</td>
<td>730</td>
</tr>
<tr>
<td>Lognormal</td>
<td>475</td>
<td>590</td>
<td>650</td>
<td>700</td>
</tr>
</tbody>
</table>
3.5 Analytical Approaches

An alternative to the graphic technique is to estimate the statistical parameters of a distribution from the sample data (refer to Sec. 3.6). Then the distribution model can be used to solve for the variate value corresponding to any desired return period or probability as

\[ x_T = F^{-1}_x\left(1 - \frac{1}{T} \bigg| \theta \right) \]  

(3.5)

in which \( F^{-1}_x(\theta) \) is the inverse cumulative distribution function with the model parameter vector \( \theta \). Equation (3.5) can be applied when the inverse distribution function forms are analytically amenable, such as for the Gumbel, generalized extreme value, generalized logistic, and generalized Pareto distributions (see Sec. 2.6.6).

**Example 3.2** Consider that the annual maximum floods follow a lognormal distribution with a mean of 490 ft³/s and a standard deviation of 80 ft³/s. Determine the flood magnitude with a 1-in-100 chance of being exceeded in any given year.

**Solution** From Eqs. (2.67a) and (2.67b), the parameters of a lognormal distribution, for annual maximum flood \( Q \), can be obtained as

\[
\begin{align*}
\sigma_{\ln Q} &= \sqrt{\ln \left( \frac{80}{490} \right)^2 + 1} = 0.1622 \\
\mu_{\ln Q} &= \ln(490) - \frac{1}{2}(0.1622)^2 = 6.1812
\end{align*}
\]

Since \( \ln(Q) \) follows a normal distribution with a mean of \( \mu_{\ln Q} = 6.1812 \) and a standard deviation of \( \sigma_{\ln Q} = 0.1622 \) as previously computed, the magnitude of the log-transformed 100-year flood can be calculated by

\[
\frac{\ln(q_{100}) - \mu_{\ln Q}}{\sigma_{\ln Q}} = \phi^{-1}\left(1 - \frac{1}{100}\right) = \phi^{-1}(0.99) = 2.34
\]

Hence \( \ln(q_{100}) = \mu_{\ln Q} + 2.34 \times \sigma_{\ln Q} = 6.5607 \), and the corresponding 100-year flood magnitude can be calculated as \( q_{100} = \exp[\ln(q_{100})] = 706.8 \) ft³/s.

For some distributions, such as Pearson type 3 or log-Pearson type 3, the appropriate probability paper or CDF inverse form is unavailable. In such a case, an analytical approach using the *frequency factor* \( K_T \) is applied:

\[ x_T = \mu_x + K_T \times \sigma_x \]  

(3.6)

in which \( x_T \) is the variate corresponding to a return period of \( T \), \( \mu_x \) and \( \sigma_x \) are the mean and standard deviation of the random variable, respectively, and \( K_T \) is the frequency factor, which is a function of the return period \( T \) or \( P(X \geq x_T) \) and higher moments, if required. It is clear that a plot of Eq. (3.6) \((x_T \text{ versus } K_T)\)
on linear graph paper will yield a straight line with slope of $\sigma_x$ and intercept $\mu_x$ at $K_T = 0$.

In order for Eq. (3.6) to be useful, the functional relationship between $K_T$ and exceedance probability or return period must be determined for the distribution to be used. In fact, the frequency factor $K_T = (x_T - \mu_x)/\sigma_x$ is identical to a standardized variate corresponding to the exceedance probability of $1/T$ for a particular distribution model under consideration. For example, if the normal distribution is considered, then $K_T = z_T = \Phi^{-1}(1 - T^{-1})$. The same applies to the lognormal distribution when the mean and standard deviation of log-transformed random variables are used. Hence the standard normal probability table (Table 2.2) provides values of the frequency factor for sample data from normal and log normal distributions. Once this relation is known, a nonlinear probability or return-period scale can be constructed to replace the linear $K_T$ scale, and thus special graph paper can be constructed for any distribution so that plot of $x_T$ versus $P$ or $T$ will be linear.

Gumbel probability paper has been printed, although it is not readily available from commercial sources. Referring to Eq. (2.85a), the relationship between $K_T$ and $T$ for this distribution can be derived as

$$K_T = -\frac{\sqrt{6}}{\pi} \left[ 0.5772 + \ln \left( \frac{T}{T-1} \right) \right]$$  \hspace{1cm} (3.7)

For Pearson and log-Pearson type 3 distributions, linearization can be accomplished according to Eq. (3.6). However, for this distribution, the frequency factor is a function of both $P$ or $T$ and the skewness coefficient $\gamma_x$. This means that a different nonlinear $P$ or $T$ scale is required for each skewness coefficient, and therefore, it is impractical to construct a probability paper for this distribution. However, it should be pointed out that if $\gamma_x = 0$ in log-space, the log-Pearson type 3 reduces to the lognormal distribution, and thus commercial lognormal probability paper can be used. The relationship between frequency factor $K_T$, $T$, and $\gamma_x$ cannot be developed in a closed form, as was done for the Gumbel distribution in Eq. (3.7). However, the relationship can be computed numerically, and the results are given in Table 3.3. For $0.99^{-1} \leq T \leq 100$ and $|\gamma_x| < 2$, the frequency-factor values are well approximated by the Wilson-Hilferty transformation (Stedinger et al., 1993):

$$K_T(\gamma_x) = \frac{2}{\gamma_x} \left\{ 1 + z_T \left( \frac{\gamma_x}{6} \right) - \left( \frac{\gamma_x}{6} \right)^2 \right\}^3 - 1$$  \hspace{1cm} (3.8)

in which $z_T$ is the standard normal quantile with exceedance probability of $1/T$.

The procedure for using the frequency-factor method is outlined as follows:

1. Compute the sample mean $\bar{x}$, standard deviation $\sigma_x$, and skewness coefficient $\gamma_x$ (if needed) for the sample.
### TABLE 3.3 Frequency Factor ($K_f$) for Pearson Type 3 Distribution

<table>
<thead>
<tr>
<th>Skewness coefficient $y_x$</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>-0.396</td>
<td>0.420</td>
<td>1.180</td>
<td>2.278</td>
<td>3.152</td>
<td>4.051</td>
<td>4.970</td>
</tr>
<tr>
<td>2.9</td>
<td>-0.390</td>
<td>0.440</td>
<td>1.195</td>
<td>2.277</td>
<td>3.134</td>
<td>4.013</td>
<td>4.909</td>
</tr>
<tr>
<td>2.8</td>
<td>-0.384</td>
<td>0.460</td>
<td>1.210</td>
<td>2.275</td>
<td>3.114</td>
<td>3.973</td>
<td>4.847</td>
</tr>
<tr>
<td>2.7</td>
<td>-0.376</td>
<td>0.479</td>
<td>1.224</td>
<td>2.272</td>
<td>3.093</td>
<td>3.932</td>
<td>4.783</td>
</tr>
<tr>
<td>2.6</td>
<td>-0.368</td>
<td>0.499</td>
<td>1.238</td>
<td>2.267</td>
<td>3.071</td>
<td>3.889</td>
<td>4.718</td>
</tr>
<tr>
<td>2.5</td>
<td>-0.360</td>
<td>0.518</td>
<td>1.250</td>
<td>2.262</td>
<td>3.048</td>
<td>3.845</td>
<td>4.652</td>
</tr>
<tr>
<td>2.4</td>
<td>-0.351</td>
<td>0.537</td>
<td>1.262</td>
<td>2.256</td>
<td>3.023</td>
<td>3.800</td>
<td>4.584</td>
</tr>
<tr>
<td>2.3</td>
<td>-0.341</td>
<td>0.555</td>
<td>1.274</td>
<td>2.248</td>
<td>2.997</td>
<td>3.753</td>
<td>4.515</td>
</tr>
<tr>
<td>2.2</td>
<td>-0.330</td>
<td>0.574</td>
<td>1.284</td>
<td>2.240</td>
<td>2.970</td>
<td>3.705</td>
<td>4.444</td>
</tr>
<tr>
<td>2.1</td>
<td>-0.319</td>
<td>0.592</td>
<td>1.294</td>
<td>2.230</td>
<td>2.942</td>
<td>3.656</td>
<td>4.372</td>
</tr>
<tr>
<td>2.0</td>
<td>-0.307</td>
<td>0.609</td>
<td>1.302</td>
<td>2.219</td>
<td>2.912</td>
<td>3.605</td>
<td>4.298</td>
</tr>
<tr>
<td>1.9</td>
<td>-0.294</td>
<td>0.627</td>
<td>1.310</td>
<td>2.207</td>
<td>2.881</td>
<td>3.553</td>
<td>4.223</td>
</tr>
<tr>
<td>1.8</td>
<td>-0.282</td>
<td>0.643</td>
<td>1.318</td>
<td>2.193</td>
<td>2.848</td>
<td>3.499</td>
<td>4.147</td>
</tr>
<tr>
<td>1.7</td>
<td>-0.268</td>
<td>0.660</td>
<td>1.324</td>
<td>2.179</td>
<td>2.815</td>
<td>3.444</td>
<td>4.069</td>
</tr>
<tr>
<td>1.6</td>
<td>-0.254</td>
<td>0.675</td>
<td>1.329</td>
<td>2.163</td>
<td>2.780</td>
<td>3.388</td>
<td>3.990</td>
</tr>
<tr>
<td>1.5</td>
<td>-0.240</td>
<td>0.690</td>
<td>1.333</td>
<td>2.146</td>
<td>2.743</td>
<td>3.330</td>
<td>3.910</td>
</tr>
<tr>
<td>1.4</td>
<td>-0.225</td>
<td>0.705</td>
<td>1.337</td>
<td>2.128</td>
<td>2.706</td>
<td>3.271</td>
<td>3.828</td>
</tr>
<tr>
<td>1.3</td>
<td>-0.210</td>
<td>0.719</td>
<td>1.339</td>
<td>2.108</td>
<td>2.666</td>
<td>3.211</td>
<td>3.745</td>
</tr>
<tr>
<td>1.2</td>
<td>-0.195</td>
<td>0.732</td>
<td>1.340</td>
<td>2.087</td>
<td>2.626</td>
<td>3.149</td>
<td>3.661</td>
</tr>
<tr>
<td>1.1</td>
<td>-0.180</td>
<td>0.745</td>
<td>1.341</td>
<td>2.066</td>
<td>2.585</td>
<td>3.087</td>
<td>3.575</td>
</tr>
<tr>
<td>1.0</td>
<td>-0.164</td>
<td>0.758</td>
<td>1.340</td>
<td>2.043</td>
<td>2.542</td>
<td>3.022</td>
<td>3.489</td>
</tr>
<tr>
<td>0.9</td>
<td>-0.148</td>
<td>0.769</td>
<td>1.339</td>
<td>2.018</td>
<td>2.498</td>
<td>2.957</td>
<td>3.401</td>
</tr>
<tr>
<td>0.8</td>
<td>-0.132</td>
<td>0.780</td>
<td>1.336</td>
<td>1.993</td>
<td>2.453</td>
<td>2.891</td>
<td>3.312</td>
</tr>
<tr>
<td>0.7</td>
<td>-0.116</td>
<td>0.790</td>
<td>1.333</td>
<td>1.967</td>
<td>2.407</td>
<td>2.824</td>
<td>3.223</td>
</tr>
<tr>
<td>0.6</td>
<td>-0.099</td>
<td>0.800</td>
<td>1.328</td>
<td>1.939</td>
<td>2.359</td>
<td>2.755</td>
<td>3.132</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.083</td>
<td>0.808</td>
<td>1.323</td>
<td>1.910</td>
<td>2.311</td>
<td>2.686</td>
<td>3.041</td>
</tr>
<tr>
<td>0.4</td>
<td>-0.066</td>
<td>0.816</td>
<td>1.317</td>
<td>1.880</td>
<td>2.261</td>
<td>2.615</td>
<td>2.949</td>
</tr>
<tr>
<td>0.3</td>
<td>-0.050</td>
<td>0.824</td>
<td>1.309</td>
<td>1.849</td>
<td>2.211</td>
<td>2.544</td>
<td>2.856</td>
</tr>
<tr>
<td>0.2</td>
<td>-0.033</td>
<td>0.830</td>
<td>1.301</td>
<td>1.818</td>
<td>2.159</td>
<td>2.472</td>
<td>2.763</td>
</tr>
<tr>
<td>0.1</td>
<td>-0.017</td>
<td>0.836</td>
<td>1.292</td>
<td>1.785</td>
<td>2.107</td>
<td>2.400</td>
<td>2.670</td>
</tr>
<tr>
<td>0.0</td>
<td>0</td>
<td>0.842</td>
<td>1.282</td>
<td>1.751</td>
<td>2.054</td>
<td>2.326</td>
<td>2.576</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.017</td>
<td>0.846</td>
<td>1.270</td>
<td>1.716</td>
<td>2.000</td>
<td>2.252</td>
<td>2.482</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.033</td>
<td>0.850</td>
<td>1.258</td>
<td>1.680</td>
<td>1.945</td>
<td>2.178</td>
<td>2.388</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.050</td>
<td>0.853</td>
<td>1.245</td>
<td>1.643</td>
<td>1.890</td>
<td>2.104</td>
<td>2.294</td>
</tr>
<tr>
<td>-0.4</td>
<td>0.066</td>
<td>0.855</td>
<td>1.231</td>
<td>1.606</td>
<td>1.834</td>
<td>2.029</td>
<td>2.201</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.083</td>
<td>0.856</td>
<td>1.216</td>
<td>1.567</td>
<td>1.777</td>
<td>1.955</td>
<td>2.108</td>
</tr>
<tr>
<td>-0.6</td>
<td>0.099</td>
<td>0.857</td>
<td>1.200</td>
<td>1.528</td>
<td>1.720</td>
<td>1.880</td>
<td>2.016</td>
</tr>
<tr>
<td>-0.7</td>
<td>0.116</td>
<td>0.857</td>
<td>1.183</td>
<td>1.488</td>
<td>1.663</td>
<td>1.806</td>
<td>1.926</td>
</tr>
<tr>
<td>-0.8</td>
<td>0.132</td>
<td>0.856</td>
<td>1.166</td>
<td>1.448</td>
<td>1.606</td>
<td>1.733</td>
<td>1.837</td>
</tr>
<tr>
<td>-0.9</td>
<td>0.148</td>
<td>0.854</td>
<td>1.147</td>
<td>1.407</td>
<td>1.549</td>
<td>1.660</td>
<td>1.749</td>
</tr>
<tr>
<td>-1.0</td>
<td>0.164</td>
<td>0.852</td>
<td>1.128</td>
<td>1.366</td>
<td>1.492</td>
<td>1.588</td>
<td>1.664</td>
</tr>
<tr>
<td>-1.1</td>
<td>0.180</td>
<td>0.848</td>
<td>1.107</td>
<td>1.324</td>
<td>1.435</td>
<td>1.518</td>
<td>1.581</td>
</tr>
<tr>
<td>-1.2</td>
<td>0.195</td>
<td>0.844</td>
<td>1.086</td>
<td>1.282</td>
<td>1.379</td>
<td>1.449</td>
<td>1.501</td>
</tr>
<tr>
<td>-1.3</td>
<td>0.210</td>
<td>0.838</td>
<td>1.064</td>
<td>1.240</td>
<td>1.324</td>
<td>1.383</td>
<td>1.424</td>
</tr>
<tr>
<td>-1.4</td>
<td>0.225</td>
<td>0.832</td>
<td>1.041</td>
<td>1.198</td>
<td>1.270</td>
<td>1.318</td>
<td>1.351</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.240</td>
<td>0.825</td>
<td>1.018</td>
<td>1.157</td>
<td>1.217</td>
<td>1.256</td>
<td>1.282</td>
</tr>
<tr>
<td>-1.6</td>
<td>0.254</td>
<td>0.817</td>
<td>0.994</td>
<td>1.116</td>
<td>1.166</td>
<td>1.197</td>
<td>1.216</td>
</tr>
</tbody>
</table>

(Continued)
2. For the desired return period, determine the associated value of $K_T$ for the distribution.

3. Compute the desired quantile value using Eq. (3.6) with $\bar{x}$ replacing $\mu_x$ and $s_x$ replacing $\sigma_x$, that is,

$$\hat{x}_T = \bar{x} + K_T \times s_x \quad \text{(3.9)}$$

It should be recognized that the basic difference between the graphic and analytical approaches is that each represents a different method of estimating the statistical parameters of the distribution being used. By the analytical approach, a best-fit line is constructed that then sets the statistical parameters. In the mathematical approach, the statistical parameters are first computed from the sample, and effectively, the line thus determined is used. The line determined using the mathematical approach is in general a poorer fit to the observed data than that obtained using the graphic approach, especially if curve-fitting procedures are applied. However, the U.S. Water Resources Council (1967) recommended use of the mathematical approach because

1. Graphic least-squares methods are avoided to reduce the incorporation of the random characteristics of the particular data set (especially in the light of the difficulty in selecting the proper plotting-position formula).

2. The generally larger variance of the mathematical approach is believed to help compensate for the typically small data sets.

### TABLE 3.3 Frequency Factor ($K_T$) for Pearson Type 3 Distribution (Continued)

<table>
<thead>
<tr>
<th>Skewness coefficient $\gamma_x$</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-1.7$</td>
<td>0.268</td>
<td>0.808</td>
<td>0.970</td>
<td>1.075</td>
<td>1.116</td>
<td>1.140</td>
<td>1.155</td>
</tr>
<tr>
<td>$-1.8$</td>
<td>0.282</td>
<td>0.799</td>
<td>0.945</td>
<td>1.035</td>
<td>1.069</td>
<td>1.087</td>
<td>1.097</td>
</tr>
<tr>
<td>$-1.9$</td>
<td>0.294</td>
<td>0.788</td>
<td>0.920</td>
<td>0.996</td>
<td>1.023</td>
<td>1.037</td>
<td>1.044</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>0.307</td>
<td>0.777</td>
<td>0.895</td>
<td>0.959</td>
<td>0.980</td>
<td>0.990</td>
<td>0.995</td>
</tr>
<tr>
<td>$-2.1$</td>
<td>0.319</td>
<td>0.765</td>
<td>0.869</td>
<td>0.923</td>
<td>0.939</td>
<td>0.946</td>
<td>0.949</td>
</tr>
<tr>
<td>$-2.2$</td>
<td>0.330</td>
<td>0.752</td>
<td>0.844</td>
<td>0.888</td>
<td>0.900</td>
<td>0.905</td>
<td>0.907</td>
</tr>
<tr>
<td>$-2.3$</td>
<td>0.341</td>
<td>0.739</td>
<td>0.819</td>
<td>0.855</td>
<td>0.864</td>
<td>0.867</td>
<td>0.869</td>
</tr>
<tr>
<td>$-2.4$</td>
<td>0.351</td>
<td>0.725</td>
<td>0.795</td>
<td>0.823</td>
<td>0.830</td>
<td>0.832</td>
<td>0.833</td>
</tr>
<tr>
<td>$-2.5$</td>
<td>0.360</td>
<td>0.711</td>
<td>0.771</td>
<td>0.793</td>
<td>0.798</td>
<td>0.799</td>
<td>0.800</td>
</tr>
<tr>
<td>$-2.6$</td>
<td>0.368</td>
<td>0.696</td>
<td>0.747</td>
<td>0.764</td>
<td>0.768</td>
<td>0.769</td>
<td>0.769</td>
</tr>
<tr>
<td>$-2.7$</td>
<td>0.376</td>
<td>0.681</td>
<td>0.724</td>
<td>0.738</td>
<td>0.740</td>
<td>0.740</td>
<td>0.741</td>
</tr>
<tr>
<td>$-2.8$</td>
<td>0.384</td>
<td>0.666</td>
<td>0.702</td>
<td>0.712</td>
<td>0.714</td>
<td>0.714</td>
<td>0.714</td>
</tr>
<tr>
<td>$-2.9$</td>
<td>0.390</td>
<td>0.651</td>
<td>0.681</td>
<td>0.683</td>
<td>0.689</td>
<td>0.690</td>
<td>0.690</td>
</tr>
<tr>
<td>$-3.0$</td>
<td>0.396</td>
<td>0.636</td>
<td>0.666</td>
<td>0.666</td>
<td>0.666</td>
<td>0.667</td>
<td>0.667</td>
</tr>
</tbody>
</table>

Example 3.3 Using the frequency-factor method, estimate the flows with return periods of 2, 10, 25, 50, and 100 years for the Boneyard Creek using the Gumbel and log-Pearson type 3 distributions. Use the historical data in Example 3.1 as a basis for the calculations.

Solution Based on the samples, the method requires determination of the frequency factor $K_T$ in

$$\hat{x}_T = \hat{x} + K_T \times s$$

For the Gumbel distribution, values of $K_T$ can be calculated by Eq. (3.7). For the log-Pearson type 3 distribution, Table 3.3 or Eq. (3.8) can be used, which requires computation of the skewness coefficient. The calculations of relevant sample moments are shown in the following table:

<table>
<thead>
<tr>
<th>Year</th>
<th>$q_i$ (ft$^3$/s)</th>
<th>$q_i^2$</th>
<th>$q_i^3$</th>
<th>$y_i = \ln(q_i)$</th>
<th>$y_i^2$</th>
<th>$y_i^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1961</td>
<td>390</td>
<td>1.52e+05</td>
<td>5.93e+07</td>
<td>5.97</td>
<td>35.59</td>
<td>212.36</td>
</tr>
<tr>
<td>1962</td>
<td>374</td>
<td>1.40e+05</td>
<td>5.22e+07</td>
<td>5.92</td>
<td>35.10</td>
<td>207.92</td>
</tr>
<tr>
<td>1963</td>
<td>342</td>
<td>1.17e+05</td>
<td>4.00e+07</td>
<td>5.83</td>
<td>34.05</td>
<td>198.65</td>
</tr>
<tr>
<td>1964</td>
<td>507</td>
<td>2.57e+05</td>
<td>1.30e+08</td>
<td>6.23</td>
<td>38.79</td>
<td>241.63</td>
</tr>
<tr>
<td>1965</td>
<td>596</td>
<td>3.55e+05</td>
<td>2.12e+08</td>
<td>6.39</td>
<td>40.84</td>
<td>260.95</td>
</tr>
<tr>
<td>1966</td>
<td>416</td>
<td>1.73e+05</td>
<td>7.20e+07</td>
<td>6.03</td>
<td>36.37</td>
<td>219.33</td>
</tr>
<tr>
<td>1967</td>
<td>533</td>
<td>2.84e+05</td>
<td>1.51e+08</td>
<td>6.28</td>
<td>39.42</td>
<td>247.50</td>
</tr>
<tr>
<td>1968</td>
<td>505</td>
<td>2.55e+05</td>
<td>1.29e+08</td>
<td>6.22</td>
<td>38.75</td>
<td>241.17</td>
</tr>
<tr>
<td>1969</td>
<td>549</td>
<td>3.01e+05</td>
<td>1.65e+08</td>
<td>6.31</td>
<td>39.79</td>
<td>251.01</td>
</tr>
<tr>
<td>1970</td>
<td>414</td>
<td>1.71e+05</td>
<td>7.10e+07</td>
<td>6.03</td>
<td>36.31</td>
<td>218.81</td>
</tr>
<tr>
<td>1971</td>
<td>524</td>
<td>2.75e+05</td>
<td>1.44e+08</td>
<td>6.26</td>
<td>39.21</td>
<td>245.49</td>
</tr>
<tr>
<td>1972</td>
<td>505</td>
<td>2.55e+05</td>
<td>1.29e+08</td>
<td>6.22</td>
<td>38.75</td>
<td>241.17</td>
</tr>
<tr>
<td>1973</td>
<td>447</td>
<td>2.00e+05</td>
<td>8.93e+07</td>
<td>6.10</td>
<td>37.24</td>
<td>227.27</td>
</tr>
<tr>
<td>1974</td>
<td>543</td>
<td>2.95e+05</td>
<td>1.60e+08</td>
<td>6.30</td>
<td>39.65</td>
<td>249.70</td>
</tr>
<tr>
<td>1975</td>
<td>591</td>
<td>3.49e+05</td>
<td>2.06e+08</td>
<td>6.38</td>
<td>40.73</td>
<td>259.92</td>
</tr>
<tr>
<td>Sum</td>
<td>7236</td>
<td>3.58e+06</td>
<td>1.81e+09</td>
<td>92.48</td>
<td>570.58</td>
<td>3522.88</td>
</tr>
</tbody>
</table>

For the Gumbel distribution,

$$\bar{q} = \frac{7236}{15} = 482.4 \text{ft}^3/\text{s}$$

$$s = \left( \frac{\sum q_i^2 - 15\bar{q}^2}{15-1} \right)^{1/2} = (6361.8)^{1/2} = 79.8 \text{ft}^3/\text{s}$$

For the log-Pearson type 3 distribution,

$$\bar{y} = \frac{\sum \ln(q_i)}{n} = \frac{92.48}{15} = 6.165$$

$$s_y = \left( \frac{\sum y_i^2 - 15\bar{y}^2}{15-1} \right)^{1/2} = (0.417/14)^{1/2} = 0.173$$

$$g_y = \frac{n}{(n-1)(n-2)} \frac{m_3}{s_y^3} = \frac{15(-0.00336)}{(14)(13)(0.173)^3} = -0.540$$
in which \( m_3 = \sum y_i^3 - 3\bar{y} \sum y_i^2 + 2n\bar{y}^3 \). The determination of the values of frequency factor corresponding to different return periods is shown in the following table:

<table>
<thead>
<tr>
<th>Return period (years)</th>
<th>Exceedance probability</th>
<th>Nonexceedance probability</th>
<th>Frequency factor by distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gumbel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Eq. (3.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LP3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Eq. (3.8)</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
<td>0.50</td>
<td>0.0892</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>0.90</td>
<td>1.3046</td>
</tr>
<tr>
<td>25</td>
<td>0.04</td>
<td>0.96</td>
<td>2.0438</td>
</tr>
<tr>
<td>50</td>
<td>0.02</td>
<td>0.98</td>
<td>2.5923</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
<td>0.99</td>
<td>3.1367</td>
</tr>
</tbody>
</table>

Based on the preceding frequency-factor values, the flood magnitude of the various return periods can be determined as

<table>
<thead>
<tr>
<th>Return period (years)</th>
<th>Frequency curves by distribution (ft³/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gumbel</td>
</tr>
<tr>
<td></td>
<td>( q_T = 482.4 + 79.8K_T ) (EV_1)</td>
</tr>
<tr>
<td></td>
<td>LP3</td>
</tr>
<tr>
<td></td>
<td>( q_T = \exp(6.165 + 0.173K_T - \text{LP3}) )</td>
</tr>
<tr>
<td>2</td>
<td>469.3</td>
</tr>
<tr>
<td>10</td>
<td>586.5</td>
</tr>
<tr>
<td>25</td>
<td>645.4</td>
</tr>
<tr>
<td>50</td>
<td>689.2</td>
</tr>
<tr>
<td>100</td>
<td>732.6</td>
</tr>
</tbody>
</table>

One could compare these results for the Gumbel distribution with those obtained from the graphic approach of Example 3.1.

### 3.6 Estimation of Distributional Parameters

For a chosen distributional model, its shape and position are completely defined by the associated parameters. By referring to Eq. (3.5), determination of the quantile also requires knowing the values of the parameters \( \theta \).

There are several methods for estimating the parameters of a distribution model on the basis of available data. In frequency analysis, the commonly used parameter-estimation procedures are the method of maximum likelihood and the methods of moments (Kite, 1988; Haan, 1977). Other methods, such as method of maximum entropy (Li et al., 1986), have been applied.

#### 3.6.1 Maximum-likelihood (ML) method

This method determines the values of parameters of a distribution model that maximizes the likelihood of the sample data at hand. For a sample of \( n \) independent random observations, \( x = (x_1, x_2, \ldots, x_n)^T \), from an identical distribution, that is,

\[
x_i \overset{iid}{\sim} f_x(x | \theta) \quad \text{for} \ i = 1, 2, \ldots, n
\]
in which \( \theta = (\theta_1, \theta_2, \ldots, \theta_m)^T \), a vector of \( m \) distribution model parameters, the likelihood of occurrence of the samples is equal to the joint probability of \( \{x_i\}_{i=1,2,\ldots,n} \) calculable by

\[
L(x | \theta) = \prod_{i=1}^{n} f_x(x_i | \theta)
\]  

(3.10)

in which \( L(x | \theta) \) is called the likelihood function. The ML method determines the distribution parameters by solving

\[
\max_{\theta} L(x | \theta) = \max_{\theta} \ln[L(x | \theta)]
\]

or more specifically

\[
\max_{\theta} \prod_{i=1}^{n} f_x(x_i | \theta) = \max_{\theta} \sum_{i=1}^{n} \ln[f_x(x_i | \theta)]
\]

(3.11)

As can be seen, solving for distribution-model parameters by the ML principle is an unconstrained optimization problem. The unknown model parameters can be obtained by solving the following necessary conditions for the maximum:

\[
\frac{\partial}{\partial \theta_j} \left\{ \sum_{i=1}^{n} \ln[f_x(x_i | \theta)] \right\} = 0 \quad \text{for} \; j = 1, 2, \ldots, m
\]

(3.12)

In general, depending on the form of the distribution model under consideration, Eq. (3.12) could be a system of nonlinear equations requiring numerical solution. Alternatively, Eq. (3.11) can be solved by a proper direct optimum search scheme, such as the conjugate gradient method or quasi-Newton method (McCormick, 1983 or see Section 8.13.2).

**Example 3.4** Referring to Eq. (2.79) for the exponential distribution as

\[
f_x(x | \beta) = \exp(-x/\beta)/\beta \quad \text{for} \; x > 0, \beta > 0
\]

determine the maximum likelihood estimate for \( \beta \) based on \( n \) independent random samples \( \{x_i\}_{i=1,2,\ldots,n} \).

**Solution** The log-likelihood function for the exponential distribution is

\[
\ln[L(x | \beta)] = \sum_{i=1}^{n} \ln \left( \frac{1}{\beta} e^{-x_i/\beta} \right) = -n \ln(\beta) - \frac{1}{\beta} \sum_{i=1}^{n} x_i
\]

The parameter \( \beta \) that maximizes the preceding log-likelihood function, according to the necessary condition, i.e., Eq. (3.12), is

\[
\frac{\partial}{\partial \beta} [\ln[L(x | \beta)]] = - \frac{n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} x_i = 0
\]
Hence the ML estimator of $\beta$ for the exponential distribution is

$$\hat{\beta}_{ML} \frac{1}{n} \sum_{i=1}^{n} x_i$$

which is the sample mean.

**Example 3.5** Consider a set of $n$ independent samples, $x = (x_1, x_2, \ldots, x_n)^t$, from a normal distribution with the following PDF:

$$f_x(x|\alpha, \beta) = \frac{1}{\beta \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \alpha}{\beta}\right)^2} \quad \text{for } -\infty < x < \infty$$

Determine the ML estimators for the parameters $\alpha$ and $\beta$.

**Solution** The likelihood function for the $n$ independent normal samples is

$$L(x|\alpha, \beta) = \left(\frac{1}{\beta \sqrt{2\pi}}\right)^n \exp\left[-\frac{\sum_{i=1}^{n} (x_i - \alpha)^2}{2\beta^2}\right]$$

The corresponding log-likelihood function can be written as

$$\ln[L(x|\alpha, \beta)] = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\beta^2) - \frac{\sum_{i=1}^{n} (x_i - \alpha)^2}{2\beta^2}$$

Taking the partial derivatives of the preceding log-likelihood function with respect to $\alpha$ and $\beta^2$ and setting them equal to zero results in

$$\frac{\partial}{\partial \alpha} [\ln[L(x|\alpha, \beta)]] = \frac{\sum_{i=1}^{n} (x_i - \alpha)}{\beta^2} = 0$$

$$\frac{\partial}{\partial \beta^2} [\ln[L(x|\alpha, \beta)]] = -\frac{n}{2\beta^2} + \frac{\sum_{i=1}^{n} (x_i - \alpha)^2}{2\beta^4} = 0$$

After some algebraic manipulations, one can easily obtain the ML estimates of normal distribution parameters $\alpha$ and $\beta$ as

$$\hat{\alpha}_{ML} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \hat{\beta^2}_{ML} = \frac{\sum_{i=1}^{n} (x_i - \alpha)^2}{n}$$

As can be seen, the ML estimation of the normal parameters for $\alpha$ is the sample mean and for $\beta^2$ is a biased variance.

### 3.6.2 Product-moments-based method

By the moment-based parameter-estimation methods, parameters of a distribution are related to the statistical moments of the random variable. The conventional method of moments uses the product moments of the random variable. Example 3.3 for frequency analysis is typical of this approach. When sample data are available, sample product moments are used to solve for the model parameters. The main concern with the use of product moments is that their reliabilities owing to sampling errors deteriorate rapidly as the order of moment
increases, especially when sample size is small (see Sec. 3.1), which is often the case in many geophysical applications. Hence, in practice only, the first few statistical moments are used. Relationships between product-moments and parameters of distribution models commonly used in frequency analysis are listed in Table 3.4.

3.6.3 L-moments-based method

As described in Sec. 2.4.1, the L-moments are linear combinations of order statistics (Hosking, 1986). In theory, the estimators of L-moments are less sensitive to the presence of outliers in the sample and hence are more robust than the conventional product moments. Furthermore, estimators of L-moments are less biased and approach the asymptotic normal distributions more rapidly and closely. Hosking (1986) shows that parameter estimates from the L-moments are sometimes more accurate in small samples than are the maximum-likelihood estimates.

To calculate sample L-moments, one can refer to the probability-weighted moments as

\[ \beta_r = M_{l,r,0} = E\{X[F_x(X)]^r\} \quad \text{for } r = 0, 1, \ldots \]  

where \( M_{l,r,0} \) is defined on the basis of nonexceedance probability or CDF. The estimation of \( \beta_r \) then is hinged on how \( F_x(X) \) is estimated on the basis of sample data.

Consider \( n \) independent samples arranged in ascending order as \( X_{(n)} \leq X_{(n-1)} \leq \cdots \leq X_{(2)} \leq X_{(1)} \). The estimator for \( F_x(X_{(m)}) \) for the \( m \)th-order statistic can use an appropriate plotting-position formula as shown in Table 3.2, that is,

\[ \hat{F}(X_{(m)}) = 1 - \frac{m - a}{n + 1 - b} \quad \text{for } m = 1, 2, \ldots, n \]

with \( a \geq 0 \) and \( b \geq 0 \). The Weibull plotting-position formula \((a = 0, b = 0)\) is a probability-unbiased estimator of \( F_x(X_{(m)}) \). Hosking et al. (1985a, 1985b) show that a smaller mean square error in the quantile estimate can be achieved by using a biased plotting-position formula with \( a = 0.35 \) and \( b = 1 \). According to the definition of the \( \beta \)-moment \( \beta_r \) in Eq. (3.13), its sample estimate \( \beta_r \) can be obtained easily as

\[ \beta_r = \frac{1}{n} \sum_{i=1}^{n} x_{(m)}[\hat{F}(x_{(m)})]^r \quad \text{for } r = 0, 1, \ldots \]  

Stedinger et al. (1993) recommend the use of the quantile-unbiased estimator of \( F_x(X_{(m)}) \) for calculating the L-moment ratios in at-site and regional frequency analyses.
TABLE 3.4 Relations between Moments and Parameters of Selected Distribution Models

<table>
<thead>
<tr>
<th>Distribution</th>
<th>PDF or CDF</th>
<th>Range</th>
<th>Product moments</th>
<th>L-Moments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( f_x(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left[-\frac{1}{2} \left( \frac{x-\mu}{\sigma}\right)^2\right] )</td>
<td>(-\infty &lt; x &lt; \infty)</td>
<td>( \lambda_1 = \mu; \lambda_2 = \sigma / \sqrt{\pi}; )</td>
<td>( \lambda_3 = 0; \lambda_4 = 0.1226 )</td>
</tr>
<tr>
<td>Lognormal</td>
<td>( f_x(x) = \frac{1}{\sqrt{2\pi} \sigma_\text{ln} x} \exp\left[-\frac{1}{2} \left( \ln x - \mu_{\ln x} \right)^2 / \sigma_{\ln x}^2\right] )</td>
<td>( x &gt; 0 )</td>
<td>( \mu_{\ln x} = \ln \mu_x - \sigma_{\ln x}^2 / 2; )</td>
<td>( \gamma_x = 3\Omega_x + \Omega_x^3 )</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>( f_x(x) = \frac{(x-\xi)}{\sigma} \exp\left[-\frac{1}{2} \left( \frac{x-\xi}{\sigma}\right)^2\right] )</td>
<td>( \xi \leq x &lt; \infty )</td>
<td>( \mu = \xi + \sqrt{2\pi} \sigma; )</td>
<td>Eq. (2.68); Eq. (2.70)</td>
</tr>
<tr>
<td>Pearson 3</td>
<td>( f_x(x) = \frac{1}{\beta \Gamma(\alpha)} \left( \frac{\alpha^\beta}{\pi^{\frac{\beta}{2}}} \right) e^{-\left(\frac{x-\xi}{\beta}\right)^{\frac{2}{\beta}}} )</td>
<td>( x &gt; 0 )</td>
<td>( \mu = \xi + \alpha \beta; )</td>
<td>( \gamma = \text{sign}(\beta) \frac{2}{\sqrt{\beta}} )</td>
</tr>
<tr>
<td>Exponential</td>
<td>( f_x(x) = e^{-x/\beta} / \beta )</td>
<td>( x &gt; 0 )</td>
<td>( \mu = \beta )</td>
<td>( \lambda_1 = \beta; \lambda_2 = \beta / 2; )</td>
</tr>
<tr>
<td>Gumbel (EV1 for maxima)</td>
<td>( f_x(x) = \frac{\beta}{\pi} \left( \frac{x-\xi}{\sigma}\right)^{\alpha-1} \exp\left[-\left(\frac{x-\xi}{\sigma}\right)^{2}\right] )</td>
<td>(-\infty &lt; x &lt; \infty )</td>
<td>( \mu = \xi + 0.5772 \beta; )</td>
<td>Eq. (2.68); Eq. (2.70)</td>
</tr>
<tr>
<td>Weibull</td>
<td>( f_x(x) = \frac{\beta}{\pi} \left( \frac{x-\xi}{\sigma}\right)^{\alpha-1} \exp\left[-\left(\frac{x-\xi}{\sigma}\right)^{2}\right] )</td>
<td>( \alpha, \beta &gt; 0; x &gt; 0 )</td>
<td>( \mu = \beta \Gamma \left( 1 + \frac{1}{\alpha} \right); )</td>
<td>( \lambda_1 = \xi + \beta \Gamma \left( 1 + \frac{1}{\alpha} \right); )</td>
</tr>
<tr>
<td>Generalized extreme-value (GEV)</td>
<td>( F_x(x) = \exp\left{-\left[1 - \left(\frac{x-\xi}{\sigma}\right)^{\frac{1}{\alpha}}\right]^{1/\alpha}\right} )</td>
<td>( \alpha &gt; 0; x &lt; \left( \xi + \frac{\beta}{\sigma} \right); \alpha &lt; 0; x &gt; \left( \xi + \frac{\beta}{\sigma} \right) )</td>
<td>( \mu = \xi + \frac{\beta}{\Gamma(1+\alpha)} \Gamma(1+\alpha) )</td>
<td>( \lambda_1 = \xi + \frac{\beta}{\Gamma(1+\alpha)} \Gamma(1+\alpha) )</td>
</tr>
<tr>
<td>Generalized pareto (GPA)</td>
<td>( F_x(x) = 1 - \left[1 - \left(\frac{x-\xi}{\sigma}\right)^{\frac{1}{\alpha}}\right]^{1/\alpha} )</td>
<td>( \alpha &gt; 0; ) ( \zeta \leq x \leq \left( \xi + \frac{\beta}{\sigma} \right) ); ( \alpha &lt; 0; \zeta \leq x &lt; \infty )</td>
<td>( \mu = \xi + \frac{\beta}{\Gamma(1+\alpha)} \Gamma(1+\alpha) )</td>
<td>( \lambda_1 = \xi + \frac{\beta}{\Gamma(1+\alpha)} \Gamma(1+\alpha) )</td>
</tr>
</tbody>
</table>
For any distribution, the L-moments can be expressed in terms of the probability-weighted moments as shown in Eq. (2.28). To compute the sample L-moments, the sample probability-weighted moments can be obtained as

\[
\begin{align*}
l_1 &= b_0 \\
l_2 &= 2b_1 - b_0 \\
l_3 &= 6b_2 - 6b_1 + b_0 \\
l_4 &= 20b_3 - 30b_2 + 12b_1 - b_0
\end{align*}
\]

(3.15)

where the \( l \) s are sample estimates of the corresponding L-moments, the \( \lambda \) s, respectively. Accordingly, the sample L-moment ratios can be computed as

\[
\begin{align*}
t_2 &= \frac{l_2}{l_1} \\
t_3 &= \frac{l_3}{l_2} \\
t_4 &= \frac{l_4}{l_2}
\end{align*}
\]

(3.16)

where \( t_2, t_3, \) and \( t_4 \) are the sample L-coefficient of variation, L-skewness coefficient, and L-kurtosis, respectively. Relationships between L-moments and parameters of distribution models commonly used in frequency analysis are shown in the last column of Table 3.4.

**Example 3.6** Referring to Example 3.3, estimate the parameters of a generalized Pareto (GPA) distribution by the L-moment method.

**Solution** Since the GPA is a three-parameter distribution, the calculation of the first three sample L-moments is shown in the following table:

<table>
<thead>
<tr>
<th>Year</th>
<th>( q_i ) (ft³/s)</th>
<th>Ordered Rank</th>
<th>( q_i ) (ft³/s)</th>
<th>( F(q_i) = \frac{q_i - 0.35}{n} )</th>
<th>( q_i \times F(q_i) )</th>
<th>( q_i \times F(q_i)^2 )</th>
<th>( q_i \times F(q_i)^3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1961</td>
<td>390</td>
<td>1</td>
<td>342</td>
<td>0.0433</td>
<td>14.82</td>
<td>0.642</td>
<td>0.0278</td>
</tr>
<tr>
<td>1962</td>
<td>374</td>
<td>2</td>
<td>374</td>
<td>0.1100</td>
<td>41.14</td>
<td>4.525</td>
<td>0.4978</td>
</tr>
<tr>
<td>1963</td>
<td>342</td>
<td>3</td>
<td>390</td>
<td>0.1767</td>
<td>68.90</td>
<td>12.172</td>
<td>2.1504</td>
</tr>
<tr>
<td>1964</td>
<td>507</td>
<td>4</td>
<td>414</td>
<td>0.2433</td>
<td>100.74</td>
<td>24.513</td>
<td>5.9649</td>
</tr>
<tr>
<td>1965</td>
<td>596</td>
<td>5</td>
<td>416</td>
<td>0.3100</td>
<td>128.96</td>
<td>39.978</td>
<td>12.3931</td>
</tr>
<tr>
<td>1966</td>
<td>416</td>
<td>6</td>
<td>447</td>
<td>0.3767</td>
<td>168.37</td>
<td>63.419</td>
<td>23.8880</td>
</tr>
<tr>
<td>1967</td>
<td>533</td>
<td>7</td>
<td>505</td>
<td>0.4433</td>
<td>223.88</td>
<td>99.255</td>
<td>44.0030</td>
</tr>
<tr>
<td>1968</td>
<td>505</td>
<td>8</td>
<td>505</td>
<td>0.5100</td>
<td>257.55</td>
<td>131.351</td>
<td>66.9888</td>
</tr>
<tr>
<td>1969</td>
<td>549</td>
<td>9</td>
<td>507</td>
<td>0.5767</td>
<td>292.37</td>
<td>168.600</td>
<td>97.2260</td>
</tr>
<tr>
<td>1970</td>
<td>414</td>
<td>10</td>
<td>524</td>
<td>0.6433</td>
<td>337.11</td>
<td>216.872</td>
<td>139.5210</td>
</tr>
<tr>
<td>1971</td>
<td>554</td>
<td>11</td>
<td>533</td>
<td>0.7100</td>
<td>378.43</td>
<td>268.685</td>
<td>190.7666</td>
</tr>
<tr>
<td>1972</td>
<td>505</td>
<td>12</td>
<td>543</td>
<td>0.7767</td>
<td>421.73</td>
<td>327.544</td>
<td>254.3922</td>
</tr>
<tr>
<td>1973</td>
<td>447</td>
<td>13</td>
<td>549</td>
<td>0.8433</td>
<td>462.99</td>
<td>390.455</td>
<td>329.2836</td>
</tr>
<tr>
<td>1974</td>
<td>543</td>
<td>14</td>
<td>591</td>
<td>0.9100</td>
<td>537.81</td>
<td>489.407</td>
<td>445.3605</td>
</tr>
<tr>
<td>1975</td>
<td>591</td>
<td>15</td>
<td>596</td>
<td>0.9767</td>
<td>582.09</td>
<td>568.511</td>
<td>555.2459</td>
</tr>
<tr>
<td>Sum</td>
<td>7236</td>
<td></td>
<td></td>
<td>4016.89</td>
<td>2805.930</td>
<td>2167.710</td>
<td></td>
</tr>
</tbody>
</table>


Note that the plotting-position formula used in the preceding calculation is that proposed by Hosking et al. (1985a) with $a = 0.35$ and $b = 1$.

Based on Eq. (3.14), the sample estimates of $\beta_j$, for $j = 0, 1, 2, 3$, are $b_0 = 428.4$, $b_1 = 267.80$, $b_3 = 187.06$, and $b_4 = 144.51$. Hence, by Eq. (3.15), the sample estimates of $\lambda_j$, $j = 1, 2, 3, 4$, are $l_1 = 482.40$, $l_2 = 53.19$, $l_3 = -1.99$, and $l_4 = 9.53$, and the corresponding sample L-moment ratios $r_j$, for $j = 2, 3, 4$, are $t_2 = 0.110$, $t_3 = -0.037$, and $t_4 = 0.179$.

By referring to Table 3.4, the preceding sample $l_1 = 482.40$, $l_2 = 53.19$, and $t_3 = -0.037$ can be used in the corresponding L-moment and parameter relations, that is,

\[
l_1 = 482.40 = \xi + \frac{\beta}{1 + \alpha} \]
\[
l_2 = 53.19 = \frac{\beta}{(1 + \alpha)(2 + \alpha)} \]
\[
t_3 = -0.037 = \frac{1 - \alpha}{3 + \alpha} \]

Solving backwards starting from $t_3$, $l_2$, and then $l_1$, then, the values of sample parameter estimates of the GPA distribution can be obtained as

\[
\hat{\alpha} = 1.154 \quad \hat{\beta} = 361.36 \quad \hat{\xi} = 314.64
\]

### 3.7 Selection of Distribution Model

Based on a given sample of finite observations, procedures are needed to help identify the underlying distribution from which the random samples are drawn. Several statistical goodness-of-fit procedures have been developed (D’Agostino and Stephens, 1986). The insensitivity to the tail portion of the distribution of the conventional chi-square test and Kolmogorov-Smirnov test has been well known. Other more powerful goodness-of-fit criteria such as the probability plot correlation coefficient (Filliben, 1975) have been investigated and advocated (Vogel and McMartin, 1991). This and other criteria are described herein.

#### 3.7.1 Probability plot correlation coefficients

The probability plot is a graphic representation of the $m$th-order statistic of the sample $x_{(m)}$ as a function of a plotting-position $\hat{F}(x_{(m)})$. For each order statistic $X_{(m)}$, a plotting-position formula can be applied to estimate its corresponding nonexceedance probability $\hat{F}(x_{(m)})$, which, in turn, is used to compute the corresponding quantile $Y_m = G^{-1}[\hat{F}(X_{(m)})]$ according to the distribution model $G(\cdot)$ under consideration. Based on a sample with $n$ observations, the probability plot correlation coefficient (PPCC) then can be defined mathematically as

\[
PPCC = \frac{\sum_{m=1}^{n}(x_{(m)} - \bar{x})(y_m - \bar{y})}{\left[\sum_{m=1}^{n}(x_{(m)} - \bar{x})^2\right]^{0.5} \left[\sum_{m=1}^{n}(y_m - \bar{y})^2\right]^{0.5}} \tag{3.17}
\]
where $y_m$ is the quantile value corresponding to $\tilde{F}(x_{(m)})$ from a selected plotting-position formula and an assumed distribution model $G(\cdot)$, that is, $y_m = G^{-1}[\tilde{F}(x_{(m)})]$. It is intuitively understandable that if the samples to be tested are actually generated from the hypothesized distribution model $G(\cdot)$, the corresponding plot of $x_{(m)}$ versus $y_m$ would be close to linear. The values of $\tilde{F}(x_{(m)})$ for calculating $y_m$ in Eq. (3.17) can be determined by using either a probability- or quantile-unbiased plotting-position formula. The hypothesized distribution model $G(\cdot)$ that yields the highest value of the PPCC should be chosen.

Critical values of the PPCCs associated with different levels of significance for various distributions have been developed. They include normal and lognormal distribution (Fillben, 1975; Looney and Gulledge, 1985; Vogel, 1986), Gumbel distribution (Vogel, 1986), uniform and Weibull distributions (Vogel and Kroll, 1989), generalized extreme-value distribution (Chowdhury et al., 1991), Pearson type 3 distribution (Vogel and McMartin, 1991), and other distributions (D’Agostino and Stephens, 1986). A distribution is accepted as the underlying random mechanism with a specified significance level if the computed PPCC is larger than the critical value for that distribution.

### 3.7.2 Model reliability indices

Based on the observed $\{x_{(m)}\}$ and the computed $\{y_m\}$, the degree of goodness of fit also can be measured by two reliability indices proposed by Leggett and Williams (1981). They are the **geometric reliability index** $K_G$, 

$$K_G = \frac{1 + \sum_{m=1}^{n} \frac{n}{n} \left[ \frac{1 - (y_m/x_{(m)})}{1 + (y_m/x_{(m)})} \right]^2}{1 - \sum_{m=1}^{n} \frac{n}{n} \left[ \frac{1 - (y_m/x_{(m)})}{1 + (y_m/x_{(m)})} \right]^2}$$  \hspace{1cm} (3.18)

and the **statistical reliability index** $K_S$,

$$K_S = \exp \left\{ \frac{1}{n} \sum_{m=1}^{n} \log \left( \frac{y_m}{x_{(m)}} \right) \right\}$$  \hspace{1cm} (3.19)

When the computed series $\{y_m\}$ perfectly matches with the observed sequence $\{x_{(m)}\}$, the values of $K_G$ and $K_S$ reach their lower bound of 1.0. As the discrepancy between $\{x_{(m)}\}$ and $\{y_m\}$ increases, the values of $K_G$ and $K_S$ increase. Again, for each of $K_G$ and $K_S$, two different values can be computed, each associated with the use of probability-unbiased and quantile-unbiased plotting-position formulas. The most suitable probability model is the one that is associated with the smallest value of the reliability index.

### 3.7.3 Moment-ratio diagrams

Relationships between product moments and the parameters of various distributions are shown in Table 3.4, which also can be found elsewhere (Patel et al., 1976; Stedinger et al., 1993). Similarly, the **product-moment ratio diagram**
based on skewness coefficient and kurtosis (Stuart and Ord, 1987, p. 211) can be used to identify the distributions. When sample data are used, sample product moments are used to solve for the model parameters. However, owing to the low reliability of sample skewness coefficient and kurtosis, use of the product-moment ratio diagram for model identification is not reliable. Alternatively, the *L-moment ratio diagram* defined in the \((\tau_3, \tau_4)\)-space (Fig. 3.3) also can be used for model identification. Namely, one can judge the closeness of the sample L-skewness coefficient and L-kurtosis with respect to the theoretical \(\tau_3 - \tau_4\) curve associated with different distribution models. Some types of distance measures can be computed between the sample point of \((t_3, t_4)\) and each theoretical \(\tau_3 - \tau_4\) curve. One commonly used distance measure is to compute the shortest distance or distance in L-kurtosis direction fixed at the sample L-skewness coefficient (Pandey et al., 2001). Although it is computationally simple, the

![Figure 3.3 L-moment ratio diagram and shortest distance from a sample point.](image-url)
distance measure could not account for the sampling error in the sample L-skewness coefficient. To consider the effect of sampling errors in both the sample L-skewness coefficient and L-kurtosis, the shortest distance between the sample point \((t_3, t_4)\) and the theoretical \(\tau_3 - \tau_4\) curve of each candidate distribution model is computed for the measure of goodness of fit. The computation of the shortest distance requires locating a point on the theoretical \(\tau_3 - \tau_4\) curve that minimizes the distance as

\[
\text{DIS} = \min_{\tau_3} \sqrt{(t_3 - \tau_3)^2 + (t_4 - \tau_4(\tau_3))^2} \quad (3.20)
\]

Since the theoretical \(\tau_3 - \tau_4\) curve for a specified distribution is unique, determination of the shortest distance was accomplished by an appropriate one-dimensional search technique such as the golden-section procedure or others.

**Example 3.7 (Goodness of Fit)** Referring to the flood data given in Example 3.3, calculate the values of the probability-unbiased PPCCs and the two reliability indices with respect to the generalized Pareto distribution (GPA).

**Solution** Referring to Table 3.4, the GPA quantile can be obtained easily as

\[
x(F) = \xi + \frac{\beta}{\alpha}[1 - (1 - F)^{1/\alpha}] - (3.15)
\]

According to the model parameter values obtained from Example 3.6, that is, \(\hat{\alpha} = 1.154, \hat{\beta} = 361.36, \hat{\xi} = 314.64\), the GPA quantile can be computed as

\[
x(F) = 314.64 + \frac{361.36}{1.154} [1 - (1 - F)^{1.154}]
\]

Using the probability-unbiased plotting position, i.e., the Weibull formula, the corresponding GPA quantiles are calculated and shown in column (4) of the following table. From data in columns (2) and (4), the correlation coefficient can be obtained as 0.9843.

To calculate the two-model reliability indices, the ratios of GPA quantiles \(y_m\) to the order flow \(q(m)\) are calculated in column (5) and are used in Eqs. (3.18) and (3.19) for \(K_G\) and \(K_S\), respectively, as 1.035 and 1.015.

<table>
<thead>
<tr>
<th>Rank ((m))</th>
<th>Ordered (q(m))</th>
<th>(F(q(m)) = m/(n + 1))</th>
<th>(y_m)</th>
<th>(y_m/q(m))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>342</td>
<td>0.0625</td>
<td>337.1</td>
<td>0.985714</td>
</tr>
<tr>
<td>2</td>
<td>374</td>
<td>0.1250</td>
<td>359.4</td>
<td>0.960853</td>
</tr>
<tr>
<td>3</td>
<td>390</td>
<td>0.1875</td>
<td>381.4</td>
<td>0.977846</td>
</tr>
<tr>
<td>4</td>
<td>414</td>
<td>0.2500</td>
<td>403.1</td>
<td>0.973676</td>
</tr>
<tr>
<td>5</td>
<td>416</td>
<td>0.3125</td>
<td>424.6</td>
<td>1.020591</td>
</tr>
<tr>
<td>6</td>
<td>447</td>
<td>0.3750</td>
<td>445.7</td>
<td>0.997162</td>
</tr>
<tr>
<td>7</td>
<td>505</td>
<td>0.4375</td>
<td>466.6</td>
<td>0.923907</td>
</tr>
<tr>
<td>8</td>
<td>505</td>
<td>0.5000</td>
<td>487.1</td>
<td>0.964476</td>
</tr>
<tr>
<td>9</td>
<td>507</td>
<td>0.5625</td>
<td>507.2</td>
<td>1.000308</td>
</tr>
<tr>
<td>10</td>
<td>524</td>
<td>0.6250</td>
<td>526.8</td>
<td>1.005368</td>
</tr>
<tr>
<td>11</td>
<td>533</td>
<td>0.6875</td>
<td>546.0</td>
<td>1.024334</td>
</tr>
<tr>
<td>12</td>
<td>543</td>
<td>0.7500</td>
<td>564.5</td>
<td>1.039672</td>
</tr>
<tr>
<td>13</td>
<td>549</td>
<td>0.8125</td>
<td>582.4</td>
<td>1.060849</td>
</tr>
<tr>
<td>14</td>
<td>591</td>
<td>0.8750</td>
<td>599.4</td>
<td>1.014146</td>
</tr>
<tr>
<td>15</td>
<td>596</td>
<td>0.9375</td>
<td>615.0</td>
<td>1.031891</td>
</tr>
</tbody>
</table>
3.7.4 Summary

As the rule for selecting a single distribution model, the PPCC-based criterion would choose the model with highest values, whereas the other two criteria (i.e., reliability index and DIS) would select a distribution model with the smallest value. In practice, it is not uncommon to encounter a case where the values of the adopted goodness-of-fit criterion for different distributions are compatible, and selection of a best distribution may not necessarily be the best course of action, especially in the presence of sampling errors. The selection of acceptable distributions based on their statistical plausibility through hypothesis testing, at the present stage, can only be done for the PPCCs for which extensive experiments have been done to define critical values under various significance levels (or type I errors) and different distributions.

3.8 Uncertainty Associated with a Frequency Relation

Consider Example 3.2 in which the annual maximum flood peak discharges over a 15-year period on the Boneyard Creek at Urbana, Illinois, were analyzed. Suppose that the annual maximum floods follow the Gumbel distribution. The estimated 25-year flood peak discharge is 656 ft³/s. It is not difficult to imagine that if one had a second set of 15 years of record, the estimated 25-year flood based on the second 15-year record likely would be different from the first 15-year record. Also, combining with the second 15 years of record, the estimated 25-year flood magnitude based on a total of 30 years of record again would not have the same value as 656 ft³/s. This indicates that the estimated 25-year flood is subject to uncertainty that is due primarily to the use of limited amount of data in frequency analysis. Furthermore, it is intuitive that the reliability of the estimated 25-year flood, based on a 30-year record, is higher than that based on a 15-year record.

From the preceding discussions one can conclude that using a limited amount of data in frequency analysis, the estimated value of a geophysical quantity of a particular return period $x_T$ and the derived frequency relation are subject to uncertainty. The degree of uncertainty of the estimated $x_T$ depends on the sample size, the extent of data extrapolation (i.e., return period relative to the record length), and the underlying probability distribution from which the data are sampled (i.e., the distribution). Since the estimated design quantity is subject to uncertainty, it is prudent for an engineer to quantify the magnitude of such uncertainty and assess its implications on the engineering design (Tung and Yen, 2005, Sec. 1.5). Further, Benson (1968) noted that the results of the U.S. Water Resources Council study to determine the “best” distribution indicated that confidence limits always should be computed for flood frequency analysis.

In practice, there are two ways to express the degree of uncertainty of a statistical quantity, namely, standard error and confidence interval (confidence limit). Because the estimated geophysical quantities of a particular return period are subject to uncertainty, they can be treated as a random variable associated with a distribution, as shown in Fig. 3.4. Similar to the standard deviation of a
random variable, the \textit{standard error of estimate} $s_e$ measures the standard deviation of an estimated statistical quantity from a sample, such as $\hat{x}_T$, about the true but unknown event magnitude. On the other hand, the \textit{confidence limit} of an estimated quantity is an interval that has a specified probability (or confidence) to include the true value.

In the context of frequency analysis, the standard error of $\hat{x}_T$ is a function of the distribution of the data series under consideration and the method of determining the distribution parameters. For example, the asymptotic (that is, as $n \to \infty$) standard error of a $T$-year event $s_e(\hat{x}_T)$ from a normal distribution can be calculated as (Kite, 1988)

$$s_e(\hat{x}_T) = \left( \frac{2 + z_T^2}{2n} \right)^{1/2} s_x$$ (3.21)

in which $z_T$ is the standard normal variate corresponding to the exceedance probability of $1/T$, that is, $\Phi(z_T) = 1 - 1/T$, $n$ is the sample size, and $s_x$ is the sample standard deviation of random variable $X$. From the Gumbel distribution, the standard error of $\hat{x}_T$ is

$$s_e(\hat{x}_T) = \left( \frac{1}{n} \left( 1 + 1.1396 K_T + 1.1 K_T^2 \right) \right)^{1/2} s_x$$ (3.22)
To construct the confidence interval for \( \hat{x}_T \) or for the frequency curve, a confidence level \( c \) that specifies the desired probability that the specified range will include the unknown true value is predetermined by the engineer. In practice, a confidence level of 95 or 90 percent is used. Corresponding to the confidence level \( c \), the significance level \( \alpha \) is defined as \( \alpha = 1 - c \); for example, if the desired confidence level \( c = 90 \) percent, the corresponding significance level \( \alpha = 10 \) percent. In determining the confidence interval, the common practice is to distribute the significance level \( \alpha \) equally on both ends of the distribution describing the uncertainty feature of estimated \( \hat{x}_T \) (see Fig. 3.4). In doing so, the boundaries of the confidence interval, called confidence limits, are defined. Assuming normality for the asymptotic sample distribution for \( \hat{x}_T \), the approximated 100(1 - \( \alpha \)) percent confidence interval for \( \hat{x}_T \) is

\[
\begin{align*}
x_{L,T,\alpha} &= \hat{x}_T - z_{1-\alpha/2} \times s_e(\hat{x}_T) \\
x_{U,T,\alpha} &= \hat{x}_T + z_{1-\alpha/2} \times s_e(\hat{x}_T)
\end{align*}
\] (3.23)

in which \( x_{L,T,\alpha} \) and \( x_{U,T,\alpha} \) are, respectively, the values defining the lower and upper bounds for the 100(1 - \( \alpha \)) percent confidence interval, and \( z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2) \). The confidence interval defined by Eq. (3.23) is only approximate and the approximation accuracy increases with sample size.

Similar to the frequency-factor method, the formulas to compute the upper and lower limits of confidence interval for \( \hat{x}_T \) has the same form as Eq. (3.6), except that the frequency-factor term is adjusted as

\[
\begin{align*}
x_{L,T,\alpha} &= \bar{x} + K_{L,T,\alpha} \times s_x \\
x_{U,T,\alpha} &= \bar{x} + K_{U,T,\alpha} \times s_x
\end{align*}
\] (3.24)

in which \( K_{L,T,\alpha} \) and \( K_{U,T,\alpha} \) are the confidence-limit factors for the lower and upper limits of the 100(1 - \( \alpha \)) percent confidence interval, respectively. For random samples from a normal distribution, the exact confidence-limit factors can be determined using the noncentral-\( t \) variates \( \xi \) (Table 3.5). An approximation for \( K_{L,T,\alpha} \) with reasonable accuracy for \( n \geq 15 \) and \( \alpha = 1 - c \geq 5 \) percent (Chowdhury et al., 1991) is

\[
K_{L,T,\alpha} \approx \xi_{T,\alpha/2} \approx \frac{z_T + z_{\alpha/2} \sqrt{\frac{1}{n} + \frac{z_T^2}{2(n-1)} - \frac{z_{\alpha/2}^2}{2(n-1)}}}{1 - \frac{z_{\alpha/2}^2}{2(n-1)}}
\] (3.25)

To compute \( K_{U,T,\alpha} \), by symmetry, one only has to change \( z_{\alpha/2} \) by \( z_{1-\alpha/2} \) in Eq. (3.25). As was the case for Eq. (3.20), the confidence intervals defined by Eqs. (3.24) and (3.25) are most appropriate for samples from populations following a normal distribution, and for nonnormal populations, these confidence limits are only approximate, with the approximation accuracy increasing with sample size.

For Pearson type 3 distributions, the values of confidence-limit factors for different return periods and confidence levels given in Eq. (3.24) can be modified by introducing the scaling factor obtained from a first-order asymptotic
<table>
<thead>
<tr>
<th>n</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_{L,T}$</td>
<td>$K_{U,T}$</td>
<td>$K_{L,T}$</td>
<td>$K_{U,T}$</td>
<td>$K_{L,T}$</td>
<td>$K_{U,T}$</td>
</tr>
<tr>
<td>15</td>
<td>-0.4468</td>
<td>0.4468</td>
<td>0.3992</td>
<td>1.4641</td>
<td>0.7908</td>
<td>2.0464</td>
</tr>
<tr>
<td>20</td>
<td>-0.3816</td>
<td>0.3816</td>
<td>0.4544</td>
<td>1.3579</td>
<td>0.8495</td>
<td>1.9101</td>
</tr>
<tr>
<td>25</td>
<td>-0.3387</td>
<td>0.3387</td>
<td>0.4925</td>
<td>1.2913</td>
<td>0.8905</td>
<td>1.8257</td>
</tr>
<tr>
<td>30</td>
<td>-0.3076</td>
<td>0.3076</td>
<td>0.5209</td>
<td>1.2447</td>
<td>0.9213</td>
<td>1.7672</td>
</tr>
<tr>
<td>40</td>
<td>-0.2647</td>
<td>0.2647</td>
<td>0.5613</td>
<td>1.1824</td>
<td>0.9654</td>
<td>1.6898</td>
</tr>
<tr>
<td>50</td>
<td>-0.2359</td>
<td>0.2359</td>
<td>0.5892</td>
<td>1.1418</td>
<td>0.9961</td>
<td>1.6398</td>
</tr>
<tr>
<td>60</td>
<td>-0.2148</td>
<td>0.2148</td>
<td>0.6100</td>
<td>1.1127</td>
<td>1.0191</td>
<td>1.6014</td>
</tr>
<tr>
<td>70</td>
<td>-0.1986</td>
<td>0.1986</td>
<td>0.6263</td>
<td>1.0906</td>
<td>1.0371</td>
<td>1.5772</td>
</tr>
<tr>
<td>80</td>
<td>-0.1855</td>
<td>0.1855</td>
<td>0.6396</td>
<td>1.0730</td>
<td>1.0518</td>
<td>1.5559</td>
</tr>
<tr>
<td>90</td>
<td>-0.1747</td>
<td>0.1747</td>
<td>0.6506</td>
<td>1.0586</td>
<td>1.0641</td>
<td>1.5385</td>
</tr>
<tr>
<td>100</td>
<td>0.1656</td>
<td>0.1656</td>
<td>0.6599</td>
<td>1.0466</td>
<td>1.0746</td>
<td>1.5240</td>
</tr>
</tbody>
</table>
approximation of the Pearson type 3 to normal quantile variance ratio \( \eta \) as (Stedinger et al., 1983)

\[
K_{T,\alpha}^L = K_T + \eta (\zeta_{T,1-\alpha/2} - z_T) \quad \text{and} \quad K_{T,\alpha}^U = K_T + \eta (\zeta_{T,\alpha/2} - z_T) \quad (3.26)
\]

where

\[
\eta = \sqrt{1 + \hat{\gamma}_x K_T + \frac{1}{2} \left( 1 + \frac{3}{4} \hat{\gamma}_x \right) K_T^2 + n \text{ var}(\hat{\gamma}_x)(\partial K_T / \partial \gamma_x)^2 / \left( 1 + (1/2)z_T^2 \right)} \quad (3.27)
\]

in which \( \hat{\gamma}_x \) is the estimated skewness coefficient, and

\[
\frac{\partial K_T}{\partial \gamma_x} \approx \frac{1}{6} (z_T^2 - 1) \left[ 1 - 3 \left( \frac{\hat{\gamma}_x}{6} \right)^2 + (z_T^2 - 6z_T) \frac{\hat{\gamma}_x}{54} + \frac{2}{3} z_T \left( \frac{\hat{\gamma}_x}{6} \right)^3 \right] \quad (3.28)
\]

A simulation study by Whitley and Hromadka (1997) showed that the approximated formula for the Pearson type 3 distribution is relatively crude and that a better expression could be derived for more accurate confidence-interval determination.

Example 3.8 Referring to Example 3.3, determine the 95 percent confidence interval of the 100-year flood assuming that the sample data are from a lognormal distribution.

Solution In this case, with the 95 percent confidence interval \( c = 0.95 \), the corresponding significance level \( \alpha = 0.05 \). Hence \( z_{0.025} = \Phi^{-1}(0.025) = -1.960 \) and \( z_{0.975} = \Phi^{-1}(0.975) = +1.960 \). Computation of the 95 percent confidence interval associated with the selected return periods is shown in the table below. Column (4) lists the values of the upper tail of the standard normal quantiles associated with each return period, that is, \( K_T = z_T = \Phi^{-1}(1 - 1/T) \). Since random floods are assumed to be lognormally distributed, columns (7) and (8) are factors computed by Eq. (3.25) for defining the lower and upper bounds of the 95 percent confidence interval of different quantiles in log-space, according to Eq. (3.24), as

\[
y_{T,0.95}^L = \bar{y} + \zeta_{T,0.025} \times s_y \quad y_{T,0.95}^U = \bar{y} + \zeta_{T,0.975} \times s_y
\]

In the original space, the 95 percent confidence interval can be obtained simply by taking exponentiation as

\[
q_{T,0.95}^L = \exp(y_{T,0.95}^L) \quad \text{and} \quad q_{T,0.95}^U = \exp(y_{T,0.95}^U)
\]

as shown in columns (11) and (12), respectively. The curves defining the 95 percent confidence interval, along with the estimated frequency curve, for a lognormal distribution are shown in Fig. 3.5.
95% CL for lognormal

Figure 3.5 95 percent confidence limits for a lognormal distribution applied to the annual maximum discharge for 1961–1975 on the Boneyard Creek at Urbana, IL.

<table>
<thead>
<tr>
<th>Return period</th>
<th>Exceedance probability</th>
<th>Nonexceedance probability</th>
<th>$K_T = z_T$</th>
<th>$y_T$</th>
<th>$q_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$ (years)</td>
<td>$1 - p = 1/T$</td>
<td>$p = 1 - 1/T$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0000</td>
<td>6.165</td>
<td>475.9</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8416</td>
<td>6.311</td>
<td>550.3</td>
</tr>
<tr>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
<td>1.2816</td>
<td>6.386</td>
<td>593.7</td>
</tr>
<tr>
<td>25</td>
<td>0.04</td>
<td>0.96</td>
<td>1.7505</td>
<td>6.467</td>
<td>643.8</td>
</tr>
<tr>
<td>50</td>
<td>0.02</td>
<td>0.98</td>
<td>2.0537</td>
<td>6.520</td>
<td>678.3</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
<td>0.99</td>
<td>2.3263</td>
<td>6.567</td>
<td>711.0</td>
</tr>
</tbody>
</table>

In order to define confidence limits properly for the Pearson type 3 distribution, the skewness coefficient must be estimated accurately, thus allowing the frequency factor $K_T$ to be considered a constant and not a statistic. Unfortunately, with the Pearson type 3 distribution, no simple, explicit formula is available for the confidence limits. The Interagency Advisory Committee on Water Data (1982) (hereafter referred to as “the Committee”) proposed that the confidence limits for the log-Pearson type 3 distribution could be approximated using a noncentral $t$-distribution. The committee’s procedure is similar to that of Eqs. (3.24) and (3.25), except that $K_{T,\alpha}^L$ and $K_{T,\alpha}^U$, the confidence-limit factors for the lower and upper limits, are computed with the frequency factor $K_T$ replacing $Z_T$ in Eq. (3.25).
Example 3.9 Referring to Example 3.3, determine the 95 percent confidence intervals for the 2-, 10-, 25-, 50-, and 100-year floods assuming that the sample data are from a log-Pearson type 3 distribution.

Solution From Example 3.3, the mean and standard deviation of the logarithms of the peak flows were 6.17 and 0.173, and the number of data $n$ is 15. For the 100-year flood, $K_T$ is 1.8164, and for the 95 percent confidence limits, $a$ is 0.05; thus $Z_a/2$ is $-1.96$. Thus $K_T^a$ is $-0.651$, and the lower 95 percent confidence bound is 427.2 ft$^3$/s. The upper and lower confidence bounds for all the desired flows are listed in the following table:

<table>
<thead>
<tr>
<th>Return Period</th>
<th>$T$ (years)</th>
<th>$K_T$ Eq. (3.7)</th>
<th>$K_{T,0.025}$ Eq. (3.25)</th>
<th>$K_{T,0.975}$ Eq. (3.25)</th>
<th>$q_{T,0.95}^L$ (ft$^3$/s)</th>
<th>$q_{T,0.95}^U$ (ft$^3$/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.0907</td>
<td>-0.6513</td>
<td>0.4411</td>
<td>427.2</td>
<td>516.1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.0683</td>
<td>0.5260</td>
<td>1.9503</td>
<td>523.7</td>
<td>670.1</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1.4248</td>
<td>0.8322</td>
<td>2.4705</td>
<td>552.2</td>
<td>733.2</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.6371</td>
<td>1.0082</td>
<td>2.7867</td>
<td>569.3</td>
<td>774.4</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.8164</td>
<td>1.1540</td>
<td>3.0565</td>
<td>583.8</td>
<td>811.4</td>
<td></td>
</tr>
</tbody>
</table>

3.9 Limitations of Hydrologic Frequency Analysis

3.9.1 Distribution Selection: Practical Considerations

Many different probability distributions have been proposed for application to hydrologic data. Some of them were proposed because the underlying concept of the distribution matched the goal of hydrologic frequency analysis. For example, the extremal distributions discussed in Sec. 2.6.4 have very favorable properties for hydrologic frequency analysis. Ang and Tang (1984, p. 206) noted that the asymptotic distributions of extremes in several cases tend to converge on certain limiting forms for large sample sizes $n$, specifically to the double-exponential form or to two single-exponential forms. The extreme value from an initial distribution with an exponentially decaying tail (in the direction of the extreme) will converge asymptotically to the extreme-value type I (Gumbel) distribution form. Distributions with such exponentially decaying tails include the normal distribution and many others listed in Sec. 2.6. This is why Gumbel (1941) first proposed this distribution for floods, and it has gained considerable popularity since then. Also, the properties of the central limit theorem discussed in Sec. 2.6.2 have made the lognormal distribution a popular choice for hydrologic frequency analysis.

In the 1960s, as the number of different approaches to flood frequency analysis were growing, a working group of U.S. government agency hydrologic experts was formed by the U.S. Water Resources Council to evaluate the best/preferred approach to flood frequency analysis. Benson (1968) reviewed the results of this working group and listed the following key results of their study:

1. There is no physical rule that requires the use of any specific distribution in the analysis of hydrologic data.
2. Intuitively, there is no reason to expect that a single distribution will apply to all streams worldwide.

3. No single method of testing the computed results against the original data was acceptable to all those on the working group, and the statistical consultants could not offer a mathematically rigorous method.

Subsequent to this study, the U.S. Water Resources Council (1967) recommended use of the log-Pearson type 3 distribution for all flood frequency analyses in the United States, and this has become the official distribution for all flood frequency studies in the United States. There are no physical arguments for the application of this distribution to hydrologic data. It has added flexibility over two-parameter distributions (e.g., Gumbel, lognormal) because the skewness coefficient is a third independent parameter, and the use of three parameters generally results in a better fit of the data. However, a number of researchers have suggested that the use of data for a single site may be insufficient to estimate the skewness coefficient properly.

Beard (1962) recommended that only average regional skewness coefficients should be applied in flood frequency analysis for a single station unless that record exceeds 100 years. This led the U.S. Water Resources Council (1967) to develop maps of regional skewness coefficient values that are averaged with the at-a-site skewness coefficient as a function of the number of years of record. For details on the procedure, see Interagency Advisory Committee on Water Data (1982). Linsley et al. (1982) noted that although regional skewness coefficients may not make for more reliable analysis, their use does lead to more consistency between values for various streams in the region.

### 3.9.2 Extrapolation problems

Most often frequency analysis is applied for the purpose of estimating the magnitude of truly rare events, e.g., a 100-year flood, on the basis of short data series. Viessman et al. (1977, pp. 175–176) note that “as a general rule, frequency analysis should be avoided...in estimating frequencies of expected hydrologic events greater than twice the record length.” This general rule is followed rarely because of the regulatory need to estimate the 100-year flood; e.g., the U.S. Water Resources Council (1967) gave its blessing to frequency analyses using as few as 10 years of peak flow data. In order to estimate the 100-year flood on the basis of a short record, the analyst must rely on extrapolation, wherein a law valid inside a range of \( p \) is assumed to be valid outside of \( p \). The dangers of extrapolation can be subtle because the results may look plausible in the light of the analyst’s expectations.

The problem with extrapolation in frequency analysis can be referred to as “the tail wagging the dog.” In this case, the “tail” is the annual floods of relatively high frequency (1- to 10-year events), and the “dog” is the estimation of extreme floods needed for design (e.g., the floods of 50-, 100-, or even higher-year return periods). When trying to force data to fit a mathematical distribution,
equal weight is given to the low end and high end of the data series when trying to determine high-return-period events. Figure 3.6 shows that small changes in the three smallest annual peaks can lead to significant changes in the 100-year peak owing to “fitting properties” of the assumed flood frequency distribution. The analysis shown in Fig. 3.6 is similar to the one presented by Klemes (1986); in this case, a 26-year flood series for Gilmore Creek at Winona, Minnesota, was analyzed using the log-Pearson type 3 distribution employing the skewness coefficient estimated from the data. The three lowest values in the annual maximum series (22, 53, and 73 ft$^3$/s) then were changed to values of 100 ft$^3$/s (as if a crest-stage gauge existed at the site with a minimum flow value of 100 ft$^3$/s), and the log-Pearson type 3 analysis was repeated. The relatively small absolute change in these three events changed the skewness coefficient from 0.039 to 0.648 and the 100-year flood from 7,030 to 8,530 ft$^3$/s. As discussed by Klemes (1986), it is illogical that the 1- to 2-year frequency events should have such a strong effect on the rare events.

Under the worst case of hydrologic frequency analysis, the frequent events can be caused by a completely different process than the extreme events. This situation violates the initial premise of hydrologic frequency analysis, i.e., to find some statistical relation between the magnitude of an event and its likelihood of occurrence (probability) without regard for the physical process of flood formation. For example, in arid and semiarid regions of Arizona, frequent events (1- to 5-year events) are caused by convective storms of limited spatial extent, whereas the major floods (>10-year events) are caused by frontal
monsoon-type storms that distribute large amounts of rainfall over large areas for several days. Figure 3.7 shows the daily maximum discharge series for the Agua Fria River at Lake Pleasant, Arizona, for 1939–1979 and clearly indicates a difference in magnitude and mechanism between frequent and infrequent floods. In this case estimating the 100-year flood giving equal weight in the statistical calculations to the 100 ft\(^3\)/s and the 26,000 ft\(^3\)/s flows seems inappropriate, and an analyst should be prepared to use a large safety factor if standard frequency analysis methods were applied.

Another problem with “the tail wagging the dog” results when the watershed experiences substantial changes. For example, in 1954 the Vermilion River, Illinois, Outlet Drainage District initiated a major channelization project involving the Vermilion River, its North Fork, and North Fork tributaries. The project was completed in the summer of 1955 and resulted in changing the natural 35-ft-wide North Fork channel to a trapezoidal channel 100 ft in width and the natural 75-ft-wide Vermilion channel to a trapezoidal channel 166 ft in width. Each channel also was deepened 1 to 6 ft (U.S. Army Corps of Engineers, 1986). Discharges less than about 8,500 ft\(^3\)/s at the outlet remain in the modified channel, whereas those greater than 8,500 ft\(^3\)/s go overbank. At some higher discharge, the overbank hydraulics dominate the flow, just as they did before the channelization. Thus the more frequent flows are increased by the improved hydraulic efficiency of the channel, whereas the infrequent events are still subject to substantial attenuation by overbank flows. Thus the frequency curve is flattened relative to the prechannelization condition, where the more frequent events are also subject to overbank attenuation. The pre- and postchannelization flood frequency curves cross in the 25- to 50-year return period range.

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{fig37}
\caption{Return periods for the annual maximum daily flow of the Agua Fria River at Lake Pleasant, Arizona, for 1939–1979.}
\end{figure}
(Fig. 3.8), resulting in the illogical result that the prechannelization condition results in a higher 100-year flood than the postchannelization condition. Similar results have been seen for flood flows obtained from continuous simulation applied to urbanizing watersheds (Bradley and Potter, 1991).

### 3.9.3 The stationarity assumption

Viessman et al. (1977, p. 158) noted that “usually, the length of record as well as the design life for an engineering project are relatively short compared with geologic history and tend to temper, if not justify, the assumption of stationarity.” On the other hand, Klemes (1986) noted that there are many known causes for nonstationarity ranging from the dynamics of the earth’s motion to human-caused changes in land use. In this context, Klemes (1986) reasons that the notion of a 100-year flood has no meaning in terms of average return period, and thus the 100-year flood is really a reference for design rather than a true reflection of the frequency of an event.

### 3.9.4 Summary comments

The original premise for the use of hydrologic frequency analysis was to find the optimal project size to provide a certain protection level economically, and the quality of the optimization is a function of the accuracy of the estimated flood level. The preceding discussions in this section have indicated that the accuracy of hydrologic frequency estimates may not be high. For example, Beard (1987) reported that the net result of studies of uncertainties of flood frequency analysis is that standard errors of estimated flood magnitudes are very high—on the order of 10 to 50 percent depending on the stream characteristics and amount of data available.

Even worse, the assumptions of hydrologic frequency analysis, namely, stationarity and homogeneous, representative data, and good statistical
modeling—not extrapolating too far beyond the range of the data—may be violated or stretched in common practice. This can lead to illogical results such as the crossing of pre- and post-change frequency curves illustrated in Fig. 3.8, and the use of such illogical results is based on “a subconscious hope that nature can be cheated and the simple logic of mathematical manipulations can be substituted for the hidden logic of the external world” (Klemes, 1986).

Given the many potential problems with hydrologic frequency analysis, what should be done? Klemes (1986) suggested that if hydrologic frequency theorists were good engineers, they would adopt the simplest procedures and try to standardize them in view of the following facts:

1. The differences in things such as plotting positions, parameter-estimation methods, and even the distribution types, may not matter much in design optimization (Slack et al., 1975). Beard (1987) noted that no matter how reliable flood frequency estimates are, the actual risk cannot be changed. Thus the benefits from protection essentially are a function of investment and are independent of uncertainties in estimating flood frequencies. Moderate changes in protection or zoning do not change net benefits greatly; i.e., the benefit function has a broad, flat peak (Beard, 1987).

2. There are scores of other uncertain factors in the design that must be settled, but in a rather arbitrary manner, so the whole concept of optimization must be taken as merely an expedient design procedure. The material covered in Chaps. 4, 6, 7, and 8 of this book provide methods to consider the other uncertain factors and improve the optimization procedure.

3. Flood frequency analysis is just one convenient way of rationalizing the old engineering concept of a safety factor rather than a statement of hydrologic truth.

Essentially, the U.S. Water Resources Council (1967) was acting in a manner similar to Klemes’ approach in that a standardized procedure was developed and later improved (Interagency Advisory Committee on Water Data, 1982). However, rather than selecting and standardizing a simple procedure, the relatively more complex log-Pearson type 3 procedure was selected. Beard (1987) suggested that the U.S. Water Resources Council methods are the best currently available but leave much to be desired.

**Problems**

Given are the significant independent peak discharges measured on the Saddle River at Lodi, NJ, for two 18-year periods 1948–1965 and 1970–1987. The Saddle River at Lodi has a drainage area of 54.6 mi² primarily in Bergen County. The total data record for peak discharge at this gauge is as follows: 1924–1937 annual peak only, 1938–1987 all peaks above a specified base value, 1988–1989 annual peak only (data are missing for 1966, 1968, and 1969, hence the odd data periods).
### Hydrologic Frequency Analysis

<table>
<thead>
<tr>
<th>Water year</th>
<th>Date</th>
<th>$Q_p$ (ft³/s)</th>
<th>Water year</th>
<th>Date</th>
<th>$Q_p$ (ft³/s)</th>
<th>Water year</th>
<th>Date</th>
<th>$Q_p$ (ft³/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8/10/65</td>
<td>1020</td>
<td></td>
<td>4/10/80</td>
<td>2470</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4/29/80</td>
<td>2370</td>
</tr>
<tr>
<td>1949</td>
<td>12/31/48</td>
<td>1030</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1950</td>
<td>3/24/50</td>
<td>452</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4/03/70</td>
<td>2130</td>
<td></td>
<td>5/12/81</td>
<td>1900</td>
</tr>
<tr>
<td>3/12/52</td>
<td>1100</td>
<td></td>
<td></td>
<td>9/12/71</td>
<td>3770</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/06/52</td>
<td>1470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6/02/52</td>
<td>1740</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/25/53</td>
<td>993</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4/16/83</td>
<td>2550</td>
</tr>
<tr>
<td>4/08/53</td>
<td>1090</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1954</td>
<td>9/12/54</td>
<td>1270</td>
<td>1974</td>
<td>12/21/73</td>
<td>2940</td>
<td>1984</td>
<td>10/24/83</td>
<td>1510</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12/13/83</td>
<td>2610</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4/05/84</td>
<td>3350</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5/30/84</td>
<td>2840</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7/07/84</td>
<td>2990</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7/14/75</td>
<td>2720</td>
<td></td>
<td>9/27/85</td>
<td>2120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9/27/75</td>
<td>2350</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1956</td>
<td>10/16/55</td>
<td>1530</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1957</td>
<td>11/02/56</td>
<td>795</td>
<td>1976</td>
<td>4/01/76</td>
<td>1590</td>
<td>1986</td>
<td>1/26/86</td>
<td>1850</td>
</tr>
<tr>
<td>4/06/57</td>
<td>795</td>
<td></td>
<td></td>
<td>7/01/76</td>
<td>2440</td>
<td></td>
<td>8/17/86</td>
<td>1660</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1958</td>
<td>1/22/58</td>
<td>964</td>
<td>1977</td>
<td>2/25/77</td>
<td>3130</td>
<td>1987</td>
<td>12/03/86</td>
<td>2310</td>
</tr>
<tr>
<td>2/28/58</td>
<td>1760</td>
<td></td>
<td></td>
<td>3/23/77</td>
<td>2380</td>
<td></td>
<td>4/04/87</td>
<td>2320</td>
</tr>
<tr>
<td>4/07/58</td>
<td>1100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1959</td>
<td>3/07/59</td>
<td>795</td>
<td>1978</td>
<td>11/09/77</td>
<td>4500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1/26/78</td>
<td>1980</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3/27/78</td>
<td>1610</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1960</td>
<td>9/13/60</td>
<td>1190</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1961</td>
<td>2/26/61</td>
<td>952</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1962</td>
<td>3/13/62</td>
<td>1670</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1963</td>
<td>3/07/63</td>
<td>824</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1964</td>
<td>1/10/64</td>
<td>702</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 3.1 Determine the annual maximum series.

#### 3.2 Plot the annual maximum series on normal, lognormal, and Gumbel probability papers.

#### 3.3 Calculate the first four product moments and L-moments based on the given peak-flow data in both the original and logarithmic scales.

#### 3.4 Use the frequency-factor approach to the Gumbel, lognormal, and log-Pearson type 3 distributions to determine the 5-, 25-, 50-, and 100-year flood peaks.
3.5 Based on the L-moments obtained in Problem 3.3, determine the 5-, 25-, 50-, and 100-year flood peaks using Gumbel, generalized extreme value (GEV), and lognormal distributions.

3.6 Determine the best-fit distribution for the annual maximum peak discharge series based on the probability-plot correlation coefficient, the two model reliability indices, and L-moment ratio diagram.

3.7 Establish the 95 percent confidence interval for the frequency curve derived based on lognormal and log-Pearson type 3 distribution models.

References


